博士論文

# Causal Machine Learning from Small Data: A Data Augmentation Approach

（小データからの因果的機械学習：
データ拡張によるアプローチ）

手嶋 毅志

博士論文

# Causal Machine Learning from Small Data: A Data Augmentation Approach

（小データからの因果的機械学習：
データ拡張によるアプローチ）

東京大学大学院
新領域創成科学研究科

手嶋 毅志

# Abstract

The twenty-first century has seen a rapid and widespread development of automated intelligent systems, such as computer-assisted diagnosis, object detection, speech recognition, and automated translation. Many of such systems are powered by *statistical machine learning*, a paradigm in which various methods for learning from data have been developed. The designs of statistical machine learning parallel *inductive inference* in logical reasoning: learning from observations — the *training data* — and drawing conclusions about the unobserved — the *testing instances*.

For inductive reasoning to be helpful, some "uniformity of nature" principle is required. In statistical machine learning, or more generally in statistics, analogues of such a *uniformity principle* are embedded in one form or another, such as an assumption that the data is an *independent and identically distributed* sample of a probability distribution or that one should be a rational decision-maker. Corresponding to each of such different premises, reasonable inferential rules — the *learning methods* — have been derived.

*Causality* is a form of such a *uniformity principle*, which is distinctive in humans' course of thinking and perception of the world. What is causality? Some philosophical theories of causation emphasize that causality is about *difference-making*: without the cause, the effect would not have happened. In particular, some view causal relations as potential routes by which the world can be manipulated or controlled, i.e., difference-making about what *potential outcome* is realized by some intervention. Others emphasize that causality is about *production*: causes *bring about* their effects. In particular, some appeal to the concept of *causal mechanisms*, i.e., complex systems producing some behaviors through invariant direct interaction of a number of parts, as is often considered in the explanatory practices of the special sciences.

Founded by these two viewpoints of philosophical theories of causality, namely *interventions* and *mechanisms*, the *statistical frameworks of causal modeling* have been developed since the end of the 20th century. Such frameworks enable natural formulations of causality-related quantities based on probability theory, such as the average difference made by an intervention.

The pragmatic utility of acquiring the knowledge of such intervention-related quantities is rather apparent: one can use it for making informed decisions about the actions, e.g., answering questions such as what intervention to perform to get a favorable result. From this viewpoint, the knowledge of detailed mechanisms is only a *means* to infer the consequences of our interventions. On the other hand, our intellectual curiosity to learn about the causality of nature seems to go beyond the pragmatic utility of knowing interventionistic quantities. Indeed, elucidating a mechanism has been a gold standard for explanations in scientific practice, even when making interventions is not necessarily an immediate target in such fields.

Then, a natural question arises: are there pragmatic motivations for finding out the detailed causal mechanisms when the knowledge may not be relevant to any interventions we can implement? This dissertation provides a partial but concrete answer: the knowledge of causal mechanisms can facilitate *learning from small data* in *statistical machine learning* for predictive modeling. We provide the answer by designing the methods to incorporate the knowledge encoded in statistical causal models into the learning process.

Learning from small data, despite the rapid progress in the methodology of machine learning, remains a fundamental challenge in the field. When data is limited in quantity, it is essential to incorporate appropriate prior knowledge about the nature of the data in order to learn an accurate predictor. In this dissertation, we approach the small-data learning problem from the perspective of exploiting known or acquired causal knowledge. The general idea is to incorporate the *statistical independence* relations implied by the statistical causal models into the machine learning procedures by developing *data augmentation* strategies.

The following is the chapter organization. Chapter 1 provides the conceptual background and declares the central statement of this dissertation, and it is followed by Chapter 2, where we review the *structural causal framework* of statistical causal modeling. Specifically, we introduce two interrelated formulations: structural causal models (SCMs) and graphical causal models (GCMs). The two types of models form a hierarchy: SCMs capture the quantitative knowledge of the data-generating mechanisms expressed using deterministic functions, and GCMs retain only the coarser qualitative knowledge of the dependency relations in the data-generating mechanisms expressed using a graphical representation.

Following these introductory chapters, in the main Chapters 3 and 4, we develop the proposals for exploiting the knowledge of the causal models for supervised machine learning. Concretely, in Chapter 3, we consider the case that the graphical representation of a GCM is either estimable or known thanks to domain experts. In Chapter 4, we consider the case that partial knowledge of the deterministic functions of an SCM is estimable from the data of a relevant domain. When the GCMs or SCMs characterizing the data-generating mechanisms are (partially) known, we can infer some properties of the probability distribution of the data, namely certain statistical independence relations. However, it is not straightforward to incorporate such knowledge into predictive modeling. Therefore, in these chapters, we introduce *data augmentation* methods that allow us to exploit the knowledge encoded in the causal models for supervised machine learning in a manner that is independent of the predictor's model class which we use. The proposed methods enjoy theoretical guarantees of *excess risk bounds* indicating that the proposed methods suppress overfitting by reducing the apparent complexity of the predictor hypothesis class. Using real-world data conforming to the problem setups, we also provide numerical experiments showing that the proposed method effectively improves the prediction accuracy, especially in the small-data regime.

We dedicate Chapter 5 to presenting a theoretical result that reinforces the justification of the method of Chapter 4, which relies on the modeling technique called *invertible neural networks* (INNs). As a recently emerged function approximation model, the INNs had not been given a theoretical guarantee of their representation power, i.e., whether the model class theoretically has sufficient flexibility to approximate various complex functions. This was a critical concern that could undermine the applicability of the proposed method of Chapter 4 to a broad range of applications. The results in Chapter 5 are affirmative: the INNs used in Chapter 4 enjoy a theoretical representation power guarantee, namely that they are *universal approximators* for a fairly large class of smooth invertible maps. We use Chapter 5 to discuss this result in length because it is also an interesting theoretical result in its own right whose scope is not limited to supplementing Chapter 4.

Finally, in Chapter 6, we summarize the overall conclusion of the dissertation and discuss further the possibilities of future research directions. To summarize, the chapters of this dissertation jointly provide the affirmation of the thesis that the causal knowledge captured by statistical causal models can be helpful in tackling the small-data learning problems in statistical machine learning. The proposed strategy for exploiting the causal knowledge is based on data augmentation, and thus the proposed methods can be readily combined with virtually any supervised learning method for learning a predictor.

# Acknowledgments

First and foremost, I would like to express my most profound appreciation to my supervisor Prof. Masashi Sugiyama for his guidance, encouragement, and continuous support. His advice throughout my graduate studies has formed my views and foundations of conducting research. In retrospect, what has brought me to his laboratory was his "Ganbarimashou" (a short phrase of encouragement in Japanese). It has been a magical phrase that helped me keep up when I face the pressure and the challenges as a novice researcher. I am, and will always be, deeply grateful for his patience and warm encouragement.

I would also like to thank the dissertation committee members: Prof. Masaaki Imaizumi, Prof. Yuki Izumida, Prof. Shohei Shimizu, and Prof. Naoto Yokoya, for their valuable comments and suggestions for improving the dissertation. Special thanks go to Prof. Jun Otsuka for reading the manuscript of this dissertation and having a constructive discussion from a science-philosophical viewpoint.

I am also very thankful to Prof. Issei Sato, whose invaluable advice has nurtured how I approach the research. I feel privileged to have worked with my exceptional collaborators, Prof. Isao Ishikawa, Dr. Masahiro Ikeda, Dr. Koichi Tojo, and Dr. Kenta Oono.

I would like to thank all the current and former members of Sugiyama Laboratory who shared the time with me. The lively atmosphere and interactions made my long journey an enjoyable experience. Thank you to Prof. Junya Honda, Prof. Takashi Ishida, Dr. Ikko Yamane, Dr. Tomoya Sakai, Masahiro Kato, Masahiro Fujisawa, Dr. Futoshi Futami, Dr. Nontawat Charoenphakdee, Hideaki Imamura, Jongyeong Lee, Han Bao, Liyuan Xu, Taira Tsuchiya, Dr. Yuko Kuroki, Yivan Zhang, Soma Yokoi, Kento Nozawa, Prof. Takayuki Osa, Dr. Gang Niu, Prof. Nobutaka Ito, Dr. Yoshihiro Nagano, Dr. Dongxian Wu, and Dr. Nan Lu. I am also indebted to Etsuko Yoshida, Yuko Kawashima, and Fumi Sato for their continuous support facilitating the lab activities.

I would like to thank RIKEN Junior Research Associate Program, Masason Foundation, Toyota-Dwango AI Scholarship, and the Japan Student Services Organization (JASSO) Scholarship Program. The opportunities and support have enabled me to complete my graduate studies.

Last but not least, I would like to thank my parents, my younger brother, and my grandparents for their love and continuous support. I am deeply grateful for Kiwa Nakajima, who has provided me with a true partnership full of trust and thoughtfulness. She has always been there for me in the most challenging times, and if it were not for her, I would not have been able to come this far.

# Related Publications

## Content

The following is the list of publications on which this dissertation is based.

### Chapter 3

1. Teshima, T., and Sugiyama, M., Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021)*.

### Chapter 4

1. Teshima, T., Sato, I., and Sugiyama, M., Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of 37th International Conference on Machine Learning (ICML 2020)*.

### Chapter 5

1. Teshima, T.*, Ishikawa, I.*, Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M., Coupling-based invertible neural networks are universal diffeomorphism approximators. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.[1]

2. Teshima, T., Tojo, K., Ikeda, M., Ishikawa, I., and Oono, K., Universal approximation property of neural ordinary differential equations. *arXiv:2012.02414 [cs.LG]*.

[1] * Equal contribution.

# Contents

**Abstract** i

**Acknowledgments** iii

**Related Publications** v

**List of Symbols and Abbreviations** xiii

**List of Figures** xv

**List of Tables** xvii

**1 Introduction** 1
- 1.1 Philosophy of Causality and Conceptual Research Question . . . 1
  - 1.1.1 Statistical Learning and Uniformity of Nature . . . 1
  - 1.1.2 Causality in Philosophy of Science . . . 2
  - 1.1.3 Statistical Frameworks of Causality . . . 5
  - 1.1.4 Conceptual Research Question . . . 7
- 1.2 Small-data Learning and Regularity of Causal Mechanisms . . . 8
  - 1.2.1 Statistical Machine Learning . . . 8
  - 1.2.2 Small-data Problems . . . 9
  - 1.2.3 Regularity of Causal Mechanisms . . . 9
- 1.3 Central Statement and Main Idea . . . 10
  - 1.3.1 Existing Attempts on Causality-informed Learning . . . 10
  - 1.3.2 This Dissertation: Scope and Main Idea . . . 11
- 1.4 Chapter Structure and Contributions . . . 12
  - 1.4.1 Division of the Problem . . . 12
  - 1.4.2 Chapter 3: When Graphical Causal Model is Known or Estimable: Causal-graph Data Augmentation . . . 12
  - 1.4.3 Chapter 4: When Structural Causal Model is Estimable: Causal Mechanism Transfer . . . 14
  - 1.4.4 Chapter 5: Theoretical Analysis of the Representation Power of Invertible Neural Networks . . . 15

**2 Preliminaries** 17
- 2.1 Introduction to Statistical Frameworks of Causal Inference . . . 17
  - 2.1.1 An Illustration . . . 17
  - 2.1.2 Statistical Frameworks of Causal Inference . . . 20
- 2.2 General Notation . . . 21
- 2.3 Structural Causal Framework . . . 23

2.3.1 Structural Causal Models (SCMs) . . . 23
2.3.2 Graphical Causal Models (GCMs) . . . 26
2.3.3 Relation between SCMs and GCMs . . . 28
2.4 Properties and Estimation of Structural Causal Framework . . . 32
2.4.1 Statistical Independences in GCMs . . . 32
2.4.2 Statistical Independences in SCMs . . . 33
2.5 Problem Setup and Approach . . . 36
2.5.1 Problem: Small-data Learning . . . 36
2.5.2 Problem Taxonomy: Two Estimable Causal Model Layers . . . 37
2.5.3 Approach: Data Augmentation . . . 38
2.6 Causal Machine Learning . . . 39
2.6.1 Causality for Machine Learning . . . 39
2.6.2 Causality by Machine Learning . . . 42
2.6.3 Causality in Machine Learning . . . 42
2.7 Conclusion . . . 43

**3 When Graphical Causal Model is Known or Estimable: Causal-graph Data Augmentation 45**
3.1 Overview . . . 45
3.1.1 Motivation . . . 45
3.1.2 Idea . . . 46
3.1.3 Contributions . . . 46
3.2 Problem Setup and Main Assumption . . . 47
3.2.1 Base Problem: Supervised Learning . . . 47
3.2.2 Main Assumption . . . 47
3.2.3 Problem Statement . . . 47
3.3 Proposed Method . . . 48
3.3.1 Overview of the Method . . . 48
3.3.2 Derivation of the Proposed Method . . . 48
3.3.3 Proposed Method: Causal-graph Data Augmentation . . . 49
3.3.4 Implementation Details . . . 50
3.4 Theoretical Analysis . . . 50
3.5 Experimental Evaluation . . . 53
3.5.1 Real-world Data Experiment . . . 53
3.5.2 Synthetic-data Experiment . . . 55
3.6 Related Work and Discussion . . . 57
3.6.1 GCMs and Predictive Modeling . . . 57
3.6.2 Causal Discovery and Transfer Learning . . . 58
3.6.3 GCMs and Efficient Estimation . . . 59
3.6.4 Cyclic CGs . . . 59
3.6.5 Users' Burden of Inputting CGs . . . 59
3.7 Conclusion . . . 60

**4 When Structural Causal Model is Estimable: Causal Mechanism Transfer 61**
4.1 Overview . . . 61
4.1.1 Motivation . . . 61
4.1.2 Idea . . . 62
4.1.3 Contributions . . . 63
4.2 Problem Setup and Main Assumption . . . 63

4.2.1 Base Problem: Few-shot Domain Adapting Regression . . . 64
4.2.2 Main Assumption . . . 64
4.2.3 Problem Statement . . . 64
4.3 Proposed Method . . . 65
4.3.1 Overview of the Method . . . 65
4.3.2 Proposed Method: Causal Mechanism Transfer . . . 65
4.3.3 Implementation Details Based on Invertible Neural Networks . . . 66
4.4 Theoretical Analysis . . . 67
4.4.1 Complete-estimation Case: Minimum Variance Property . . . 67
4.4.2 Incomplete-estimation Case: Excess Risk Bound . . . 68
4.4.3 Representation Power of Invertible Neural Networks . . . 69
4.5 Experimental Evaluation . . . 69
4.5.1 Design of the Experiment . . . 70
4.5.2 Details of the Experiment . . . 71
4.5.3 Experimental Results . . . 72
4.5.4 Synthetic-data Experiment . . . 72
4.6 Related Work and Discussion . . . 74
4.6.1 Existing Transfer Assumptions . . . 74
4.6.2 Causality for Transfer Learning . . . 75
4.6.3 Plausibility of the Assumptions . . . 76
4.7 Conclusion . . . 77

**5 Theoretical Analysis of the Representation Power of Invertible Neural Networks 79**
5.1 Overview . . . 79
5.1.1 Motivation . . . 79
5.1.2 Idea . . . 80
5.1.3 Contributions . . . 80
5.2 Problem Setup . . . 81
5.2.1 Definitions of Models . . . 81
5.2.2 Notions of Universality . . . 82
5.2.3 Goal . . . 83
5.3 Main Results . . . 83
5.3.1 General Result: Universality of Invertible Models . . . 83
5.3.2 Application to Specific Architectures . . . 84
5.4 Proof Outline . . . 85
5.4.1 Proof Outline for Theorem 5.1 . . . 85
5.4.2 Proof Outline for Theorem 5.2 . . . 87
5.5 Related Work and Discussion . . . 88
5.5.1 Normalizing Flows . . . 88
5.5.2 Other Invertible Neural Network Architectures . . . 88
5.5.3 The Strength of the Representation Power of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ . . . 88
5.6 Conclusion . . . 89

**6 Conclusion and Future Prospects 91**
6.1 Conclusion . . . 91
6.2 Future Prospects . . . 91
6.2.1 Reducing Computational Complexity . . . 92
6.2.2 Relaxing Model Assumptions . . . 92
6.2.3 Evaluating and Mitigating Model Misspecification . . . 92

6.2.4 Exploiting Other Aspects of SCMs . . . 93
6.2.5 Characterizing the Limitation of Causality-informed Learning . . . 94
6.2.6 Application to Other Statistical Inference Tasks . . . 94
6.2.7 More Detailed Levels of the Data-Generating Processes . . . 95
6.2.8 Exploiting the Potential Outcome Framework . . . 95
6.2.9 Extension to Continual Learning . . . 95

**A Appendices for Chapter 2** **97**
A.1 Example of Figure 2.1 . . . 97
A.2 Supplementary on Causal Models and Proofs . . . 98
A.2.1 Preparation from Probability Theory . . . 98
A.2.2 GCMs . . . 99
A.2.3 SCMs . . . 101
A.2.4 Compatibility of SCMs and GCMs . . . 102

**B Appendices for Chapter 3** **107**
B.1 Real-world Data Experiment Details . . . 107
B.1.1 Data Set Details . . . 107
B.1.2 Predictor Model Details . . . 109
B.1.3 Proposed Method Implementation Details . . . 109
B.1.4 Causal Discovery Method Configuration . . . 109
B.1.5 Supplementary Figures . . . 109
B.2 Synthetic-data Experiment Details . . . 110
B.3 Details and Proof of the Theoretical Analysis . . . 111
B.3.1 Notation and Problem Setup . . . 111
B.3.2 Main Theorem . . . 113
B.3.3 Comparison of Complexity Measures . . . 121
B.4 Computational Complexity of Algorithm 2 . . . 125

**C Appendices for Chapter 4** **127**
C.1 Nonlinear ICA . . . 128
C.2 Details of Real-world Data Experiment . . . 129
C.2.1 Dataset Details . . . 129
C.2.2 Model Details: Invertible Neural Networks . . . 129
C.2.3 Model Details: Penultimate Layer Networks . . . 129
C.2.4 Training Details . . . 129
C.2.5 Compared Methods Details . . . 129
C.3 Details of Synthetic-data Experiment . . . 130
C.3.1 Data-generating Process . . . 130
C.3.2 Proposed Method Configuration . . . 131
C.3.3 Evaluation . . . 131
C.3.4 Supplementary Figures . . . 131
C.4 Details and Proofs of Theorem 4.2 . . . 131
C.4.1 Notation . . . 132
C.4.2 Problem Setup . . . 133
C.4.3 Assumptions . . . 134
C.4.4 Theorem Statement . . . 135
C.4.5 V-statistic and U-statistic . . . 136
C.4.6 Proof of Pseudo Estimation Error Bound . . . 138

C.4.7 Proof of Approximation Error Bound . . . 140
C.4.8 Comparison of Rademacher Complexities . . . 145
C.4.9 Remark on Higher-order Sobolev Norms . . . 147
C.5 Details and Proofs of Theorem 4.1 . . . 148
C.6 Further Comparison with Related Work . . . 149
C.6.1 Comparison with Magliacane et al. [176] . . . 149
C.6.2 Comparison with Gong et al. [90] . . . 150
C.6.3 Comparison with Arjovsky et al. [8] . . . 151

**D Appendices for Chapter 5** **153**
D.1 Proof of Lemma 5.1: From $L^p$-universality to Distributional Universality . . . 153
D.2 Proof of Theorem 5.1: Equivalence of Universality Properties . . . 156
D.2.1 From $\mathcal{D}^2$ to $\mathrm{Diff}_{\mathrm{c}}^2$ . . . 157
D.2.2 From $\mathrm{Diff}_{\mathrm{c}}^2$ to $\mathcal{S}_{\mathrm{c}}^\infty$ and Permutations . . . 157
D.3 Properties of Diffeomorphisms on $\mathbb{R}^d$: From $\mathrm{Diff}_{\mathrm{c}}^2$ to Nearly-Id . . . 160
D.4 Proof of Theorem 5.2: $L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ . . . 162
D.4.1 Approximation of General Elements of $\mathcal{S}_{\mathrm{c}}^0$ . . . 162
D.4.2 Special Case: Approximation of Coordinate-wise Independent Transformation 164
D.5 Locally Bounded Maps and Piecewise Diffeomorphisms . . . 165
D.5.1 Definition of Locally Bounded Maps . . . 165
D.5.2 Definition and Properties of Piecewise $C^1$-maps . . . 165
D.6 Compatibility of Approximation and Composition . . . 168
D.7 Examples of Flow Architectures Covered in Chapter 5 . . . 170
D.7.1 Neural Autoregressive Flows (NAFs) . . . 170
D.7.2 Sum-of-squares Polynomial Flows (SoS flows) . . . 174
D.8 Using Permutation Matrices Instead of Aff in the Definition of $\mathrm{INN}_{\mathcal{G}}$ . . . 175
D.9 Other Related Work . . . 176

**Bibliography** **177**

# List of Symbols and Abbreviations

**Abbreviations**

| | |
|---|---|
| ADMG | Acyclic Directed Mixed Graph |
| CDA | Causal-graph Data Augmentation |
| CG | Causal Graph |
| CMT | Causal Mechanism Transfer |
| DAG | Directed Acyclic Graph |
| e.g. | for example |
| ERM | Empirical Risk Minimization |
| GCL | Generalized Contrastive Learning |
| GCM | Graphical Causal Model |
| i.e. | that is |
| ICA | Independent Component Analysis |
| ICM | Independent Component Model |
| NLICA | NonLinear Independent Component Analysis |
| ODE | Ordinary Differential Equation |
| PAG | Partial Ancestral Graph |
| RSE | Reduced-form Structural Equation |
| RSF | Reduced-form Structural Function |
| s.t. | such that |
| SCM | Structural Causal Model |
| SE | Structural Equation |
| SEM | Structural Equation Model |
| SF | Structural Function |

**Basic**

| | |
|---|---|
| $[i]$ | the set $\{1, 2, \dots, i\}$ |
| $\lvert\cdot\rvert$ | cardinality |
| $\mathcal{O}(\cdot)$ | order of computational complexity |
| $o(\cdot)$ | small-order of computational complexity |
| $\coprod$ | disjoint union |
| $[\cdot]$ | natural projection to quotient space |
| $\overset{\text{i.i.d.}}{\sim}$ | independent and identically distributed |
| $\overset{\text{i.i.g.}}{\Leftarrow}$ | independent and identically generated |
| $\mathbb{1}$ | indicator function |
| $\boldsymbol{a}^S$ | subvector $(a^{s_1}, \dots, a^{s_m})$ $(\boldsymbol{a} = (a^1, \cdots, a^d)^\top, S = \{s_1, \dots, s_m\} \subset [d])$ |
| $\top$ | transpose |

**Graphs**

| | |
|---|---|
| desc | descendants |
| dis | district |
| $\overline{\text{pa}}$ | generalized parents |
| mb | Markov blanket |
| mp | Markov pillow |

non-desc non-descendants

pa parents

pred Predecessors

$\mathfrak{B}$ bi-directed edges

$\mathfrak{D}$ directed edges

$\mathfrak{V}$ vertices

**Learning**

$d$ dimensionality of data

$K$ number of domains

$n$ sample size

$\mathbb{P}_{\boldsymbol{Z}}$ joint distribution of input data $X$ and label $Y$

$\mathcal{D}$ data set

$\boldsymbol{Z} = \mathcal{X} \times \mathcal{Y}$ joint space of labeled data $(X, Y)$

$\hat{\mathcal{R}}_{\mathrm{ERM}}$ empirical risk

$\ell$ loss function

$\mathcal{H}$ predictor hypothesis class

$p_{\boldsymbol{Z}}$ probability density of data

$\mathcal{R}$ risk functional

$Z = (X, Y)$ labeled data

**Number sets**

$\mathbb{E}$ expectation operator

$\emptyset$ empty set

$\hat{\mathbb{E}}$ empirical average operator

$\mathbb{N}$ positive integers

$\mathbb{R}_{\geq 0}$ non-negative reals

$\mathbb{N}_0$ non-negative integers

$\mathbb{R}_{>0}$ positive reals

$\mathbb{R}$ real numbers

$\Omega$ sample space of a probability space

$\mathbb{Z}$ integers

# List of Figures

1.1 Philosophical theories of causality (not exhaustive). . . . . 2
1.2 Illustrations of the two philosophical concepts of causation: counterfactuals and mechanisms. . . . . 4
1.3 Organization of the chapters. . . . . 13
1.4 Problem setup of Chapter 3. . . . . 13
1.5 Problem setup of Chapter 4. . . . . 14
1.6 Universal approximation property of function models. . . . . 15

2.1 Example: same distribution from different random variables . . . . . 18
2.2 Distributions of $(X, Y)$ after intervention . . . . . 19
2.3 Examples of structural equations. . . . . 26
2.4 Relations among SCMs and GCMs. . . . . 28
2.5 Constraints imposed on semi-Markovian GCMs. . . . . 33
2.6 General idea of this dissertation . . . . . 38

3.1 Visualization of the idea in trivariate case . . . . . 46
3.2 Illustration of the proposed method . . . . . 49
3.3 Probability tree of the proposed method . . . . . 51
3.4 Reference CGs for the data sets . . . . . 54
3.5 Experimental results . . . . . 55
3.6 Results of synthetic-data experiments . . . . . 56

4.1 Independent component model . . . . . 62
4.2 Assumption of common generative mechanism . . . . . 63
4.3 Schematic illustration of proposed method . . . . . 67
4.4 Illustration of the proposed method . . . . . 68
4.5 Fitting the inflated data . . . . . 69
4.6 Results of synthetic data experiments . . . . . 74

5.1 Illustration of coupling-based flow layers . . . . . 81
5.2 Decomposition of nearly-Id transformations . . . . . 86
5.3 Illustration of the proof for the $L^p$-universality . . . . . 87

B.1 Average relative improvement in percentage . . . . . 110
B.2 Ground-truth CGs of the synthetic data sets . . . . . 111

C.1 Results of synthetic data experiments . . . . . 132

D.1 Outline of propositions and lemmas . . . . . 154

# List of Tables

1.1 Division of the problem and the corresponding chapters. . . . . 12

2.1 Terminology of structural causal models. . . . . 23

3.1 Summary of data sets . . . . . 54
3.2 Summary of synthetic data sets . . . . . 56

4.1 Results of the real-world data experiments . . . . . 73
4.2 Comparison of transfer assumptions for domain adaptation . . . . . 75

5.1 CF-INN architectures analyzed in this chapter . . . . . 84

6.1 Three levels of model misspecification in causality-informed machine learning. . . . . 93

B.1 List of abbreviations and symbols used in the chapter. . . . . 108

C.1 List of abbreviations and symbols used in the chapter. . . . . 127

D.1 List of abbreviations and symbols used in the chapter. . . . . 153

# Chapter 1

# Introduction

The quest for *causality* is inherent in humans' cognitive behavior. If causal concepts are so important to us, should artificial intelligence internalize them as well? In particular, is causal knowledge useful when manipulation or explanation is *not* the goal, and when the goal is to *make accurate predictions*? Prompted by these questions, this dissertation presents the author's attempts to establish concrete *causality-informed machine learning* methods. In this introductory chapter, let us begin by framing a conceptual research question in the language of the philosophy of sciences (Section 1.1). We then continue by introducing the general idea of this dissertation (Section 1.2), phrasing the central statement (Section 1.3), and explaining the outline of the subsequent chapters (Section 1.4).

## 1.1 Philosophy of Causality and Conceptual Research Question

The twenty-first century has seen a rapid and widespread development of automated intelligent systems, such as computer-aided diagnosis [229], object detection [170], speech recognition [306], and automated translation [149, 128]. Many of such systems are powered by *statistical machine learning*, a paradigm in which various methods for learning from data have been developed.

### 1.1.1 Statistical Learning and Uniformity of Nature

The designs of statistical machine learning parallel *inductive inference* in logical reasoning: learning from observations — the *training data* — and drawing conclusions about the unobserved — the *testing instances*. As the famous philosopher David Hume saw through, logical inductive reasoning requires certain "uniformity of nature" for it to be useful (Hume [118, IV.II.32], Salmon [228], Henderson [102]). He further argued that such a uniformity principle is not derived from within the *relations of ideas*, i.e., pure a priori logic, but that our "Custom or Habit" is at the basis of such a principle (Hume [118, V.I.35-36], Morris and Brown [189]).

In statistical machine learning, or more generally in statistics, analogues of such a *uniformity principle* are embedded in the formulation in one form or another, such as an assumption that the data is an *independent and identically distributed* (i.i.d.) sample from a probability distribution (e.g., [184]) or that one should be a rational decision-maker (e.g., [22]). Of course, none of such assumptions is a priori justified; the justification of such assumptions would have to invoke an inductive argument (or its analogue, statistical induction), falling into a circularity. Nevertheless, corresponding to each of such different premises, reasonable inferential rules — the *learning methods* — have been derived (e.g., [22, 279]). In other words, devising different assumptions have led to the development of a variety of learning algorithms that are *conditionally reasonable* given the premise.

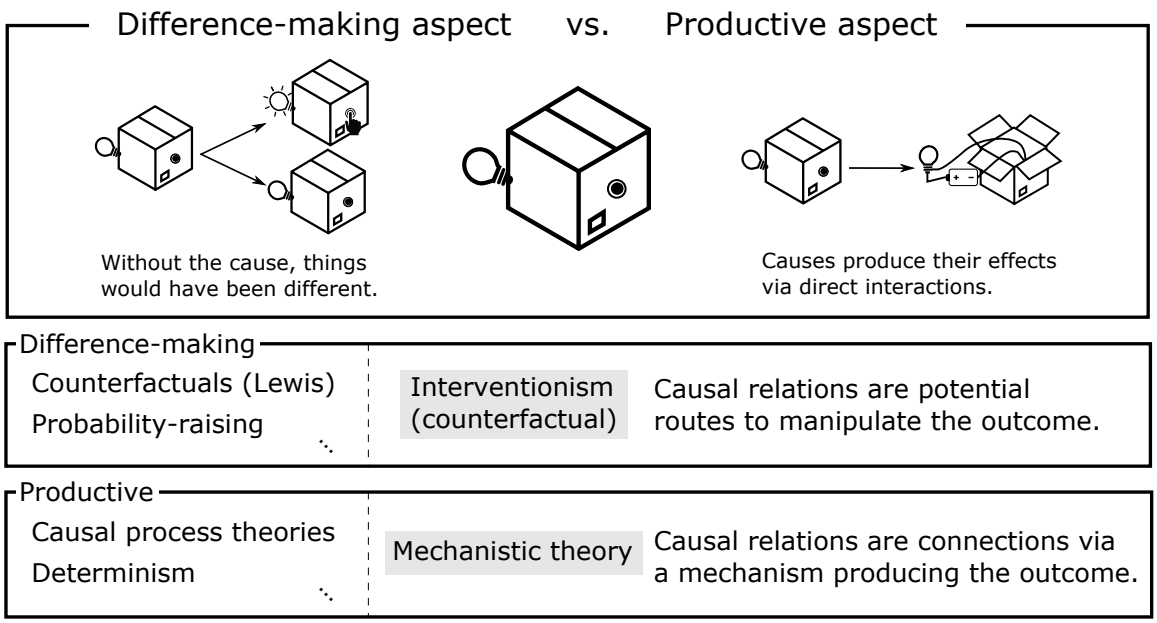



Figure 1.1: Philosophical theories of causality (not exhaustive).

In this dissertation, we discuss the usage of causal information for statistical machine learning (Section 1.3). Our discussion is based on statistical causal models: the concrete methodology built on top of probability theory that materializes the philosophical concepts of causality. Before we turn to the concrete frameworks of statistical causal modeling, let us briefly review the philosophical theories on the concept of causality.

### 1.1.2 Causality in Philosophy of Science

*Causality* is a form of such a *uniformity principle*, which is distinctive in humans' perception of the world (Hume [118, V.I.35-36], De Pierris and Friedman [60]). Unfortunately, the very *definition* of causation has never seen a complete consensus in the philosophy of sciences, not to mention its ontology ("does it exist?") and its epistemology ("how can we recognize it?"). Thus, it is beyond the capacity of the author to revisit all such competing theories here (for more comprehensive accounts, see, e.g., Beebee et al. [17], Mumford and Anjum [190], and Kutach [154]). Instead, let us review two key concepts: *interventionism* and *mechanisms*, the two philosophical notions that play central roles in the contemporary *statistical frameworks of causal modeling* (Hitchcock [105]).

To organize the philosophical theories of causation, the dichotomy of "*difference-making* versus *production*" is useful (Kutach [154, p.13]). Some theories of causation emphasize that causality is about *difference-making*: without the cause, the effect would not have happened. Others emphasize that causality is about *production*: causes *bring about* their effects. Between the two, the interventionistic theory of causation is more focused on the difference-making aspect, whereas the mechanistic theory has more emphasis on the productive aspect of causality.[1] See Figure 1.1 for an overview.

**Interventionistic theory of causation.** The basic idea of counterfactual theories of causation is that the causal claims can be explained in terms of statements about counterfactual events such as "if event c had occurred, event e would have occurred" (Menzies and Beebee [182]). It is an

[1] Other theories attending to the difference-making aspect include the *probability-raising* theories (e.g., [154, Chapter 6]), and others attending to the productive aspect include the *causal process theories* and the *determinism* (e.g., [154, Chapters 3 and 5]).

approach to characterize causality based on our intuition that causality has a *difference-making* nature [154]; as Lewis [165] put it, "[human beings] think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it." The best-known counterfactual analysis of causation is due to Lewis [165] (Menzies and Beebee [182]).

In particular, *interventionism* is a prominent version of the counterfactual theories with a focus on the difference-making about what *potential outcome* is realized by an intervention,[2] and it is arguably the most prevalent in the literature of statistical causal models (Woodward [294]). The central idea of interventionism is that causal relations are potential routes by which the world can be manipulated or controlled. Thus, it asserts that the most relevant counterfactuals are those that describe the altered behavior of variables under interventions that change the value of another (Woodward [294, p.15]). In other words, interventionism (and its precursor known as *manipulationism*) materialized the central theme that one of the main reasons we care about causation is that causes are often means by which we can bring about effects or make them more likely (Kutach [154, p.138]).

For example, consider a box with a button and a light bulb (Figure 1.1). If we press the button, the bulb will glow. If we do not press the button, the bulb will not glow. In the interventionistic view, A (e.g., a button being pressed) is a cause of B (e.g., a light) if an *intervention* in A (e.g., pressing the button) results in a difference in B (e.g., whether the bulb glows). The existence of a connection between A and B is technically not required: in this view, it does not matter if the box is empty and any connection between the button and the bulb is missing [154].

**Mechanistic theory of causation.** The *mechanistic* view of causality emphasizes the *productive* aspect of causality [175, 154, 83]. As Glennan [86] put it, "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations." In this view, "events are causally related when there is a mechanism that connects them" (Glennan [84]).

Let us recall the light bulb example (Figure 1.1). The mechanistic theory of causality considers A (e.g., a button) to be a cause of B (e.g., a glowing bulb) when there is a mechanism from A to B that *produces* B via direct interactions of some parts (e.g., the button switches the circuit connecting a battery to the light bulb, the battery starts an electric current, and the electrical energy is converted to light energy). In this view, what is inside the box (e.g., a circuit and a battery) is important [154]. Of course, what counts as a *direct* interaction or a *part* depends on the *level* of the analysis (e.g., Kutach [154, p.53]), and the parts as well as the interactions may consist of lower-level ones (e.g., how a battery produces electric current). Thus, mechanisms usually have black-box components which themselves are causal in nature. In this sense, the mechanistic account of causality is not completely reductive (Craver and Tabery [54, 2.3.2]). Nevertheless, mechanisms do characterize causal regularities in more detail than merely citing causes and effects (Kutach [154, p.60]).

The philosophy of *mechanisms* emerged around the turn of the twenty-first century reflecting a synthesis of the philosophy and the history of science [54]. What sparked its emergence was the observation that many significant discoveries in science involved the discoveries and descriptions of mechanisms (e.g., Machamer et al. [175], Glennan [83], and Craver and Tabery [54]) and that elucidating a mechanism has been a gold standard for explanations in scientific practice (Glennan [83]).

[2] Interventionism is a type of counterfactual theory because, to define and understand the "difference" due to an intervention, one needs to (at least conceptually) compare the result under intervention (the factual) with *what would have been the result had the intervention not taken place* (the counterfactual).

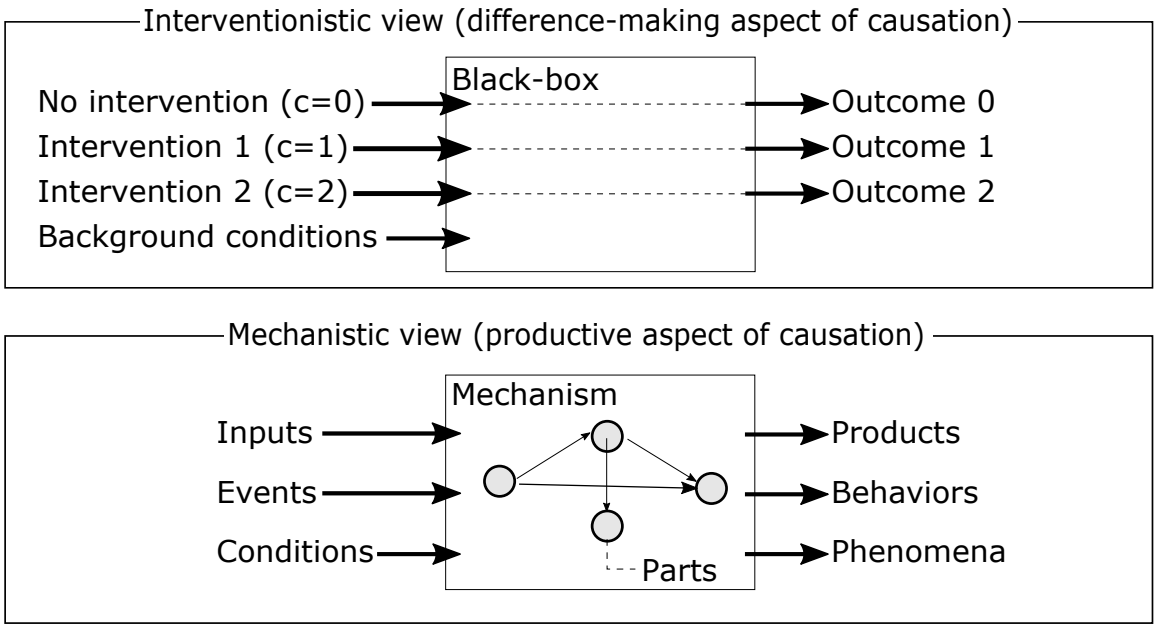



Figure 1.2: Illustrations of the two philosophical concepts of causation: counterfactuals and mechanisms.

**Relation of the two theories.** The two theories are alike in many respects:

1. They are both understood to have the characteristic of being "invariant, change-relating generalizations" (Glennan [86] and Woodward [296]).

2. They are not required to be *exceptionless* or *non-local*: the qualities required of legitimate *laws of nature* in philosophical terms [86, 296]. The relevant generalizations may be conditional on the context, time, and/or location (Woodward [296]).[3] Yet, they are expected to be relatively stable or invariant enough to play the role of "causal laws" (Glennan [86] and Woodward [296]).[4]

3. They are ontologically silent, i.e., they maintain neither that causation is part of the fundamental reality waiting to be discovered (in which case it is called a *natural kind* [154, p.58]) nor that it is merely our useful fiction for understanding the world (in which case it is called an *artificial kind* [154, p.59]). The two theories can be maintained as long as there is enough structure in fundamental reality to vindicate the reasonableness of the counterfactuals or the mechanisms (Kutach [154, pp.60,71]).

4. They are *non-reductionistic*; they have not been successful in reducing their causal notions to other fundamental concepts that are not causal. The concept of mechanism ineliminably contains a causal element (Craver and Tabery [54, 2.3.2]).[5] Interventionism uses the concept of *manipulation* or *intervention* as a primitive of the arguments, but they are apparently causal concepts. Counterfactual accounts commonly struggle to determine the value of the counterfactual quantities. However, some authors such as Woodward [295, 1.7] explicitly defend non-reductionistic standpoints.

[3] Classical examples of such generalizations include Mendel's law in biology (Woodward [296] and Glennan [86]).
[4] Here, the term *generalization* is used in contrast to the term "law" (as in *paradigmatic laws* of nature) that are commonly understood to be exceptionless and non-local in the philosophy of sciences (Woodward [296]).
[5] Indeed, if a causal relation is fundamental, then by definition, there is no mechanism for it (e.g., [54, 2.5.2] and [84]).

Thanks to the immediate connections to statistical causal frameworks, these are among the most prevalent philosophical stances in various fields of special sciences [126, 204, 295].

### 1.1.3 Statistical Frameworks of Causality

Along with the philosophical foundations (namely counterfactuals and mechanisms), the end of the 20th century saw the rapid development of statistical[6] causal frameworks (Hitchcock [106]). Such frameworks allow us to naturally formulate the causal quantities of interest, which could not be fully characterized within the conventional statistical frameworks.

**Potential outcome framework.** The *potential outcome framework* (POF) is a statistical causal framework pioneered by Splawa-Neyman [255] and developed by Rubin [226] and Robins [221], among others. In the formulation of the POF, counterfactuals play a central role (Hitchcock [105, 4.10]) as it employs *counterfactual random variables* (also known as *potential outcomes*) as primitives (Imbens and Rubin [126]).

The basic idea of the POF is to introduce certain random variables called the *potential outcomes*, which represent the outcomes that we would observe in the *counterfactual* scenario (as well as those of the factual scenario, which we can observe). For example, consider a *treatment* variable[7] $t$ (0: "no medication", 1: "take aspirin") and an outcome variable $Y$ (0: "headache persisted", 1: "headache cured"). If a person is given a treatment $t = 1$, and $Y = 1$ was observed, then the factual scenario (which we actually observed) is $(t = 1, Y = 1)$. Then, in the POF, we introduce the *potential outcome* variables $Y_{t=1}$ and $Y_{t=0}$, which are random variables representing the outcome given $t = 1$ and $t = 0$ (in this example, corresponding to the factual and the counterfactual scenarios), respectively. These random variables are used to model the causal quantities of interest; for instance, a causal effect can be defined as the outcome differences under different intervention states, e.g., $Y_{t=1} - Y_{t=0}$.

The framework is strongly tied with *interventionism* since the potential outcomes are usually introduced for some treatment variable. The POF was developed in the fields where the primary goal is to predict or estimate the results of interventions, e.g., medicine, epidemiology, econometrics, and political science [126]. For a history of the potential outcome approach to causal inference, please refer to Imbens and Rubin [126, Chapter 2].

In this framework, various practical conditions and methods for estimating such causal estimands have been developed: weak/strong ignorability (e.g., Hernán and Robins [103, Chapter 3]), propensity score methods (e.g., Hernán and Robins [103, Chapter 15]), instrumental variables (e.g., Hernán and Robins [103, Chapter 16]), regression discontinuity designs (e.g., Abadie and Cattaneo [1, Section 7]), and difference-in-differences and synthetic controls (e.g., Abadie and Cattaneo [1, Section 5]).

**Structural causal framework.** The structural causal framework (SCF) is a statistical causal framework developed by Pearl [204] and Spirtes et al. [253], among others. This framework attends more to the concept of mechanism (Menzies [181]; see also Woodward [295], Pearl [204], and Spirtes et al. [253], and Glennan [84]), where deterministic functions are used to represent the causal mechanisms (Hitchcock [105]).

The basic idea is to introduce certain deterministic functions that represent *causal mechanisms*, and to use them as so-called *structural equations* to describe how each random variable depends on its immediate causal predecessors. The machinery of the SCF will be elaborated in Chapter 2 because it is the main framework that the ideas of this dissertation are based on.

[6] By "statistical," in this dissertation, we indicate that the model is built on top of (measure-theoretical) probability theory.

[7] Interventions are also called *exposure* or *treatment*, depending on the context (e.g., [126, pp.4,19]).

The framework is also tied with the *interventionist* view of counterfactuals by some authors, e.g., Woodward [296] and Woodward [295] (Craver and Tabery [54, 2.3.4], Glennan [86] and Craver [55]). The central commitment of this view is that models of mechanisms describe variables that make a difference to the values of other variables in the model and to the phenomenon (Craver and Tabery [54, 2.3.4]).

In this framework, various conditions and methods for estimating (partial knowledge of) the causal mechanisms have been developed: constraint-based causal discovery (e.g., Glymour et al. [88]), score-based causal discovery (e.g., [88, 114]), function-based causal discovery (e.g., [239, 209, 210]), among others (e.g., [187, 133, 132]). Under the SCF, after such knowledge of causal mechanisms has been defined and estimated, it can be used to perform causal inference such as predicting the results of interventions [204].

**Common philosophical stances.** The following are some additional detailed philosophical stances that are shared by the two frameworks. In both frameworks, the models are introduced as potentially useful devices for expressing the causality behind the data that we observe in some domains of limited scope and not as the models that can represent all causal laws of nature having an unlimited scope. It is sensible that different application domains have different suited concepts of causality, as discussed by Woodward [296]. Epistemologically, they primarily support *type-level* causal (or *general causation* [106]) claims as opposed to *singular* causal (or *token-level* causal or *actual causation* [106]) claims (Woodward [294, Section 2.7], Hitchcock [106]). This focus on the *type-level* causality may be considered to be in part due to the "Fundamental Problem of Causal Inference" (Holland [109, p.947]) i.e., it is impossible to observe the counterfactual, in both cases of the POF and the SCF. The singular causal claims are made only as secondary derivatives of the estimated causal models ([97, 126]). This is in stark contrast to how philosophical frameworks of counterfactual approaches to causation have been devoted to analyzing singular causation (Hitchcock [106]).

**Interrelation of the two frameworks.** Some authors prefer to view the POF as a convenient notation system derived from the SCF (Pearl [202]), where the truth conditions of the counterfactuals are delivered by the structural equations. Some others (Hitchcock [107], Woodward [295], and Menzies and Beebee [182]) regard the SCF to be a concise way to declare the value assignments of some basic counterfactual quantities. In this sense, the two frameworks are reciprocal to each other (see, e.g., Menzies and Beebee [182, Section 5.2]), and they are practically selected depending on the application field.

In some fields, one framework may be preferred to the other, partly because of the difference in the conceptual emphasis: the POF attends more to the counterfactual theory and the SCF to the mechanistic theory. As a result, the two frameworks tend to focus on different goals; the SCF has a tendency to aim at building a large model incorporating all relevant variables, and it seems to be more often employed where the primary interest is in understanding a complex mechanism (e.g., biology [88, Section 7], ecosystem study [262], ecological studies [77], psychiatry [171, 21], clinical epidemiology [140, 21]). On the other hand, the POF tends to aim at imposing the minimum assumptions (e.g., [178, Introduction]), [103, Technical Point 6.2] required for estimating the specific quantity of interest (e.g., intervention effects), and it seems to be more prevalently employed where the primary interest is in estimating the effect of interventions rather than understanding the mechanism (e.g., program evaluation in econometrics [1] and evidence-based decision making in medicine and public policy [126]). Such a viewpoint is also evident in the following quote from Imbens and Rubin [126]: "In our own work, perhaps influenced by the type of examples arising in social and medical sciences, we have not found this [structural causal framework] approach to aid drawing of causal inferences [...]" In this dissertation, we do not attempt to argue which framework better serves which purpose. However, as we explain in Section 1.1.4, our focus is regarding the utility of

the knowledge of the detailed mechanisms, this dissertation is largely based on the SCF.

### 1.1.4 Conceptual Research Question

The pragmatic utility of knowing counterfactual (in particular, interventionistic) quantities, such as average treatment effects, is somewhat self-explanatory: it can be used for interventional decision making (e.g., [126, 103]). From such a viewpoint, differences in the detailed mechanisms make no difference as long as they satisfy the abstracted conditions such as *strong ignorability* (e.g., [103, Section 8.2]).

On the other hand, our intellectual curiosity seems to go even beyond the benefits of interventions; we often want to understand the world deeper, even when such an endeavor is not necessarily tied with interventional decision making. Much of the practice of contemporary science is driven by the search for mechanisms, and many of the grand achievements in the history of science are discoveries of mechanisms, particularly in special sciences [78] such as biology, neuroscience, and psychology (Craver and Tabery [54, Section 1]). Human beings seem to have an instinctive urge to understand how the world works.

Then, what is the pragmatic utility of finding out the details of the underlying mechanism? If the estimation of counterfactual quantities is the sole target of such an endeavor, the approach to abstract away from the detailed situation may be preferred by virtue of being ontologically parsimonious (e.g., Ockham's razor [10]). Attempts to formulate or estimate further details of an underlying mechanism may be unnecessary or should be avoided in the interest of ontological parsimony.

Thus, our conceptual question is the following: *is there some pragmatic utility in the knowledge of mechanisms, putting aside the estimation of counterfactual/interventionistic quantities?* Mechanisms have been a tool prevalent in many scientific fields used to organize the findings and communicate the knowledge (e.g., [269, Table 7.2]), even when such mechanistic explanations are not immediately tied with specific ideas of interventions. Then, what could be the pragmatic motivations of discovering mechanisms?

In this dissertation, we argue that the knowledge of the mechanisms estimated in the SCF can be pragmatically beneficial for statistical machine learning, *if* appropriate methods are designed to incorporate such knowledge into the learning procedures. We show, by providing concrete methods and examples, that an estimated structural causal model can be used to facilitate machine learning.[8]

Our intuitive argument is as follows. One reason why causal knowledge can be useful is that it is the knowledge of the *data-generating process*, i.e., the mechanisms that resulted in the events we observe, or in the context of statistics, the process through which the realized values of random variables are generated. Such knowledge of the generating processes of random variables could be as useful as the data themselves, in statistics or statistical machine learning, where *statistical induction* from data is embedded as an operating principle.

More concretely, we discuss how causal knowledge can be beneficial in the problem of learning from small data, as introduced in the next section. Importantly, causal mechanisms are believed to be *stable* or *invariant*, as opposed to the case of merely accidental generalizations (Woodward [295, p.15], Glennan [83], Cartwright [35]). This constitutes the conceptual reason why causality makes an interesting candidate of the regularity of nature to presuppose, estimate, and exploit for machine learning. A mechanism is believed to be typically stable in the absence of an intervention[9] (Craver

[8] Of course, the presented results are independent of the ontology of causality. The results do not inform the philosophical dispute over the reality of causality, i.e., whether it is a natural kind or an artificial kind (such as our psychological *projection* to *regularly cojoined events*). However, this dissertation is intended to reinforce the pragmatic motivation to understand causal structures or the fundamentality of causality from a different viewpoint from interventionism.

[9] To put it more precisely, we only consider such stable mechanisms in this dissertation and claim that such *ephemeral mechanisms* (Craver and Tabery [54, Section 2.4.6]) that appear in historical sciences, such as archaeology,

and Tabery [54, 2.2]). For example, as Glennan [86] put it, "mechanisms [...] are systems consisting of relatively stable configurations of parts that give rise to robust behaviors which can be expressed by invariant generalizations," [86, S348] and that "[the parts of a mechanism] must have a relatively high degree of robustness or stability; that is, in the absence of interventions, their properties must remain relatively stable" [86, S344]. Especially when the data is scarce, such *stability* (or *invariance*) that makes the knowledge valid in a range of similar contexts would be useful.

## 1.2 Small-data Learning and Regularity of Causal Mechanisms

*Machine learning* refers to the automated detection of meaningful patterns in data [236], and it is arguably one of the most powerful contemporary approaches to realizing artificial intelligence. The process of automatically extracting useful patterns from data is called *learning* or *training*. It is especially effective in those application fields where useful patterns are so complex that a human programmer cannot provide an explicit specification of how information should be processed to perform an intellectual task [236].

### 1.2.1 Statistical Machine Learning

The central feature of machine learning is its use of *data*. Data are the physical representation of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means [276]. It is collected and stored in various formats, e.g., images, audio, and tabular formats. In this dissertation, we presume the data are stored in the tabular format, such as the user data in a company, the results of social surveys (e.g., [110]), or electronic health records [304].

*Statistical machine learning* extracts meaningful patterns based on statistical concepts (Vapnik [279]). A typical problem setup is called the *supervised learning* problem. In this setup, some paired data $\{(x_i, y_i)\}_i$ of input $x_i$ and its corresponding output $y_i$ is given, and the task is to predict the output value $y$ given a previously unseen input $x$, i.e., one that was not contained in the training data. One standard approach to such a task is to construct a *function*, which is called the *predictor* or the *hypothesis*, that takes $x$ as input and outputs a prediction $\hat{y}$. Analogously to the inductive inference in logical reasoning, supervised learning is a problem of drawing conclusions about unseen events from the previous experience (i.e., the training data). In parallel to how inductive inference requires a uniformity principle, one needs an expedient assumption that connects the training data and the unseen input-output pairs in order to rationalize the inference.

A prototypical approach to learning from data is the *empirical risk minimization* principle [279, 1.5]. The principle instructs that, in order to find a predictor with a small prediction error with respect to the data distribution (called the *risk*), one should find a predictor with a small prediction error on the training data (called the *empirical risk*). Its justification is based on the premise that the training data and the unseen input-output pairs are independent and identically distributed samples from some unknown probability distribution (the *data distribution*). The main rationale of statistical learning theory is, roughly speaking, based on the (uniform) *law of large numbers*: if the training data is abundant, the empirical risk will provide an accurate estimate of the expectation of the risk, and hence the predictor learned by empirical risk minimization would yield a small risk.

history, and evolutionary biology (Glennan [85],Craver and Tabery [54, 2.4.6]) would require a different account.

### 1.2.2 Small-data Problems

The amount of data it takes for the empirical risk to be a good estimate of the risk depends on the predictors' model class, called the *hypothesis class*, as well as the data distribution. When the hypothesis class is complex and the data is limited, the (uniform) law of large numbers is rather helpless due to the scarcity of data, and the learned predictor typically fails to produce accurate predictions even if it has a small empirical risk. Such a phenomenon is called *overfitting* ([279, p.124]). Although it is difficult to estimate the amount of data required for learning an accurate predictor, as a general guide, the more data we have, the more complex hypothesis class may be learned without the fear of overfitting [236, p.21], and hence one may be able to learn a complex predictor that exploits more complex patterns, leading to improved accuracy of the predictions.

When data is limited in quantity, in general, we need an additional source of information to complement the knowledge that can be extracted solely from the data (e.g., [236, p.21, Chapter 5]). In the context of statistical machine learning, such additional information is referred to as *prior knowledge*. For example, *regularization* techniques (see, e.g., Shalev-Shwartz and Ben-David [236, Chapter 13])introduce prior knowledge that essentially restricts $\mathcal{H}$ to be "small," thereby reducing the risk of overfitting.

Learning from small data, despite the rapid progress in the methodology, remains an important challenge in many potential application fields of machine learning, such as social sciences (e.g., [110]). In such fields, data size is often limited due to costly data acquisition methods such as in-person surveys (e.g., [110, pp.46,72,68,139,143,195]).

### 1.2.3 Regularity of Causal Mechanisms

What kind of prior knowledge could we exploit for statistical learning when we have only small data? In this dissertation, we discuss the possibility of exploiting the knowledge of causality, or more precisely, that of the data-generating mechanism, to tackle the small-data learning problem.

Causal knowledge is deemed particularly useful for small-data learning because causality is believed to be relatively stable or invariant [86, 296]. For example, intuitively, one can consider incorporating the causal knowledge acquired thanks to domain experts' experiences, believing that the knowledge is still valid. Likewise, if we can obtain some knowledge about the mechanism of some system and apply the knowledge to another system having similar causal mechanisms, it can be useful for learning from small data. One can alternatively think of these as a form of knowledge transfer, either from human to machine or from a learned machine learning model to another.

In particular, we focus on the *productive* aspect of causal concepts, and in particular, the mechanistic views, as opposed to the *difference-making* aspect or the *counterfactual* view. We focus on the knowledge of the data-generating mechanisms and demonstrate that causal knowledge is not only useful for counterfactual inference but also for making inferences in the *factual world*.

Intuitively, such usage is sensible; the knowledge of data-generating mechanisms is knowledge about the regularity of nature, and thus it may have the potential to serve as a form of a uniformity principle to be exploited in statistical learning. However, it does require the development of specialized techniques; standard machine learning methods are not accompanied by canonical ways to incorporate causal knowledge. In this dissertation, we establish such concrete methodologies for combining the causal knowledge with standard machine learning methods based on a unified approach called *data augmentation*.

## 1.3 Central Statement and Main Idea

Causal models, estimated or known a priori, contain various kinds of information about the data-generating process. As we introduced in Section 1.1.4, there are two major aspects of causal knowledge: the difference-making aspect and the productive aspect.

The most common use case of causal models is arguably the facilitation of causal inference, especially predicting the effects of interventions and estimating the counterfactual quantities (Hitchcock [106]). Otherwise, the qualitative information about the causality has been used by domain experts to deepen the understanding and to generate explanations. In this dissertation, we discuss how causal models can add another layer to the machine learning methodology to leverage the uniformity of nature.

### 1.3.1 Existing Attempts on Causality-informed Learning

In the existing literature, the idea of exploiting causality for machine learning (CML) has been discussed [208, 230, 233]. A majority of existing work on CML focuses on the difference-making aspect of causality. Such an aspect has been exploited to tackle the change-related challenges in machine learning problems, e.g., robustness against the change in the data distribution based on the notion of interventions [233, I.A.] and efficient adaptation to different environments by adapting few modules reflecting the modularity of causal mechanisms [233, I.B.]. See Peters et al. [208], Schölkopf [230], Schölkopf et al. [233], and Section 2.6 for further details.

On the other hand, many fundamental problem setups of machine learning do not necessarily involve change-related concepts. Learning in a fixed environment (i.e., unaltered data generating process) is one of the most standard problem setups of supervised machine learning [279, Chapter 1]. In such setups, the problem is not necessarily change-related, and one is concerned with making an inference in the *factual* world as opposed to the *counterfactual* world. Even in such problem setups, we can anticipate that the causal knowledge is generally useful because the productive aspect of causal knowledge describes the factual world and not only the counterfactual world. Therefore, it is desirable to develop the methods to enable incorporating such productive aspects of the causal knowledge into the learning procedure.

In a nutshell, the CML literature that pursues this line of research, as a whole, attempts to support the following statement.

**Broad Theme Statement**

*The productive aspect of causality, or in particular, the knowledge of causal mechanisms, can be exploited to enhance machine learning in the factual world, i.e., for those tasks that do not necessarily involve counterfactuals.*

Providing partial support to this statement is also a goal of the present dissertation. Much of the existing CML literature only provided intuition-based arguments to bridge between the causal knowledge and the methodologies [232, 313, 311, 91, 90, 223], while many others only considered how to leverage highly problem-specific structures or model-specific structures [231, 211, 156] (see Section 2.6). Thus, the idea of leveraging the productive aspect of the causal knowledge, despite its importance, remains largely unexplored when it comes to the methods with formal justifications and with a general scope of problems. Providing formal justifications is important as it opens the possibility of theoretical analyses, which is often a key to improving the transparency and understanding the behavior of the developed methodologies.

### 1.3.2 This Dissertation: Scope and Main Idea

In this work, we aim to develop general and formally justifiable methods to incorporate the productive aspects of causal knowledge into machine learning. In other words, we aim to contribute partial support to the Broad Theme Statement.

In providing the support to the Broad Theme Statement, we take a formal approach: in contrast to the existing work, we develop our methods based on the formal mathematical representations of causal knowledge and their implications rather than intuitive arguments. In particular, we focus on leveraging the *statistical independence* relations implied by the (partial) knowledge of the causal mechanisms estimated in the SCF (namely, the *causal graphs* and the *reduced-form structural functions*) because it is a particularly well-studied type of such a formal implication [253, 204, 217, 218, 72]. We review such implications in Chapter 2.

As the machine learning task, we focus on supervised learning for learning a predictor, arguably the most prevalent task among other possible machine-learning tasks that may potentially benefit from incorporating the causal knowledge [236]. It is a common understanding that the incorporation of prior knowledge (or the *inductive bias*) to bias the learning process is inevitable for the success of learning algorithms [236, p.21], and also that the commitment to use a stronger prior assumption generally makes the learning from data easier while making the learning process less flexible [236, p.21]. Incorporating causality as a form of uniformity of nature inevitably involves a stronger prior assumption about the data-generating process: we assume the existence of some underlying data-generating mechanism and leverage its properties to enhance the learning methods. Therefore, in the CML regime, we anticipate that the derived machine learning methods enjoy some enhanced properties, such as improved sample efficiency to enable small-data learning at the cost of introducing additional assumptions.

To summarize, the goal of the present dissertation is to provide affirmative evidence to support the following central statement.

**Central Statement**

*The knowledge of causal mechanisms estimated in the SCF can enhance machine learning in small-data problems. Specifically, the statistical independence relations implied by the causal models of the SCF can be exploited through data augmentation.*

The Central Statement provides partial support of the Broad Theme Statement.

As a generic methodology to incorporate the knowledge captured by statistical causal models into the process of machine learning, we propose to use *data augmentation* as an approach to exploit the statistical independence induced by the causal models.

Designing an algorithm that uses data as an interface has two major advantages. First, the derived methodologies will not depend on a specific machine learning model or a learning method, and hence they will be easier to combine with a wide range of machine learning methods. Second, the algorithms will be relatively easy to mathematically analyze based on statistical learning theory because the complication of the analysis tends to be confined into the dependency structure due to the data augmentation. Therefore, once the dependency structure is appropriately taken into account, standard theoretical devices (e.g., Rademacher complexity [184]) can often be employed. With the aid of such theories, we can obtain theoretical insights into the properties of the proposed methods and clarify the mechanism through which they may yield favorable results.

Table 1.1: Division of the problem and the corresponding chapters.

| | Known (K) | Estimable (E) | (Unknown) |
|---|---|---|---|
| (G) GCM | Chapter 3 | Chapter 3 | – (Non-causal learning) |
| (S) SCM | – (Unrealistic) | Chapters 4,5 | – (Non-causal learning) |

## 1.4 Chapter Structure and Contributions

In this section, we explain the chapter structure of this dissertation. This chapter and Chapter 2 introduce the conceptual and mathematical backgrounds of this dissertation. In Chapter 6, we summarize the overall conclusion of the dissertation and discuss further the possibilities of future research directions. The rest of the chapters are structured based on the following division of the problem.

### 1.4.1 Division of the Problem

We consider the situations in which we apply the proposed approach by dividing the situation into the following 4 cases, divided based on whether the causal model is known (K) or estimable (E) in relevant domains, and whether the considered causal model is a graphical causal model (G) or a structural causal model (S).

(G-K) The graphical causal model (GCM) is known from domain knowledge.

(G-E) The GCM is estimable in relevant problem domains.

(S-K) The structural causal model (SCM) is known: we consider this case to be unrealistic and discard this case since the detailed form of a structural model is usually not elucidated in application domains of interest.

(S-E) The SCM is estimable in relevant problem domains.

Table 1.1 summarizes this division of the problem. We consider the above 4 cases where we design the data augmentation method for each concrete problem setup of machine learning. We discuss the theoretical analysis of the proposed methods based on statistical learning theory, as well as experimental evaluations. Yet, we disregard the case (S-K) since the assumption that a structural causal model is known is not realistic.

The problem setups and contributions of the main chapters are summarized below towards the end of this chapter. Figure 1.3 visualizes the overall structure of this dissertation.

### 1.4.2 Chapter 3: When Graphical Causal Model is Known or Estimable: Causal-graph Data Augmentation

**Problem and Motivation.** In this chapter, we consider the case that a GCM is either known by domain knowledge or estimable from the data of a relevant domain and discuss how to leverage the knowledge (Figure 1.4). A causal graph (Remark 2.1) is a compact representation used in the SCF to describe the influence relations among random variables, e.g., which variable is a direct cause of which effect. It is a quantity derived from random variables having an underlying causal structure and captures an aspect of the behavior of the random variables, similarly to how a joint probability distribution is derived from random variables and describes their behavior. In particular, the joint distribution and the causal graph have quantitative relations: from a causal graph, one can read out certain *conditional independence* relations that the joint distribution should hold.

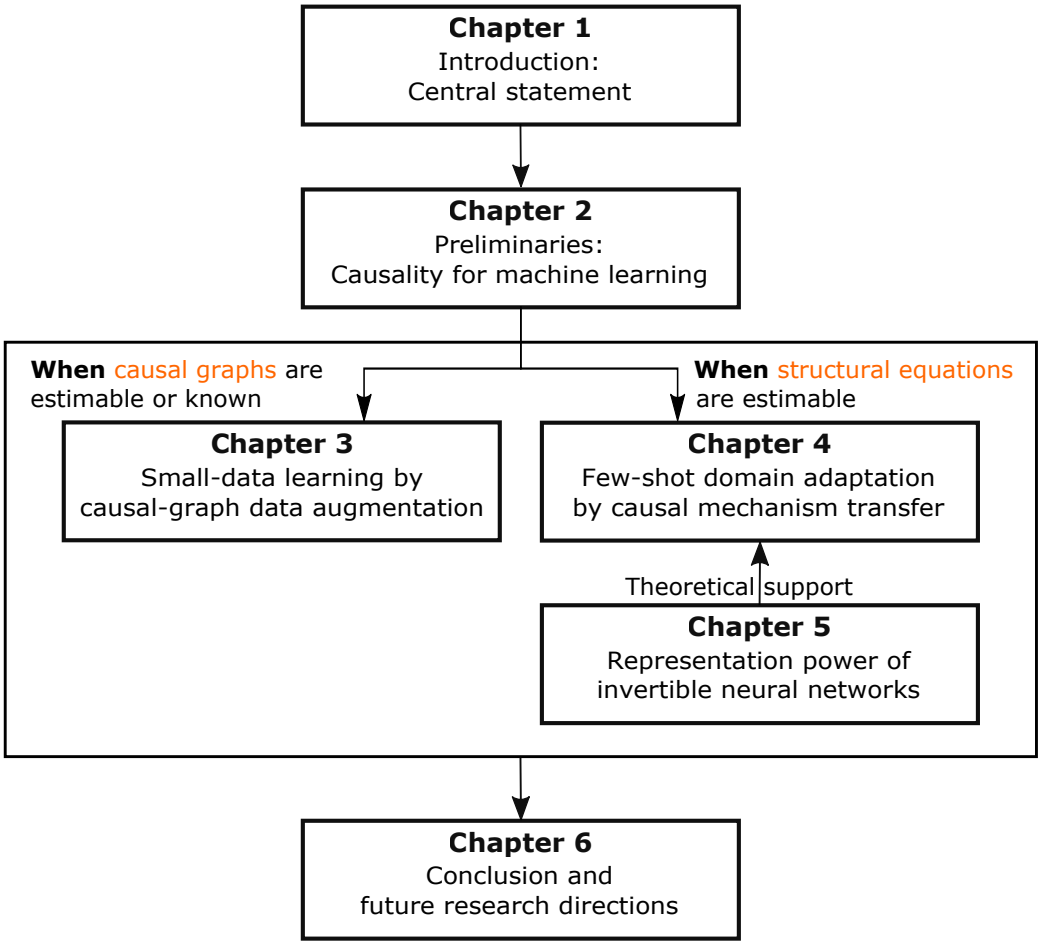



Figure 1.3: Organization of the chapters.

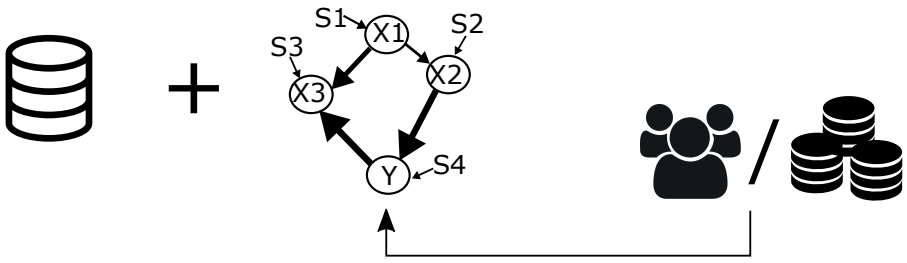



Figure 1.4: Problem setup of Chapter 3. We consider the case that a GCM is either known by domain knowledge or estimable from the data of a relevant domain.

In those application domains where data is scarce, e.g., due to costly data acquisition methods such as social surveys [110], it is essential to incorporate stronger prior knowledge into the learning process in order to enhance the data-efficiency [236, p.21]. If the knowledge of the causal graph is available, its implications to the conditional independence relations may provide effective prior knowledge to support the learning of predictors from small data.

Causal graphs may be obtained from the domain knowledge such as accumulated research on the subject matter [66, 227, 265] or the tacit knowledge of domain experts in some application domains. Indeed, thanks to their semantics as the description of direct causal relations, they have been used to communicate expert knowledge in various fields [292, 166, 103, 227, 66, 204, 117, 136].

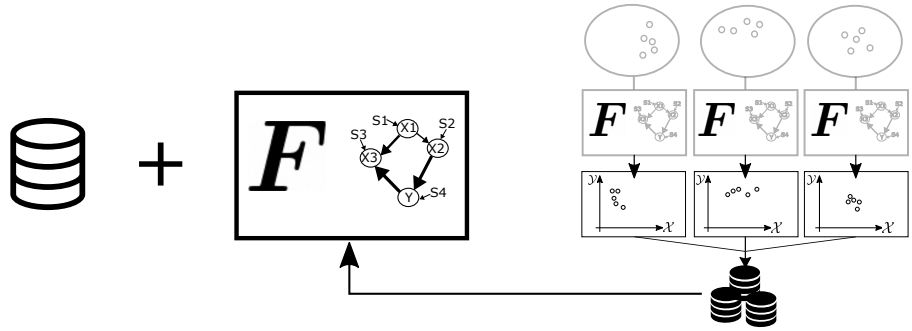



Figure 1.5: Problem setup of Chapter 4. We consider the case that an SCM is estimable from the data of a relevant domain sharing the same data-generating mechanism.

For examples of causal graphs provided by domain experts, see Figure 3.4. In case such domain knowledge is not readily available, causal discovery methods for estimating the causal graphs from data have been studied [88, 136, 25].

Thus, we will consider a problem that roughly corresponds to the following:

**Problem Sketch 1.1** (G-K and G-E)**.** *Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ be a data set sampled independently from some unknown causal structure, where $n$ is small (i.e., it is a* small-data *regime). Let $\hat{\mathcal{G}}$ be an estimated causal graph. Using the data as well as $\hat{\mathcal{G}}$, find a good predictor $h : \mathcal{X} \to \mathcal{Y}$.*

The problem is more precisely stated in Problem 2.1 of Section 2.5.2, and its slightly generalized version will be tackled in Chapter 3.

To exploit such relations, we design a data augmentation procedure that can be used to incorporate the knowledge encoded in the causal graph into statistical learning procedures (Definition 2.15).

**Contributions.** Our key contributions can be summarized in three points as follows.

1. We propose a method to augment data based on the prior knowledge expressed by a causal graph, assuming that an estimated causal graph is available.

2. We theoretically justify the proposed method via an excess risk bound based on the Rademacher complexity [15]. The bound indicates that the proposed method suppresses overfitting at the cost of introducing additional complexity and bias into the problem.

3. We empirically show that the proposed method yields consistent performance improvements, especially in the small-data regime, through experiments using real-world data with causal graphs obtained from the domain knowledge.

### 1.4.3 Chapter 4: When Structural Causal Model is Estimable: Causal Mechanism Transfer

**Problem and Motivation.** In this chapter, we consider the case that an SCM is estimable from the data of a relevant domain and discuss how to leverage the knowledge (Figure 1.5). More precisely, we consider a situation where we can estimate the *reduced-form structural function* (Reiss and Wolak [215]) of an SCM from the data in a relevant problem domain.

Such a situation is sensible in the applications where we can believe the existence of an underlying data-generating mechanism that is common across different domains, e.g., poverty nowcasting [36, 5, 177], economic studies in fragile countries [110], and health record analysis [304]. In such problem domains, it is often the case that data collection is costly, and hence developing the methodology of

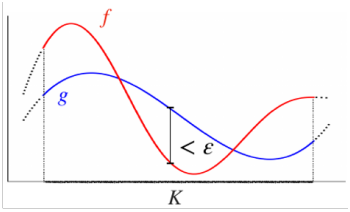



Figure 1.6: Universal approximation property of function models.

learning from relatively small data is highly important. For example, in medical record analysis for disease risk prediction, it can be reasonable to assume that there is a pathological mechanism that is common across regions or generations. If we can estimate and exploit the knowledge of such a hidden stable structure (in this case, the pathological mechanism), it can be useful for obtaining accurate predictors even if the data is scarce in the target domain (e.g., minor regions or demographic groups) by incorporating stronger prior knowledge about the data-generating process.

The problem corresponds to a *domain adaptation* problem, where we have only small data in a target domain of interest and have access to relatively large data from relevant problem domains. The central assumption in domain adaptation is the *transfer assumption* (TA) that specifies the relation between the target domain of interest and the *source* domains, i.e., the relevant problem domains from which the knowledge is transferred. Our assumption, namely the estimability of the causal mechanism from source domain data, gives rise to the novel transfer assumption of a *shared causal mechanism*, i.e., that the distributions are derived from SCMs whose structural equations are identical.

Thus, we will consider a problem that roughly corresponds to the following:

**Problem Sketch 1.2** (S-E)**.** *Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ be a data set independently generated by some unknown causal structure (called the* target domain*), where $n$ is small (i.e., it is a* small-data *regime). Let $\mathcal{D}'$ be another data set independently generated by a unknown causal structure (called the* source domain*). Assume that the target and source domains share the same causal mechanism. Identify some appropriate assumption with which the causal mechanism can be (partially) estimated from $\mathcal{D}'$, and using $\mathcal{D}$ and $\mathcal{D}'$, find a good predictor $h : \mathcal{X} \to \mathcal{Y}$.*

The problem is more precisely stated in Problem 2.2 of Section 2.5.2, and its slightly generalized version will be tackled in Chapter 4.

**Contributions.** Our key contributions can be summarized in three points as follows.

1. We formulate the flexible yet intuitively accessible TA of shared generative mechanism and develop a few-shot regression DA method (Section 4.3.2).
2. We theoretically justify the augmentation procedure by invoking the theory of generalized U-statistics [162].
3. We experimentally demonstrate the effectiveness of the proposed algorithm (Section 4.5). The real-world data we use is taken from the field of *econometrics*, for which structural equation models have been applied in previous studies [93].

### 1.4.4 Chapter 5: Theoretical Analysis of the Representation Power of Invertible Neural Networks

**Problem and Motivation.** In this chapter, we reinforce a theoretical justification of the methodology proposed in Chapter 4.

The method proposed in Chapter 4 uses a recently emerged technology called *coupling-flow-based invertible neural networks* (CF-INNs). It is a class of neural networks endowed with easy invertibility and the tractability of the Jacobian, and it has been widely used in various machine learning applications such as generative modeling [63, 145, 195, 143, 315], probabilistic inference [16, 288, 173], solving inverse problems [7], and feature extraction and manipulation [145, 192, 267].

Despite the growing popularity, due to the special architecture designs to maintain the invertibility, CF-INNs lacked a theoretical understanding. In particular, it was not clear whether CF-INNs, as function approximators (Figure 1.6), have sufficient representation power to approximate a wide range of invertible maps.

Chapter 5 sheds light on this problem by considering the following research question:

**Problem Sketch 1.3.** *How expressive is the set of invertible neural networks? Can they approximate diverse invertible maps?*

Since the result is itself an interesting contribution to the machine learning methodology in its own right and its scope is not limited to supporting the methodology of Chapter 4, we dedicate a chapter to introduce the results.

**Contributions.** Our contributions are summarized as follows.

1. We present a theorem to show the equivalence of universal approximation properties for certain classes of functions. The result enables the reduction of the task of proving the universality for general diffeomorphisms to that for much simpler coordinate-wise ones.

2. We leverage the result to show that some flow architectures, in particular even *affine coupling flows* that are the least expressive architectures among the ones appearing in this dissertation, can be used to construct a CF-INN with the universality for approximating a fairly general class of diffeomorphisms. This result can be seen as a convenient criterion to check the universality of a CF-INN: if the flow designs can reproduce the ACFs as a special case, it is universal.

3. As a corollary, we give an affirmative answer to a previously unsolved problem, namely the *distributional universality* [115, 131] of ACF-based CF-INNs.

**Theoretical implications to causal mechanism transfer.** The result of this chapter adds another layer to the theoretical guarantee of *causal mechanism transfer* (Algorithm 3), whose feasibility relies on the availability of a flexible model of invertible maps equipped with a tractable inverse.

# Chapter 2

# Preliminaries

Many questions in social and biomedical sciences are causal in nature (Imbens and Rubin [126]). In order to formulate and answer such questions, formal methods of quantitative causal inference have been developed since the 20th century [226, 253, 205, 204]. In this chapter, we review the formal treatment of the statistical causal frameworks. We mainly explain the *structural causal models* and their implications, which are to be exploited in the subsequent chapters. We also explain the general motivation of causal machine learning and provide a brief literature review. Towards the end of the chapter, we introduce the problem setup and the general approach of this dissertation. The readers who are familiar with the structural causal framework may well skip to Section 2.4, where we describe the properties that we exploit in this dissertation or directly go to Section 2.5, where we describe the problem setups and the approach of this dissertation.

## 2.1 Introduction to Statistical Frameworks of Causal Inference

In this section, we provide an intuitive introduction to the statistical frameworks of causal inference. Such frameworks have been developed in the previous half-century, corresponding to the demand in various *special sciences* such as econometrics, epidemiology, political science, and computer science, where an in-depth understanding of the mechanism or making interventions in a system is of major importance [253, 204, 126, 103].

### 2.1.1 An Illustration

Typical statistical analysis formulates the problem of interest in terms of probability distributions, and it is primarily concerned with the characteristics of the distributions (e.g., [289, Section 6.1]). Random variables, on the other hand, are usually treated as mere implementations of such distributions. Even though different random variables (as measurable maps) can yield the same distribution, they often need not be distinguished (e.g., [147, p.50]).

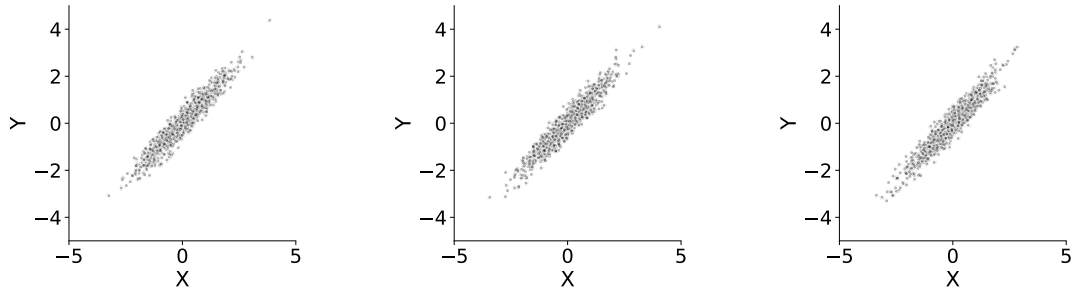
Causal inference frameworks [126, 253, 204] take into account what is called the *data-generating processes*, the additional structures in the random variables which are (in general) not captured by their joint distributions. By explicitly taking into account such additional structures, causal models enable a natural formulation of the quantities of interest called *causal estimands* [126], such as *interventional distributions* [204] or *average treatment effect* [127].

A simple example below demonstrates that the distinction of different random variables is important in certain situations, even if they have identical joint distributions (Figures 2.1 and 2.2).

```
## (a) X -> Y
X = normal()
Y = X + a * normal()
```

```
## (b) Y -> X
Y = sqrt(1+a^2) * normal()
X = Y / (1+a^2) + a / sqrt(1+a^2) * normal()
```

```
## (c) X <- Z -> Y
Z = normal()
X = Z / b + sqrt(1 - 1/(b^2)) * normal()
Y = b * Z + sqrt(1 + a^2 - b^2) * normal()
```




(a) $X \to Y$ (b) $Y \to X$ (c) $X \leftarrow Z \to Y$

Figure 2.1: Three different random vectors $(X, Y)$ following the identical joint distribution. The Python-like pseudocodes outline the sampling scripts for the corresponding figures. The graphs indicate the orders in the definitions of the random variables. The function normal() samples from the standard normal distribution, and sqrt() is the square-root function. The configuration was $(a, b) = (0.3, 1.03)$.

**Data-generating process (Figure 2.1).** We consider the construction of a random vector $(X, Y)$. To begin with, let $(e_1, e_2, e_3)$ be random variables with distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We can define random variables $X$ and $Y$ by

$$\begin{aligned} X(\cdot) &= f_X(e_1(\cdot)), \\ Y(\cdot) &= f_Y(X(\cdot), e_2(\cdot)), \end{aligned}$$

where $f_X$ and $f_Y$ are measurable functions. We may also define random variables $X$ and $Y$ by

$$\begin{aligned} Y(\cdot) &= g_Y(e_2(\cdot)), \\ X(\cdot) &= g_X(Y(\cdot), e_1(\cdot)), \end{aligned}$$

where $g_Y$ and $g_X$ are measurable functions. Also, we may well define random variables $X$, $Y$, and $Z$ by

$$\begin{aligned} Z(\cdot) &= h_Z(e_3(\cdot)), \\ X(\cdot) &= h_X(Z(\cdot), e_1(\cdot)), \\ Y(\cdot) &= h_Y(Z(\cdot), e_2(\cdot)), \end{aligned}$$

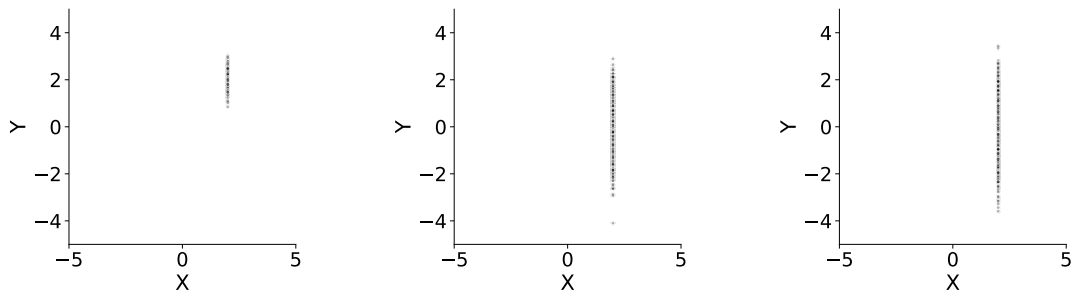
where $h_Z, h_X$, and $h_Y$ are measurable functions.

By choosing $f_X, \ldots, h_Z$ carefully, we can make the joint distributions of $(X, Y)$ the same in all three cases (Figure 2.1). That is, the precise implementations of the three random vectors $(X, Y)$ do not matter as far as we are only interested in the joint distribution. However, the detailed

```
## (a) X -> Y
X = normal()
X = const
Y = X + a * normal()
X = const
```

```
## (b) Y -> X
Y = sqrt(1+a^2) * normal()
X = const
X = Y / (1+a^2) + a / np.sqrt(1+a^2) * normal()
X = const
```

```
## (c) X <- Z -> Y
Z = normal()
X = const
X = Z / b + sqrt(1 - 1/(b^2)) * normal()
X = const
Y = b * Z + sqrt(1 + a^2 - b^2) * normal()
X = const
```




(a) $X \to Y$ (b) $Y \to X$ (c) $X \leftarrow Z \to Y$

Figure 2.2: Distributions of $(X, Y)$ after an intervention. Configurations are the same as Figure 2.1, and const $= 2$.

implementation gets into play as we start to consider interventions.

**Data-generating process after intervention (Figure 2.2).** Now, we consider an intervention to fix $X = \xi$ (a constant function). In the three definitions of $(X(\cdot), Y(\cdot))$ above, we alter the definition of $X(\cdot)$ to a constant function. In the first case,

$$\begin{aligned} X(\cdot) &= \xi, \\ Y(\cdot) &= f_Y(X(\cdot), e_2(\cdot)). \end{aligned}$$

In the second case,

$$\begin{aligned} Y(\cdot) &= g_Y(e_2(\cdot)), \\ X(\cdot) &= \xi, \end{aligned}$$

In the third case,

$$\begin{aligned} Z(\cdot) &= h_Z(e_3(\cdot)), \\ X(\cdot) &= \xi, \\ Y(\cdot) &= h_Y(Z(\cdot), e_2(\cdot)). \end{aligned}$$

Sampling from these $(X, Y)$ can be simulated by altered Python scripts where an assignment statement $X = \xi$ is inserted after every line of the original script (Figure 2.2). The resulting joint distributions are, in fact, different between the first case and the other two cases. While they are all concentrated on $X = \xi$, only the left-hand case has a different distribution of $Y$. That is, the difference in the original data-generating processes can lead to different consequences when we intervene in the data-generating process.

This example demonstrates that certain important aspects of the dependency among random variables may not be captured by the joint distribution. Indeed, when an intervention is performed, the details of implementation do matter; depending on how $X$ and $Y$ are defined, the responses of $Y$ to assigning a constant to $X(\cdot)$ may differ. Causal models, as models of data-generating processes, have been developed to capture such detailed characteristics of the random variables that are not reflected in the distribution [253, 204]. More precisely, causal models make a conjecture that the random variables representing the real-world data have such a structure behind the data distributions and leverage it to perform the causal inference.[1]

The *structural causal framework* [253, 204] formulates certain deterministic relations or the generative mechanisms of the random variables. The formulation captures the concept of the data-generating process we have seen in Section 2.1.1 through an example. The framework is based on the idea that cause-effect relations may be captured by deterministic functional relations. We elaborate on the formulation in Section 2.3 as this is the framework that this dissertation is built on.

### 2.1.2 Statistical Frameworks of Causal Inference

Essentially, statistical causal frameworks can be understood as extensions of standard statistical methodology, both of which are built on top of probability theory. In many circumstances in statistics and statistical machine learning, probability distributions are considered to be the full description of the problem setup. However, in measure-theoretical probability, probability distributions are secondary objects; random variables are defined to be measurable maps from a probability space to a measurable space, and the probability distribution of a random variable is the push-forward measure by that random variable. In this sense, even if some random variables are different as measurable maps, they can have an identical probability distribution. In other words, in conventional problem

[1] Of course, it is not always the case that such a structure exists in the random variables. Therefore, causal models are employed when one can accept the belief that such a structure exists behind the distributions.

setups of statistical frameworks, certain aspects of the random variables are ignored so long as they implement the joint probability distribution of interest, although a distribution only contains some footprints of the original random variables.

On the other hand, when it comes to making *causal inferences*, as opposed to performing classical statistical inferences, the distinction between random variables (random vectors) with the same joint probability distribution becomes crucial. In causal modeling, random variables with identical (joint) probability distributions may have different behavior under certain systematic manipulations. The more detailed properties than what is encoded in the probability distribution come into play. In a contemporary probabilistic view, studies of statistical causal models are the studies of the properties of random variables beyond their probability distributions (when they have such a structure).

We provide a detailed preliminary on the structural causal framework in Section 2.3.

## 2.2 General Notation

In this section, we describe the general notation system used throughout the dissertation. An informed reader may skip this section, except the base notation and the definition of Markov pillow, which are not necessarily commonly used outside of this dissertation.

**Basic sets and operations.** We use $\mathbb{R}$ (resp. $\mathbb{R}_{>0}, \mathbb{R}_{\geq 0}, \mathbb{N}, \mathbb{N}_0$) to denote the set of real (resp. positive real, non-negative real, positive integral, non-negative integral) numbers. We define $[i] := \{1, 2, \dots, i\}$ for $i \in \mathbb{N}$. The empty set is denoted by $\emptyset$. The cardinality of a finite set $A$ is denoted by $|A|$. We use $\langle \cdot, \cdot, \dots, \cdot \rangle$ to denote a tuple, i.e., a finite sequence. The set difference is denoted by "$\setminus$". We use $\coprod$ to denote the disjoint union of sets. Let $d, m \in \mathbb{N}$. For a vector $\boldsymbol{a} = (a^1, \dots, a^d)$ and a subset $S = \{s_1, \dots, s_m\} \subset [d]$ where $s_1 < \cdots < s_m$, we define the *subvector* $\boldsymbol{a}^S := (a^{s_1}, \cdots, a^{s_m})$. For simplicity, we also use $\boldsymbol{a}^s := \boldsymbol{a}^{\{s\}}$. By obvious extension, we use this subvector notation for finite-dimensional vector-valued functions as well as for product spaces of finitely many spaces. For operations involving sets, we write an element $b$ in lieu of a singleton set $\{b\}$ when there is no possibility of confusion. For example, if $b \in A$ and $B \subset A$, we define $B \setminus b := B \setminus \{b\}$.

**Random variables.** Let $X, Y, Z$ be random variables with a joint distribution $P$. We say that $X$ is *conditionally independent* of $Y$ given $Z$, denoted by $X \perp\!\!\!\perp Y \mid Z$, if, for any measurable set $A$ in the sample space of $X$, there exists a version of the conditional probability $P(A|Y, Z)$ which is a function of $Z$ alone. If $Z$ is trivial, we say that $X$ is independent of $Y$, and write $X \perp\!\!\!\perp Y$.[2]

**Directed mixed graph.** A directed mixed graph[3] is a tuple $\mathcal{G} = \langle \mathfrak{V}, \mathfrak{D}, \mathfrak{B} \rangle$ where

- $\mathfrak{V}$ is a finite set,
- $\mathfrak{D}$ and $\mathfrak{B}$ are disjoint subsets of $\mathfrak{V} \times \mathfrak{V}$, and
- $\mathfrak{B}$ satisfies $(u, v) \in \mathfrak{B} \Rightarrow (v, u) \in \mathfrak{B}$ for any $u, v \in \mathfrak{V}$.

We call the elements of $\mathfrak{V}$ *vertices*, those of $\mathfrak{D}$ *uni-directed edges* or simply *directed edges*, and those of $\mathfrak{B}$ *bi-directed edges*. Each element $(u, v) \in \mathfrak{D}$ is denoted by an arrow $(u \to v)$, and each element $(u, v) \in \mathfrak{B}$ is denoted by a bi-directed arrow $(u \leftrightarrow v)$.

[2] The notation is from Dawid [57] and Dawid [58].
[3] The term "mixed" refers to the possible existence of bi-directed edges in addition to uni-directed ones.

**Path.** A path from $u$ to $v$ ($u, v \in \mathfrak{V}$) in a directed mixed graph $\mathcal{G}$ is a finite sequence $\langle v_0, \varepsilon_1, v_1, \varepsilon_2, v_2, \ldots, \varepsilon_k, v_k \rangle$ of alternating nodes and edges in $\mathcal{G}$ for some $k \in \mathbb{N}_0$ satisfying:

- $v_0 = u$, $v_k = v$,
- $\{v_l\}_{l=1}^k \subset \mathfrak{V}$, $\{\varepsilon_l\}_{l=1}^k \subset \mathfrak{D} \cup \mathfrak{B}$,
- for all $l \in [k]$, $\varepsilon_l$ is one of $(v_{l-1} \to v_l)$, $(v_l \to v_{l-1})$, or $(v_{l-1} \leftrightarrow v_l)$, and
- $v_0, \ldots, v_k$ are distinct.

A path of the form $u \to \cdots \to v$ is called a *directed path* from $u$ to $v$. A path of the form $u \leftrightarrow \cdots \leftrightarrow v$ is called a *bi-directed path* between $u$ and $v$.

**Acyclic graphs.** A directed mixed graph $\mathcal{G}$ is called *cyclic* if there exist $u, v \in \mathfrak{V}$ such that there exists a directed path $u \to \cdots \to v$ and a uni-directed edge $(v \to u) \in \mathfrak{D}$.[4] A directed mixed graph $\mathcal{G}$ is called *acyclic* if it is not cyclic.

**Abbreviations.** We abbreviate *acyclic directed mixed graphs* (Richardson [217] and Richardson et al. [218]) as ADMGs. An ADMG is called a *directed acyclic graph* (DAG) if it contains no bi-directed edges.

**Topological ordering.** Let $\mathcal{G} = \langle \mathfrak{V}, \mathfrak{D}, \mathfrak{B} \rangle$ be an ADMG. A total order $\prec$ over $\mathfrak{V}$ is called a *topological ordering* with respect to $\mathcal{G}$ if $u \to v$ implies $u \prec v$ (Koller and Friedman [150, Definition 2.19]).[5] Given a topological ordering $\prec$, we define $\preceq$ in an obvious manner.

**Kinship-based nomenclature.** Let $\mathcal{G} = \langle \mathfrak{V}, \mathfrak{D}, \mathfrak{B} \rangle$ be an ADMG.

- For $v \in \mathfrak{V}$, we define its *parents* $\mathrm{pa}_{\mathcal{G}}(v) \subset \mathfrak{V}$ as

$$\mathrm{pa}_{\mathcal{G}}(v) := \{u \in \mathfrak{V} : \text{there exists a uni-directed edge } u \to v\}.$$

- For $v \in \mathfrak{V}$, we define its *district*[6] $\mathrm{dis}_{\mathcal{G}}(v) \subset \mathfrak{V}$ as

$$\mathrm{dis}_{\mathcal{G}}(v) := \{w \in \mathfrak{V} : \text{there exists a bi-directed path } v \leftrightarrow \cdots \leftrightarrow w\} \cup \{v\}.$$

- For $v \in \mathfrak{V}$, we define its *generalized parents*, denoted by $\overline{\mathrm{pa}}(v) \subset \mathfrak{V}$, as

$$\overline{\mathrm{pa}}_{\mathcal{G}}(v) := \left( \bigcup_{w \in \mathrm{dis}(v)} \mathrm{pa}_{\mathcal{G}}(w) \right) \cup \mathrm{dis}_{\mathcal{G}}(v).$$

- For $v \in \mathfrak{V}$ and a topological ordering $\prec$, we define $\mathcal{G}_{\preceq v}$ as the induced subgraph of $\mathcal{G}$ obtained by restricting the vertices to $v$ and its predecessors, i.e., $\mathrm{pred}(v, \prec) := \{v' \in \mathfrak{V} : v' \preceq v\}$.
- For $v \in \mathfrak{V}$ and a topological ordering $\prec$, we define the *Markov pillow*[7] of $v$ with respect to $\prec$ as

$$\mathrm{mp}_{\mathcal{G}}(v; \prec) := \overline{\mathrm{pa}}_{\mathcal{G}_{\preceq v}}(v) \setminus v.$$

[4] The cyclicity/acyclicity of a directed mixed graph $\mathcal{G}$ is not affected by the existence of bi-directed edges.

[5] Equivalently, $\prec$ is a topological ordering with respect to $\mathcal{G}$ if $u \prec v$ implies that there is no directed path from $v$ to $u$.

[6] The term "district" is adopted from Richardson et al. [218]. Tian and Pearl [273] termed districts as "c-components".

[7] The Markov pillow (as well as generalized parents) is a generalization of the concept of parents in the case where there are bi-directed edges. Indeed, when $\mathcal{G}$ is a DAG (i.e., $\mathfrak{B} = \emptyset$), for any topological ordering $\prec$, we have $\mathrm{mp}(v; \prec) = \mathrm{pa}(v)$.

Table 2.1: Terminology of structural causal models.

| Symbol | Terminology |
|---|---|
| $\mathcal{I}$ | Index set of *endogenous* variables |
| $\mathcal{J}$ | Index set of *exogenous* variables |
| $\{\mathcal{Z}^v\}_{v\in\mathcal{I}}$ | Domains of the endogenous variables |
| $\{\mathcal{E}^u\}_{u\in\mathcal{J}}$ | Domains of the exogenous variables |
| $\boldsymbol{f}$ | Structural function (SF)[8] |
| $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ | Exogenous distribution |
| $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ | Observational distribution |
| $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ | Joint solution distribution |
| $\boldsymbol{F}$ | Reduced-form structural function (RSF) |

- For $v \in \mathfrak{V}$, we define the descendants of $v$ as

$$\begin{aligned}\mathrm{desc}_{\mathcal{G}}(v) := \{u \in \mathfrak{V} :&\text{a directed path } v \to \cdots \to u \text{ exists and}\\ &\text{no directed path} u \to \cdots \to v \text{ exists}\}.\end{aligned}$$

- For $v \in \mathfrak{V}$, we define the non-descendants of $v$ as $\text{non-desc}_{\mathcal{G}}(v) := \mathfrak{V} \setminus (\mathrm{desc}(v) \cup v)$.

When $\mathcal{G}$ is obvious from the context, we omit $\mathcal{G}$ from the notation of $\mathrm{pa}_{\mathcal{G}}, \mathrm{dis}_{\mathcal{G}}$, $\overline{\mathrm{pa}}_{\mathcal{G}}$, $\mathrm{mp}_{\mathcal{G}}$, $\mathrm{desc}_{\mathcal{G}}$, and $\text{non-desc}_{\mathcal{G}}$.

## 2.3 Structural Causal Framework

In this section, we introduce the formal treatment of the structural causal framework, which serves the subsequent chapters with a unified ground. The structural causal modeling framework defines a suite of generative models, namely structural causal models (SCMs) and graphical causal models (GCMs). An SCM captures the detailed data-generating process using deterministic functions, while a GCM captures the more qualitative cause-effect relations in the data-generating process. Based on both SCM and GCM, the notions of *interventional distributions* can be naturally defined, and they are compatible, which is one of the initial motivations for which these models were developed (Pearl [204]).

### 2.3.1 Structural Causal Models (SCMs)

In this dissertation, we adopt the definition of SCMs from Bongers et al. [29] for its clarity of exposition. SCMs characterize random variables by some deterministic functional equations which they satisfy, such as the equations used to define the random vectors $(X, Y)$ in the examples of Section 2.1.1.

**Definition 2.1** (Structural Causal Model [297, 94, 204, 29])**.** *A* structural causal model *(SCM)*[9] *is a tuple* $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*, where*

1. *$\mathcal{I}$ and $\mathcal{J}$ are disjoint finite sets,*
2. *$\boldsymbol{\mathcal{Z}} = \prod_{v\in\mathcal{I}} \mathcal{Z}^v$, $\boldsymbol{\mathcal{E}} = \prod_{u\in\mathcal{J}} \mathcal{E}^u$, and each $\mathcal{Z}^v$ and $\mathcal{E}^u$ is a standard measurable space,*[10, 11]

[8] To the best of the author's knowledge, often in the literature, this object is not given a terminology. The term *structural function* is not a standard one, but it is used in this dissertation for convenience.

[9] SCMs are also known as functional causal models (FCMs) or structural equation models (SEMs) [204].

[10] A standard measurable space is a measurable space that is isomorphic (as measurable spaces) to a Polish space endowed with the Borel $\sigma$-algebra (Çinlar [47, Section I.2]).

[11] We temporarily consider a fixed ordering over $\mathcal{I}$ and $\mathcal{J}$ when we refer to $\prod_{v\in\mathcal{I}}$ or $\prod_{u\in\mathcal{J}}$.

3. $\boldsymbol{f} : \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Z}}$ *is a measurable map,*

4. $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{u \in \mathcal{J}} \mathbb{P}_{\mathcal{E}^u}$ *is a product measure, and each* $\mathbb{P}_{\mathcal{E}^u}$ *is a probability measure on* $\mathcal{E}^u$.

*Given an SCM, the equation*

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}) \quad \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}} \tag{2.1}$$

*is called the structural equation (SE) of* $\mathcal{M}$.

Table 2.1 summarizes the terminology used to refer to each constituent of an SCM. Conversely, random variables that are described by a given SCM are called a *solution*.

**Definition 2.2** (Solution of an SCM [29])**.** *A pair* $(\boldsymbol{Z}, \boldsymbol{E})$ *of random variables* $\boldsymbol{Z} : \Omega \to \boldsymbol{\mathcal{Z}}, \boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$, *where* $\Omega$ *is the sample space of a probability space, is a* solution *of the SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *if*

- $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}_{\boldsymbol{E}}$ *where* $\mathbb{P}_{\boldsymbol{E}}$ *is the distribution of* $\boldsymbol{E}$, *and*
- *the SE is satisfied, i.e.,* $\boldsymbol{Z} = \boldsymbol{f}(\boldsymbol{Z}, \boldsymbol{E})$ *a.s.*

*The joint distribution* $\mathbb{P}_{\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}}$ *of a solution* $(\boldsymbol{Z}, \boldsymbol{E})$ *is called a* joint solution distribution *of* $\mathcal{M}$.[12] *For convenience,* $\boldsymbol{Z}$ *is also called a* solution *of* $\mathcal{M}$ *if there exists a random variable* $\boldsymbol{E}$ *such that* $(\boldsymbol{Z}, \boldsymbol{E})$ *is a solution of* $\mathcal{M}$. *If* $\boldsymbol{Z}$ *is a solution of* $\mathcal{M}$, *we also say that* $\boldsymbol{Z}$ *is generated by* $\mathcal{M}$, *and we write* $\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}$. *Analogously, if* $\{(\boldsymbol{Z}_i, \boldsymbol{E}_i)\}_{i=1}^n$ *are solutions of* $\mathcal{M}$ *defined on the same probability space, and if* $\{\boldsymbol{E}_i\}_{i=1}^n$ *are independent, then we write* $\{\boldsymbol{Z}_i\}_{i=1}^n \overset{i.i.g.}{\leftarrow} \mathcal{M}$.[13] *The distribution of a solution* $\boldsymbol{Z}$ *is called an* observational distribution *of* $\mathcal{M}$.

We define the following graphical objects by extracting the dependency structure from the SF of an SCM. They reflect the qualitative dependency structure that is intrinsic in the SF.

**Definition 2.3** (Graph and Augmented Graph of an SCM [29])**.** *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM. The* augmented graph *of* $\mathcal{M}$ *is the directed mixed graph* $\mathrm{graph}(\mathcal{M}) = \langle \mathfrak{V}^{\mathrm{a}}, \mathfrak{D}^{\mathrm{a}}, \emptyset \rangle$ *defined as follows:*

- *the vertex set is* $\mathfrak{V}^{\mathrm{a}} = \mathcal{I} \coprod \mathcal{J}$, *and*
- *the uni-directed edge set* $\mathfrak{D}^{\mathrm{a}}$ *is defined by*

$$(u \to v) \in \mathfrak{D}^{\mathrm{a}}$$
$$\Leftrightarrow \begin{cases} v \in \mathcal{I}, \text{ and} \\ \nexists \tilde{f} : \boldsymbol{\mathcal{Z}}^{\mathcal{I} \setminus u} \times \boldsymbol{\mathcal{E}}^{\mathcal{J} \setminus u} \to \mathcal{Z}^v \text{ s.t. } \tilde{f}(\boldsymbol{z}, \boldsymbol{e}) = \boldsymbol{f}^v(\boldsymbol{z}, \boldsymbol{e}) \ \forall \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}. \end{cases}$$

*Also, the* graph *of* $\mathcal{M}$ *is the directed mixed graph* $\mathrm{obsGraph}(\mathcal{M}) = \langle \mathfrak{V}, \mathfrak{D}, \mathfrak{B} \rangle$ *obtained from* $\mathrm{graph}(\mathcal{M})$ *as follows:*[14]

- *the vertex set is* $\mathfrak{V} = \mathcal{I}$,
- *the bi-directed edge set* $\mathfrak{B}$ *is defined by*

$$(u \leftrightarrow v) \in \mathfrak{B} \Leftrightarrow \exists w \in \mathcal{J} \text{ s.t. } (w \to u), (w \to v) \in \mathfrak{D}^{\mathrm{a}}, \text{ and}$$

[12] The term *joint solution distribution* is not common and may only be used in this dissertation for convenience.
[13] Here, i.i.g. stands for "independently and identically generated."
[14] Note that, by definition, $\mathcal{J}$ has no parents in $\mathrm{graph}(\mathcal{M})$. As a result, the procedure to obtain $\mathrm{obsGraph}(\mathcal{M})$ from $\mathrm{graph}(\mathcal{M})$ in Definition 2.3 is the same as the *latent projection* (Algorithm 1).

- *the uni-directed edge set* $\mathfrak{D}$ *is defined by*

$$(u \to v) \in \mathfrak{D} \Leftrightarrow (u \to v) \in \mathfrak{D}^{\mathrm{a}} \text{ and } (u \leftrightarrow v) \notin \mathfrak{B}.$$

In this dissertation, we only consider acyclic SCMs, defined as follows. The examples in Section 2.1.1 fall into this category.

**Definition 2.4** (Acyclic SCM [29])**.** *We say that an SCM* $\mathcal{M}$ *is* acyclic *(or* semi-Markovian*) if and only if its graph* obsGraph($\mathcal{M}$) *is an ADMG.*[15] *Moreover, a semi-Markovian SCM* $\mathcal{M}$ *is called* Markovian *if* obsGraph($\mathcal{M}$) *is a DAG. We denote the set of semi-Markovian SCMs for* $\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}$ *as* SSCM($\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}$)*.*

In some cases, we can solve the SE for $\boldsymbol{z}$, i.e., find a measurable function $\boldsymbol{F}$ such that $\boldsymbol{z} = \boldsymbol{F}(\boldsymbol{e})$. Such functions are called *reduced-form SFs*. Figure 2.3 shows examples of a structural-form SE and its corresponding reduced-form SE.

**Definition 2.5** (Reduced-form SFs [215])**.** *Let* $\mathcal{M}$ *be an SCM,* $\boldsymbol{f}$ *be its structural function, and* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ *be its exogenous distribution. A measurable map* $\boldsymbol{F} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Z}}$ *is called a* reduced-form structural function *(RSF)*[16] *of* $\mathcal{M}$ *if it satisfies*

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}) \;\Rightarrow\; \boldsymbol{z} = \boldsymbol{F}(\boldsymbol{e}), \quad \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\text{-a.s.}(\boldsymbol{e}).$$

If an RSF exists, it is unique in the following sense.

**Proposition 2.1** (Pairwise Uniqueness of RSF)**.** *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM. If* $\boldsymbol{F}$ *and* $\boldsymbol{F}'$ *are both RSFs of* $\mathcal{M}$*, they are* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost surely equal on* $\boldsymbol{\mathcal{E}}_{\boldsymbol{f}} := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e})\}$.[17] *That is, there exists a* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-negligible set* $\boldsymbol{\mathcal{E}}_{\neg} \subset \boldsymbol{\mathcal{E}}$ *such that* $\boldsymbol{F} = \boldsymbol{F}'$ *on* $\boldsymbol{\mathcal{E}}_{\boldsymbol{f}} \setminus \boldsymbol{\mathcal{E}}_{\neg}$*.*

*Proof.* Let $\boldsymbol{\mathcal{E}}_0$ and $\boldsymbol{\mathcal{E}'}_0$ be the two $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-negligible sets corresponding to $\boldsymbol{F}$ and $\boldsymbol{F}'$, respectively. If we let $\boldsymbol{\mathcal{E}}_{\neg} := \boldsymbol{\mathcal{E}}_0 \cup \boldsymbol{\mathcal{E}'}_0$, then $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{E}}_{\neg}) = 0$, and for any pair $(\boldsymbol{z}, \boldsymbol{e}) \in \boldsymbol{\mathcal{Z}} \times (\boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{E}}_{\neg})$ satisfying $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e})$, we have $\boldsymbol{F}(\boldsymbol{e}) = \boldsymbol{z} = \boldsymbol{F}'(\boldsymbol{e})$. □

In the case of acyclic SCMs, we can show the existence of an RSF. In the proof in Appendix A.2, the RSF is constructed by solving Equation (2.1) for $\boldsymbol{z}$ by iterative elimination of variables.

**Proposition 2.2** (Existence of RSF)**.** *If* $\mathcal{M}$ *is an acyclic SCM,* $\mathcal{M}$ *has an RSF.*

When an RSF exists, the joint solution distribution and the observational distribution of $\mathcal{M}$ are obtained by transforming $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ by the RSF.

**Proposition 2.3.** *Let* $\boldsymbol{F}$ *be an RSF of an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*. Then, for any solution* $(\boldsymbol{Z}, \boldsymbol{E})$*, we have*

$$\begin{aligned} &\boldsymbol{Z} = \boldsymbol{F}(\boldsymbol{E}) \text{ a.s.}, \\ &\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\mathrm{d}\boldsymbol{z}, \mathrm{d}\boldsymbol{e}) = \delta_{\boldsymbol{F}(\boldsymbol{e})}(\mathrm{d}\boldsymbol{z})\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\mathrm{d}\boldsymbol{e}), \\ &\mathbb{P}_{\boldsymbol{\mathcal{Z}}} = \boldsymbol{F}_{\sharp}(\mathbb{P}_{\boldsymbol{\mathcal{E}}}) = \mathbb{P}_{\boldsymbol{\mathcal{E}}} \circ \boldsymbol{F}^{-1}, \end{aligned}$$

*where* $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ *and* $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ *are the distributions of* $(\boldsymbol{Z}, \boldsymbol{E})$ *and* $\boldsymbol{Z}$*, respectively. In particular,* $\mathcal{M}$ *has a unique joint solution distribution and unique observational distribution.*

[15] Equivalently, an acyclic SCM is one for which graph($\mathcal{M}$) is acyclic.
[16] Given a reduced-form SF $\boldsymbol{F}$, the equation $\boldsymbol{z} = \boldsymbol{F}(\boldsymbol{e})(\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}})$ is called a *reduced-form structural equation* (RSE) of $\mathcal{M}$. In contrast, we call the original equations (Equation (2.1)) the *structural-form* SEs.
[17] The notion of the uniqueness of RSF shown here is weaker than the claim that the RSF is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost surely unique on $\boldsymbol{\mathcal{E}}_{\boldsymbol{f}}$, which would mean that there exists a single $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-negligible set $\boldsymbol{\mathcal{E}}_{\neg}$ on which all RSFs match.

$$\boldsymbol{z} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \boldsymbol{z} + \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix} \boldsymbol{e}$$

(a) Structural-form structural equation

$$\boldsymbol{z} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix} \boldsymbol{e}$$

(b) Reduced-form structural equation

Figure 2.3: Examples of structural equations.

*Proof.* Let $\boldsymbol{\mathcal{E}}_{\neg}$ be the $\boldsymbol{\mathcal{E}}$-negligible set corresponding to $\boldsymbol{F}$. Since $\boldsymbol{Z} = \boldsymbol{f}(\boldsymbol{Z}, \boldsymbol{E})$ and $\boldsymbol{E} \notin \boldsymbol{\mathcal{E}}_{\neg}$ are almost surely satisfied, we almost surely have $\boldsymbol{Z} = \boldsymbol{F}(\boldsymbol{E})$. Applying Lemma A.1, we have the assertion. Uniqueness follows from Fact A.2. □

In light of Proposition 2.2 and Proposition 2.3, we are guaranteed that acyclic SCMs have unique joint solution distributions and unique observational distributions.

**Definition 2.6.** *Let $\mathcal{M}$ be an SCM with a unique joint solution distribution. Define* $\mathrm{dist}(\mathcal{M})$ *to be the joint solution distribution $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ of $\mathcal{M}$. Also define* $\mathrm{obsDist}(\mathcal{M})$ *to be the observational distribution $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ of $\mathcal{M}$.*

### 2.3.2 Graphical Causal Models (GCMs)

**Definition 2.7** (Recursive Factorization [142, 160])**.** *A probability measure $P$ over $\boldsymbol{\mathcal{Z}} = \prod_{v \in \mathcal{I}} \mathcal{Z}^v$ is said to* recursively factorize[18] according to *a DAG $\mathcal{G}$ if, for each $v \in \mathfrak{V}$, there exists a Markov kernel*[19] *$K^v$ from $\boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)}$ to $\boldsymbol{\mathcal{Z}}^v$ such that*

$$\mathbb{P}_{\boldsymbol{\mathcal{Z}}}(\mathrm{d}\boldsymbol{z}) = \prod_{v \in \mathcal{I}} K^v(\boldsymbol{z}^{\mathrm{pa}(v)}, \mathrm{d}z^v).$$

**Definition 2.8** (Probabilistic graphical models)**.** *A probabilistic graphical model (PGM) is a tuple $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$, where*

- *$\mathcal{I}$, $\mathcal{J}$, $\boldsymbol{\mathcal{Z}}$, and $\boldsymbol{\mathcal{E}}$ satisfy Conditions 1 and 2 of Definition 2.1,*
- *$\mathcal{G}^{\mathrm{a}}$ is a DAG with vertex set $\mathcal{I} \coprod \mathcal{J}$, and*
- *$\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ is a probability measure over $\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}$ that recursively factorizes according to $\mathcal{G}^{\mathrm{a}}$.*

*We call $\mathcal{I}$ the* observed *index set and $\mathcal{J}$ the* unobserved *index set.*[20] *Define* $\mathrm{graph}(\mathcal{M}) := \mathcal{G}^{\mathrm{a}}$, $\mathrm{dist}(\mathcal{M}) := \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$, *and* $\mathrm{obsDist}(\mathcal{M}) := \mathbb{P}_{\boldsymbol{\mathcal{Z}}}$, *where $\mathbb{P}_{\boldsymbol{\mathcal{Z}}} = \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\cdot, \boldsymbol{\mathcal{E}})$ is the marginal distribution.*

This measure-theoretic definition of the PGMs, namely the definition using factorization in terms of *Markov kernels* instead of conditional *densities*, can also be found in Wu et al. [298]. For more details on Markov kernels and conditional distributions, see Çinlar [47, Chapter I, Section 6] and Çinlar [47, Chapter IV, Section 2], respectively. For a comprehensive account and historical remarks on PGMs, see Lauritzen [160].

In the context of the structural causal framework, PGMs are often endowed with an operator to model the interventional distributions, which we define later in Definition 2.13.

[18] The terminology "recursively factorize" is adopted from Lauritzen [160]. In the literature, there are other expressions to refer to the same notion, e.g., a distribution is said to be *Markov relative to* a graph when the recursive factorization property is satisfied (Pearl [204, Definition 1.2.2]).

[19] See Definition A.1 in Appendix A.2.1 for the definition of Markov kernels.

[20] At this point, the distinction between $\mathcal{I}$ and $\mathcal{J}$ is unnecessary. However, the notation to distinguish the two sets can help us clarify the relations among different causal models.

**Algorithm 1** Latent projection (Verma [283])

**Input:** DAG $\mathcal{G}^{\mathrm{a}} = \langle \mathcal{I} \coprod \mathcal{J}, \mathfrak{D} \rangle$

**for** $v, v' \in \mathcal{I}$ such that $v \neq v'$
  **if** $(v \to v') \in \mathfrak{D}$ **then**
    Add $(v \to v')$ to $\tilde{\mathfrak{D}}$.
  **if** there exists a path $(v \to \cdots \to v')$ in $\mathcal{G}^{\mathrm{a}}$ such that all the internal nodes in the path are the elements of $\mathcal{J}$ **then**
    Add $(v \to v')$ to $\tilde{\mathfrak{D}}$.
  **if** there exists a path $(v \leftarrow \cdots \leftarrow u \to \cdots \to v')$ in $\mathcal{G}^{\mathrm{a}}$ such that all the internal nodes in the path are the elements of $\mathcal{J}$ (including $u$ itself) **then**
    Add $(v \leftrightarrow v')$ to $\tilde{\mathfrak{B}}$.

**Output:** ADMG $\mathcal{G} = \langle \mathcal{I}, \tilde{\mathfrak{D}}, \tilde{\mathfrak{B}} \rangle$

**Definition 2.9** (Markovian GCMs)**.** *When a PGM is endowed with the* $\mathrm{do}(\cdot, \cdot)$ *operator (Definition 2.13), we call it a* Markovian graphical causal model *(Markovian GCM; a.k.a. Markovian causal graphical models).*[21] *We denote the set of Markovian GCMs for* $\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}$ *as* $\mathrm{MGCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}})$.

The modifier "Markovian" connotes that the model explicitly includes the unobserved variables, namely by $\mathcal{J}$, $\mathcal{G}^{\mathrm{a}}$, and $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ (in a way, all variables are "observed" by this model). On the other hand, *semi-Markovian* GCMs, as defined below, are the models in which only the observed variables explicitly appear. They are essentially the equivalence classes of Markovian GCMs with respect to an equivalence relation which "squashes" the unobserved variables by marginalization and graph manipulation. Each Markovian model in the equivalence class may have different situations (definitions and distributions) of unobserved variables, but they share certain aspects of the behavior of the observed random variables. The equivalence relation is defined by the following operation called the *latent projection* (Verma [283] and Tian [271], and Evans [74]).

**Definition 2.10** (Latent projection [283, 271, 74])**.** *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$ *be a Markovian GCM. Then, define*

$$\pi_{\mathrm{SGCM}}(\mathcal{M}) := \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle,$$

*where* $\mathcal{G}$ *is obtained by Algorithm 1 from* $\mathcal{G}^{\mathrm{a}}$*, and* $\mathbb{P}_{\boldsymbol{\mathcal{Z}}} = \mathrm{obsDist}(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}})$ *is the marginal distribution over* $\boldsymbol{\mathcal{Z}}$*. By slight abuse of notation, we also write* $\pi_{\mathrm{SGCM}}(\mathcal{G}^{\mathrm{a}}) = \mathcal{G}$ *and* $\pi_{\mathrm{SGCM}}(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) = \mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ *when* $\mathcal{M}$ *is clear from the context.*

Semi-Markovian GCMs are the sets of Markovian GCMs that are projected to the same tuple.

**Definition 2.11** (Semi-Markovian GCM)**.** *Let* MGCM *denote the set of all Markovian GCMs. Consider the equivalence relation (*projection equivalence*) for* MGCM *defined by*

$$\mathcal{M} \sim \mathcal{M}' \Leftrightarrow \pi_{\mathrm{SGCM}}(\mathcal{M}) = \pi_{\mathrm{SGCM}}(\mathcal{M}').$$

*It is easy to confirm that* $\sim$ *is indeed an equivalence relation. Then, a* semi-Markovian graphical causal model *(semi-Markovian GCM) is an equivalence class of* MGCM *with respect to* $\sim$.

*Remark* 2.1 (Specification of a semi-Markovian GCM). A semi-Markovian GCM is specified by a tuple $\mathcal{M} = \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$, where $\mathcal{I}$ and $\boldsymbol{\mathcal{Z}}$ satisfy Conditions 1 and 2 of Definition 2.1, $\mathcal{G}$ is an ADMG with vertex set $\mathcal{I}$, and $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ is a probability measure over $\boldsymbol{\mathcal{Z}}$. Indeed, if $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{Z}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\mathcal{G}}^{\mathrm{a}}, \tilde{\mathbb{P}}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$

[21] The distinction between PGMs and Markovian GCMs is similar in spirit to the distinction between a metrizable space and a metric space.

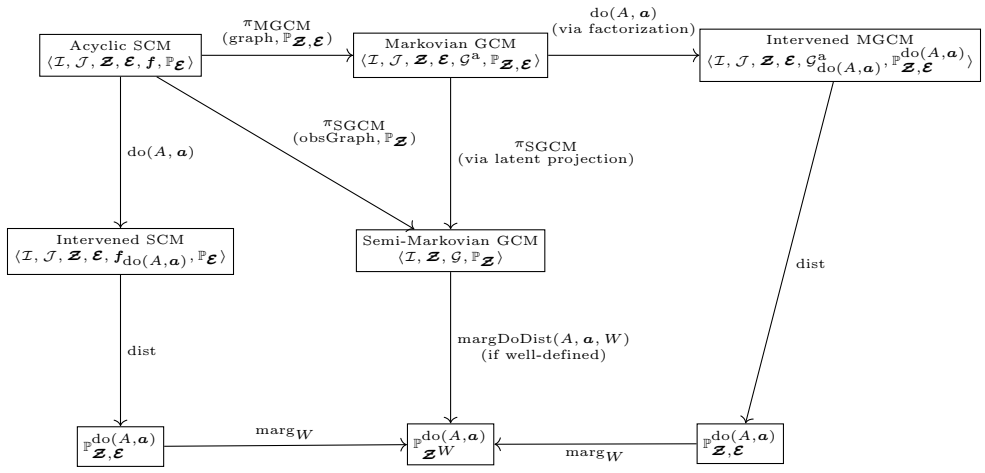



Figure 2.4: Relations among SCMs and GCMs.

is a Markovian GCM,

$$\tilde{\mathcal{M}} \in \mathcal{M} \Leftrightarrow \begin{cases} \tilde{\mathcal{I}} = \mathcal{I}, \quad \tilde{\boldsymbol{\mathcal{Z}}} = \boldsymbol{\mathcal{Z}}, \\ \pi_{\text{SGCM}}(\tilde{\mathcal{G}}^{\text{a}}) = \mathcal{G}, \quad \pi_{\text{SGCM}}(\tilde{\mathbb{P}}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) = \mathbb{P}_{\boldsymbol{\mathcal{Z}}}. \end{cases}$$

To conform to the standard notation of equivalence classes, the semi-Markovian GCM to which a Markovian GCM $\mathcal{M}$ belongs is denoted by $[\mathcal{M}]$. In fact, $[\mathcal{M}] = \pi_{\text{SGCM}}(\mathcal{M})$. Given a semi-Markovian GCM $\mathcal{M} = \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$, we call $\mathcal{G}$ the *causal graph* of $\mathcal{M}$. We denote the set of semi-Markovian GCMs for $\mathcal{I}, \boldsymbol{\mathcal{Z}}$ as SGCM$(\mathcal{I}, \boldsymbol{\mathcal{Z}})$.

*Remark* 2.2 (Natural embedding). A Markovian GCM $\langle \mathcal{I}, \emptyset, \boldsymbol{\mathcal{Z}}, \emptyset, \mathcal{G}^{\text{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ can be naturally mapped to a semi-Markovian GCM $\langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}^{\text{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$, since we have $\pi_{\text{SGCM}}(\mathcal{G}^{\text{a}}) = \mathcal{G}^{\text{a}}$ and $\pi_{\text{SGCM}}(\mathbb{P}_{\boldsymbol{\mathcal{Z}}}) = \mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ due to $\mathcal{J} = \emptyset$. It is immediate that this map is injective (but not surjective).

Note that a single distribution $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ may be coupled with different $\mathcal{G}$ to form different semi-Markovian GCMs. In other words, semi-Markovian GCMs distinguish distributions by annotating them with the graph to preserve a certain aspect of the data-generating process.

### 2.3.3 Relation between SCMs and GCMs

Here, we elaborate on the relation between SCMs and GCMs, shedding light on their hierarchical nature. While the contents of this section do not directly relate to the development of the subsequent chapters, they are useful in clarifying the hierarchical relation between SCMs and GCMs. The reader may well skip to Section 2.4, where we describe the properties of the GCMs and the SCMs on which we focus in this dissertation. The relation is summarized in Figure 2.4. GCMs can be seen as a coarsening of SCMs where the graphical dependency structure that can be used for calculating the interventional distributions is extracted from the SFs.

**Modeling Interventional Distributions.** We define the notions of *interventional distributions* on SCMs and GCMs.[22] The operations defined here are summarized in Figure 2.4.[23] The following is a definition of *perfect interventional distributions* derived from an SCM, with which an SCM can

[22] Sometimes in the literature, the interventional distributions are called the *causal effects* (Pearl [204, Definition 3.2.1]).

[23] Nonstandard notation for various operators (such as doDist, doGraph, margDoDist) is devised in this section. We believe they compactly convey the meaning and are easier to remember than the conventional notation.

be used as a model of interventional distributions. It formalizes the intervention operation of the example in Section 2.1.1.

**Definition 2.12** (Perfect interventional distributions for SCMs)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an acyclic SCM. For $A \subset \mathcal{I}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{Z}}^A$, we define a map* $\mathrm{do}(A, \boldsymbol{a})$ *as*

$$\mathrm{do}(A, \boldsymbol{a}) : \mathcal{M} \mapsto \mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})},$$

*where $\mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}_{\mathrm{do}(A,\boldsymbol{a})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is the* intervened SCM *defined by*

$$\boldsymbol{f}^v_{\mathrm{do}(A,\boldsymbol{a})} = \begin{cases} \boldsymbol{f}^v(\boldsymbol{z}, \boldsymbol{e}) & \text{if } v \notin A, \\ \boldsymbol{a}^v & \text{if } v \in A. \end{cases}$$

*For convenience, we define*[24]

$$\begin{aligned} \mathrm{doFunc}(A, \boldsymbol{a})(\mathcal{M}) &:= \mathrm{doFunc}(A, \boldsymbol{a})(\boldsymbol{f}) := \boldsymbol{f}_{\mathrm{do}(A,\boldsymbol{a})}, \\ \mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}) &:= \mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} := \mathrm{dist}(\mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})}). \end{aligned}$$

*To summarize, with these definitions,*

$$\begin{aligned} \mathrm{do}(A, \boldsymbol{a})(\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle) &= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathrm{doFunc}(A, \boldsymbol{a})(\boldsymbol{f}), \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle, \\ \mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}) &= \mathrm{dist}(\mathrm{do}(A, \boldsymbol{a})(\mathcal{M})). \end{aligned}$$

*The marginal distribution $\mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}}} := \mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\cdot, \boldsymbol{\mathcal{E}})$ is called the* interventional distribution induced by $\mathcal{M}$ *under the perfect intervention* $\mathrm{do}(A, \boldsymbol{a})$*.*

Given a Markovian GCM, one can define the following operation, with which a Markovian GCM can be used as a model of interventional distributions.

**Definition 2.13** (Perfect interventional distributions for Markovian GCMs)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$ be a Markovian GCM. Recall that $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ recursively factorizes according to $\mathcal{G}^{\mathrm{a}}$ (Definition 2.7), and let $\{K^v\}_{v \in \mathcal{I} \coprod \mathcal{J}}$ be Markov kernels (where $K^v$ is from $\boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)}$ to $\boldsymbol{\mathcal{Z}}^v$) such that*

$$\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\mathrm{d}\boldsymbol{\xi}) = \prod_{v \in \mathcal{I} \cup \mathcal{J}} K^v(\boldsymbol{\xi}^{\mathrm{pa}(v)}, \mathrm{d}\boldsymbol{\xi}^v).$$

*For $A \subset \mathcal{I}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{Z}}^A$, we define a map* $\mathrm{do}(A, \boldsymbol{a})$ *as*

$$\mathrm{do}(A, \boldsymbol{a}) : \mathcal{M} \mapsto \mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})}, = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}_{\mathrm{do}(A,\boldsymbol{a})}, \mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$$

*where $\mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})}$ is a Markovian GCM called the* intervened MGCM *defined by*

- *the intervened graph $\mathcal{G}^{\mathrm{a}}_{\mathrm{do}(A,\boldsymbol{a})}$, which is identical to $\mathcal{G}^{\mathrm{a}}$ except that we remove the edges whose arrow heads are pointed to an element of $A$, and*
- *the intervened distribution $\mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ defined by*

$$\mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\mathrm{d}\boldsymbol{\xi}) = \prod_{v \in \mathcal{I} \cup \mathcal{J}} \left( \begin{cases} K^v(\boldsymbol{\xi}^{\mathrm{pa}(v)}, \mathrm{d}\boldsymbol{\xi}^v), & \text{if } v \notin A, \\ \delta_{\boldsymbol{a}^v}(\mathrm{d}\boldsymbol{\xi}^v), & \text{if } v \in A. \end{cases} \right) \tag{2.2}$$

*with $\boldsymbol{\xi} = (\boldsymbol{z}, \boldsymbol{e})$, and $\delta_x(\cdot)$ is the Dirac measure centered at $x$.*

[24] Since $\mathcal{M}$ is assumed to be acyclic, $\mathcal{M}^{\mathrm{do}(A,\boldsymbol{a})}$ is also acyclic.

*For convenience, we define*

$$\mathrm{doGraph}(A, \boldsymbol{a})(\mathcal{M}) := \mathrm{doGraph}(A, \boldsymbol{a})(\mathcal{G}^{\mathrm{a}}) := \mathcal{G}^{\mathrm{a}}_{\mathrm{do}(A,\boldsymbol{a})},$$
$$\mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}) := \mathrm{doDist}(A, \boldsymbol{a})(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) := \mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}.$$

*To summarize, with these definitions,*

$$\mathrm{do}(A, \boldsymbol{a})(\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle)$$
$$= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathrm{doGraph}(A, \boldsymbol{a})(\mathcal{G}^{\mathrm{a}}), \mathrm{doDist}(A, \boldsymbol{a})(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) \rangle.$$

*The distribution* $\mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}}} := \mathbb{P}^{\mathrm{do}(A,\boldsymbol{a})}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}(\cdot, \boldsymbol{\mathcal{E}})$ *is called the* interventional distribution *induced by* $\mathcal{M}$ *under the perfect intervention* $\mathrm{do}(A, \boldsymbol{a})$.[25]

Given a semi-Markovian GCM, one can define the following operation, with which a semi-Markovian GCM can be used as a model of interventional distributions. However, unlike the case of SCMs or Markovian GCMs, not all interventional distributions are well-defined; that is, the interventional distributions may differ for different Markovian GCMs in the same semi-Markovian GCM.

**Definition 2.14** (Perfect interventional distributions for semi-Markovian GCMs)**.** *Let* $[\mathcal{M}]$ *be a semi-Markovian GCM containing* $\mathcal{M}$*. For* $A, W \subset \mathcal{I}$ *and* $\boldsymbol{a} \in \boldsymbol{\mathcal{Z}}^A$*, we define a map* $\mathrm{margDoDist}(A, \boldsymbol{a}, W)$ *as*

$$\mathrm{margDoDist}(A, \boldsymbol{a}, W) : [\mathcal{M}] \mapsto \mathrm{marg}_W(\mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}))$$

*whenever it is well-defined, i.e., when the right-hand side does not depend on the choice of* $\mathcal{M}$*. Here,* $\mathrm{marg}_W$ *is an operator to marginalize a probability distribution for* $W$*. When* $\mathrm{margDoDist}(A, \boldsymbol{a}, W)([\mathcal{M}])$ *is well-defined, we say that the marginal perfect interventional distribution of* $W$ *is* identifiable *in* $[\mathcal{M}]$ *under* $\mathrm{do}(A, \boldsymbol{a})$.[26]

*Remark* 2.3 (Identifiability of causal quantities). In the earlier literature on the identifiability of interventional distributions, various sufficient conditions for the identifiability were explored (for a brief review, see, e.g., Shpitser and Pearl [245]), such as the *backdoor criterion* [205, 204]. Halpern [96] showed that the rules of *do-calculus* proposed by Pearl [203] are complete for the identification. However, this was an axiomatic approach from which an explicit algorithm was not obtained. Later, Tian and Pearl [272] proposed the necessary and sufficient condition for identifying the *joint* interventional distribution, i.e., $\mathrm{margDoDist}(A, \boldsymbol{a}, \mathcal{I})(\mathcal{M})$. For marginal interventional distributions as defined in Definition 2.14, i.e., $\mathrm{margDoDist}(A, \boldsymbol{a}, W)(\mathcal{M})$ where $W$ is a strict subset of $\mathcal{I}$, the condition in Tian and Pearl [272] was only a sufficient condition and not a necessary condition for the identification. Soon after, Shpitser and Pearl [245] and Huang and Valtorta [116] concurrently provided the algorithms to decide whether $\mathrm{margDoDist}(A, \boldsymbol{a}, W)(\mathcal{M})$ is identifiable, one of which was a modified version of an algorithm found in Tian [271]. For a detailed review of the identifiability of interventional distributions, see Shpitser and Tian [246]. Similarly to the perfect intervention, there are other operators used for modeling causal quantities. One such example is the conditional interventional distribution, for which Shpitser and Pearl [244] provided a complete algorithm for deciding the identifiability.

*Remark* 2.4 (Naturalness of the definitions of interventional distributions). Here, we have provided a constructive definition of the interventional distributions, but historically these definitions have

[25] Equation (2.2) is a measure-theoretical version of what is known under the names of *g-formula* (Robins [221]), the *manipulated distribution* (Spirtes et al. [253]), or the *truncated factorization* (Pearl [204]).

[26] In contrast, it is said that all interventional distributions are identified in a Markovian GCM because it is a map and hence trivially "well-defined" (Pearl [204, Corollary 3.2.6]).

been derived based on conceptual considerations such as *modularity* or *autonomy* of causality (see, e.g., Woodward [295], Craver and Tabery [54, 2.4.3], Menzies [181]).

*Remark* 2.5 (Notation of the do operator). The operator $\mathrm{do}(A, \boldsymbol{a})$ is also denoted by placing $\mathrm{do}(A = \boldsymbol{a})$ or $\hat{\boldsymbol{a}}$ in the conditioning part of the usual notation of the conditional distribution (e.g., $p(\boldsymbol{z}|\mathrm{do}(A = \boldsymbol{a}))$ or $p(\boldsymbol{z}|\hat{\boldsymbol{a}})$; Pearl [204])

**Compatibility of the definitions.** The following proposition, which immediately follows from Corollary 8.3 of Bongers et al. [29], clarifies the relation between SCMs and GCMs. The relation is that an acyclic SCM induces a Markovian GCM as well as its corresponding semi-Markovian GCM. A proof is provided in Appendix A.2.4.

**Proposition 2.4** (Acyclic SCM induces a GCM)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an acyclic SCM. Then,*

- $\pi_{\mathrm{MGCM}}(\mathcal{M}) := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle$ *becomes a Markovian GCM where* $\mathcal{G}^{\mathrm{a}} = \mathrm{graph}(\mathcal{M})$ *and* $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} = \mathrm{dist}(\mathcal{M})$,
- $\pi_{\mathrm{SGCM}}(\mathcal{M}) := \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ *becomes a semi-Markovian GCM where* $\mathcal{G} = \mathrm{obsGraph}(\mathcal{M})$ *and* $\mathbb{P}_{\boldsymbol{\mathcal{Z}}} = \mathrm{obsDist}(\mathcal{M})$.

*Moreover,* $\pi_{\mathrm{MGCM}}(\mathcal{M}) \in \pi_{\mathrm{SGCM}}(\mathcal{M})$ *holds.*

Furthermore, the two definitions of the interventional distributions are compatible. A proof is provided in Appendix A.2.4.

**Proposition 2.5** (Compatibility of interventional distributions for Markovian models)**.** *For any $A \subset \mathcal{I}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{Z}}^A$, the following diagram commutes:*

$$\begin{array}{ccc} \mathrm{SSCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}) & \xrightarrow{\pi_{\mathrm{MGCM}}} & \mathrm{MGCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}) \\ \Big\downarrow{\scriptstyle \mathrm{do}(A,\boldsymbol{a})} & & \Big\downarrow{\scriptstyle \mathrm{do}(A,\boldsymbol{a})} \\ \mathrm{SSCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}) & \xrightarrow[\pi_{\mathrm{MGCM}}]{} & \mathrm{MGCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}) \end{array}$$

*In particular, if $\mathcal{M} \in \mathrm{SSCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}})$ and $\mathcal{M}_1 = \pi_{\mathrm{MGCM}}(\mathcal{M})$, then*

$$\mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}) = \mathrm{doDist}(A, \boldsymbol{a})(\mathcal{M}_1), \quad A \subset \mathcal{I}, \boldsymbol{a} \in \boldsymbol{\mathcal{Z}}^A.$$

*That is, the definitions of the interventional distributions of SCMs and GCMs are compatible.*

Therefore, a Markovian GCM can be considered as a coarsening of a Markovian SCM where only the graphical information is retained and the details of the SF are forgotten.

*Remark* 2.6. Acyclicity is not the most general condition under which the compatibility of the SCMs and the GCMs can be shown. There are other more general conditions that may be imposed on the SCMs, such as the *simplicity* in Bongers et al. [29].

*Remark* 2.7 (Other causal frameworks with graphical representations). Conversely, GCMs (or their variants that make fewer independence assumptions) do not necessarily need to be founded on SCMs (Robins and Richardson [222]). Instead, other generative models such as the finest fully randomized causally interpreted structured tree graph (FFRCISTG; Robins [221]) may well provide the foundation of GCMs while introducing weaker assumptions on the data-generating process. See also Technical Point 6.2 in Hernán and Robins [103].

*Remark* 2.8. If we fix $\mathcal{I}$ and $\boldsymbol{\mathcal{Z}}$ and omit them from the notation, it becomes clearer that a semi-Markovian GCM $\langle \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ is a model that annotates the distribution $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ with a graph $\mathcal{G}$ that is a distilled representation of the data-generating process.

## 2.4 Properties and Estimation of Structural Causal Framework

Here, we describe the properties of the GCMs and the SCMs on which we focus in this dissertation. We also briefly review the (partial) estimation of these models.

### 2.4.1 Statistical Independences in GCMs

In a semi-Markovian GCM $\mathcal{M} = \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$, the graph $\mathcal{G}$ imposes certain constraints on $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$, i.e., there are some known properties of $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ that are shared by all Markovian GCMs in a given semi-Markovian GCM. One such constraint is the following equality constraint on the distribution.

**Definition 2.15** (Topological ADMG factorization)**.** *Let $p_{\boldsymbol{\mathcal{Z}}}$ be a probability density function with respect to a product measure on $\boldsymbol{\mathcal{Z}} = \prod_{v \in \mathcal{I}} \mathcal{Z}^v$, and $\mathcal{G} = \langle \mathcal{I}, \mathfrak{D}, \mathfrak{B} \rangle$ be an ADMG. Also let $\prec$ be a topological ordering over $\mathcal{I}$ with respect to $\mathcal{G}$. Then, $p_{\boldsymbol{\mathcal{Z}}}$ is said to satisfy the* topological ADMG factorization *property with respect to $(\mathcal{G}, \prec)$ (Bhattacharya et al. [24]) if*

$$p_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{z}) = \prod_{v \in \mathcal{I}} p_{\boldsymbol{\mathcal{Z}}}(z^v | \boldsymbol{z}^{\mathrm{mp}(v;\prec)}) \tag{2.3}$$

*holds. In particular, if $\mathcal{G}$ is a DAG, then Equation (2.3) becomes*

$$p_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{z}) = \prod_{v \in \mathcal{I}} p_{\boldsymbol{\mathcal{Z}}}(z^v | \boldsymbol{z}^{\mathrm{pa}(v)}). \tag{2.4}$$

**Proposition 2.6** (Tian and Pearl [273, Corollary 1], Bhattacharya et al. [24])**.** *Let $\mathcal{M} = \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ be a semi-Markovian GCM. Assume $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ has a density function $p_{\boldsymbol{\mathcal{Z}}}$. Let $\prec$ be a topological ordering over $\mathcal{I}$ with respect to $\mathcal{G}$. Then, $p_{\boldsymbol{\mathcal{Z}}}$ satisfies the topological ADMG factorization property with respect to $(\mathcal{G}, \prec)$.*

*Remark* 2.9. In the special case that the ADMG is *uninformative*, i.e., when $\mathcal{G}$ is complete and all edges are bi-directed, Equation (2.3) reduces to the ordinary *chain rule* of probability: $p_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{z}) = \prod_{v \in \mathcal{I}} p_{\boldsymbol{\mathcal{Z}}}(z^v | \boldsymbol{z}^{\{u \in \mathcal{I} \setminus v : u \prec v\}})$, since $\mathrm{mp}(v; \prec) = \{u \in \mathcal{I} \setminus v : u \prec v\}$ in this case. From Equation (2.4), we can see that Definition 2.15 is a generalization of the recursive factorization property (Definition 2.7) to ADMGs. Interestingly, however, unlike the case of Markovian GCMs that are fully characterized by conditional independence relations (Proposition A.1), there are more constraints imposed on the distribution in a semi-Markovian GCM. That is, there are more constraints shared by all Markovian GCMs in a given semi-Markovian GCM.

*Remark* 2.10 (Equality and Inequality Constraints in semi-Markovian GCMs). Some of such constraints are known as *equality constraints*,[27] the constraints imposed by equalities between certain functionals of the density function. Tian and Pearl [273] studied such equality constraints systematically and obtained an algorithm to enumerate the equality constraints given the ADMG. Indeed, Proposition 2.6 follows from one of such equality constraints (Tian and Pearl [273, Corollary 1]). In the case of categorical observed variables, the algorithm of Tian and Pearl [273] has been shown to output all equality constraints (Evans [75]). Richardson et al. [218] refined the approach and provided four characterizations of the equality constraints found in a semi-Markovian GCM, one of which is the criterion provided by Tian and Pearl [273]. The tuples $\langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ satisfying one of the four characterizations are called *nested Markov models* (Richardson et al. [218]). By definition, a

[27] An early example of such constraints is known as the *Verma constraint* (Verma and Pearl [282] and Robins [221]; also see Evans [72, 2.4.1]).

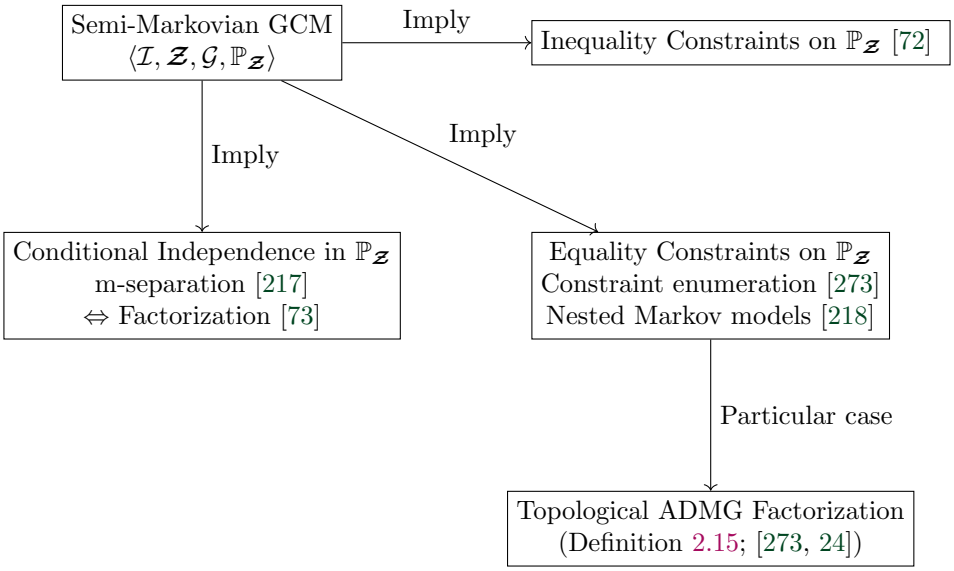



Figure 2.5: Constraints imposed on semi-Markovian GCMs.

nested Markov model $\langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$ imposes no more constraints on $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ than the corresponding semi-Markovian GCM $\langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}} \rangle$, i.e., every semi-Markovian GCM can be seen as a nested Markov model. For more details on nested Markov models, see Shpitser et al. [243] and Richardson et al. [218]. Bhattacharya et al. [24] proposed a simple sufficient condition called the *mb-shieldedness* (*mb* stands for "Markov blanket") under which the topological ADMG factorization captures all the equality constraints. The relation is summarized in Figure 2.5.

Also, quite interestingly, certain inequality constraints are known to entail semi-Markovian GCMs (see, e.g., Evans [72] for details). Evans [74] proposed another layer of equivalence classes based on *marginalized directed acyclic graphs* (mDAG), which is more granular than semi-Markovian GCMs, and showed a limitation of the type of properties that can be retained by ADMG-based GCMs. Forré and Mooij [80] introduced *directed graphs with hyperedges* (HEDGes) that generalize mDAGs and directed mixed graphs and studied their properties, namely the relations of several different versions of *Markov properties* for the corresponding probability distributions (Lauritzen et al. [158] and Lauritzen [160]).

*Remark* 2.11 (Estimation of GCMs). The property in Proposition 2.6 can be used to estimate the graph of a Markovian SCM. Also, the interpretation of a GCM as an induced object of an SCM is important in the estimation of the causal graph.

In this dissertation, we exploit the topological ADMG factorization (Definition 2.15 and Proposition 2.6) as prior knowledge about the data distribution when we have access to an estimator of $\mathcal{G}$. Intuitively, one can expect the knowledge to be useful when the data is so small that it is insufficient for reliable statistical tests of *conditional independence* since the topological ADMG factorization generalizes the recursive factorization property (Definition 2.7) which in turn is equivalent to a series of conditional independence relations (Proposition A.1).

### 2.4.2 Statistical Independences in SCMs

SCMs give rise to a certain family of generative models, called *independent component models*.

**Definition 2.16** (Independent component model; e.g., [49] and [121])**.** *An* independent component model *(ICM) is a tuple* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*, where*

- $\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}$, *and* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ *satisfy Conditions 1, 2, and 4 of Definition 2.1, and*
- $\boldsymbol{F} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Z}}$ *is a measurable map.*

*For a random variable* $\boldsymbol{Z}$ *taking values in* $\boldsymbol{\mathcal{Z}}$*, we say* $\boldsymbol{Z}$ *is* generated by $\mathcal{M}$ *(denoted by* $\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}$*) if there exists a random variable* $\boldsymbol{E}$ *defined on the same probability space such that* $\boldsymbol{E} \sim \mathbb{P}_{\boldsymbol{\mathcal{E}}}$ *and* $\boldsymbol{Z} = \boldsymbol{F}(\boldsymbol{E})$ *almost surely hold. Analogously, if* $\boldsymbol{Z}_i = \boldsymbol{F}(\boldsymbol{E}_i)$ *and* $\{\boldsymbol{E}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}_{\boldsymbol{\mathcal{E}}}$*, we write* $\{\boldsymbol{Z}_i\}_{i=1}^n \overset{i.i.g.}{\leftarrow} \mathcal{M}$*. We call* $\boldsymbol{E}$ *the* independent components *of* $\boldsymbol{Z}$*, and we call* $\boldsymbol{F}$ *the mixing map.*

If an SCM $\mathcal{M}$ has an RSF (e.g., if $\mathcal{M}$ is acyclic; Proposition 2.2), the random variables generated by $\mathcal{M}$ can be considered to be generated from an ICM.

**Proposition 2.7** (SCM as an ICM [137])**.** *Let* $\boldsymbol{F}$ *be an RSF of* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*. Then,* $\mathcal{M}' := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *is an ICM. Moreover,* $(\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}) \Rightarrow (\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}')$ *holds.*

*Proof.* In light of Proposition 2.3, $\boldsymbol{Z} \overset{\text{gen}}{\leftarrow} \mathcal{M}$ implies that $\boldsymbol{Z} = \boldsymbol{F}(\boldsymbol{E})$ almost surely holds for some random variable $\boldsymbol{E} \sim \mathbb{P}_{\boldsymbol{\mathcal{E}}}$ defined on the same probability space. Thus, by definition, $\boldsymbol{Z} \overset{\text{gen}}{\leftarrow} \mathcal{M}'$. □

Moreover, if there are multiple SCMs sharing the same SF, and if each SCM has an RSF, then they can be considered as ICMs sharing the same mixing map.

**Proposition 2.8** (SCMs with identical SFs as ICMs with identical mixing maps)**.** *Let* $K \in \mathbb{N}$ *and let*

$$\begin{aligned} \mathcal{M}_1 &= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}},1} \rangle, \\ &\vdots \\ \mathcal{M}_K &= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}},K} \rangle \end{aligned}$$

*be SCMs.*[28] *Assume that each* $\mathcal{M}_k (k \in [K])$ *has an RSF* $\boldsymbol{F}_k$*. Then, there exists a measurable map* $\boldsymbol{F} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Z}}$ *that is an RSF simultaneously for all* $\mathcal{M}_k (k \in [K])$*. In particular, if we define the ICMs based on this* $\boldsymbol{F}$ *as*

$$\begin{aligned} \mathcal{M}'_1 &= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},1} \rangle, \\ &\vdots \\ \mathcal{M}'_K &= \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},K} \rangle, \end{aligned}$$

*then we have* $(\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}_k) \Rightarrow (\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}'_k)$ *for all* $k \in [K]$*.*

*Proof.* For each $k \in [K]$, let $\boldsymbol{\mathcal{E}}_{\neg}^k$ be the $\mathbb{P}_{\boldsymbol{\mathcal{E}},k}$-negligible set corresponding to $\boldsymbol{F}_k$. Let $\boldsymbol{\mathcal{E}}_k := \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{E}}_{\neg}^k$. Then, for any distinct $k, l \in [K]$, for any $(\boldsymbol{z}, \boldsymbol{e}) \in \boldsymbol{\mathcal{Z}} \times (\boldsymbol{\mathcal{E}}_k \cap \boldsymbol{\mathcal{E}}_l)$, we have $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}) \Rightarrow \boldsymbol{F}_k(\boldsymbol{e}) = \boldsymbol{z} = \boldsymbol{F}_l(\boldsymbol{e})$. Thus, we can define

$$\boldsymbol{F}(\boldsymbol{e}) = \begin{cases} \boldsymbol{F}_k(\boldsymbol{e}) & \text{if } \boldsymbol{e} \in \boldsymbol{\mathcal{E}}_{\boldsymbol{f}} \cap \boldsymbol{\mathcal{E}}_k, \\ \tilde{\boldsymbol{F}}(\boldsymbol{e}) & \text{if } \boldsymbol{e} \notin \boldsymbol{\mathcal{E}}_{\boldsymbol{f}} \cap \left( \bigcup_{k \in [K]} \boldsymbol{\mathcal{E}}_k \right), \end{cases}$$

where $\boldsymbol{\mathcal{E}}_{\boldsymbol{f}} := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e})\}$ and $\tilde{\boldsymbol{F}} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Z}}$ is an arbitrary measurable map. Then, for any $(\boldsymbol{z}, \boldsymbol{e}) \in \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}_k$ such that $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e})$, we have $\boldsymbol{z} = \boldsymbol{F}_k(\boldsymbol{e}) = \boldsymbol{F}(\boldsymbol{e})$. Since $\mathbb{P}_{\boldsymbol{\mathcal{E}},k}(\boldsymbol{\mathcal{E}}_k) = 1$ and $\boldsymbol{F}$ is measurable, by definition, $\boldsymbol{F}$ is an RSF of $\mathcal{M}_k$. The last half of the assertion follows immediately from Proposition 2.7. □

As a result, under certain conditions, the RSF can be estimated by using the methods of *independent component analysis* (ICA).

[28] The acyclicity requirement can be relaxed to the requirement that each SCM has an RSF.

**Proposition 2.9** (Identifiability of ICM; [124, Theorem 1])**.** *Let $d, K \in \mathbb{N}$. Let*

$$\begin{aligned}\mathcal{M}_1 &= \langle [d], [d], \mathbb{R}^d, \mathbb{R}^d, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},1} \rangle, \\ &\vdots \\ \mathcal{M}_K &= \langle [d], [d], \mathbb{R}^d, \mathbb{R}^d, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},K} \rangle\end{aligned}$$

*be ICMs. Moreover, assume the following.*

- *$\mathbb{P}_{\boldsymbol{\mathcal{E}},1}, \dots, \mathbb{P}_{\boldsymbol{\mathcal{E}},K}$ have density functions $p_{\boldsymbol{\mathcal{E}},1}, \dots, p_{\boldsymbol{\mathcal{E}},K}$.*
- *$p_{\boldsymbol{\mathcal{E}},k}$ is sufficiently smooth ($k \in [K]$),*
- *$\boldsymbol{F}$ is a $C^2$-diffeomorphism from $\mathbb{R}^d$ to itself,*
- *for any $\boldsymbol{e} \in \mathbb{R}^d$, there exist distinct values $k_0, \dots, k_{2d} \in [K]$ such that*

$$\{w(\boldsymbol{e}; k_j) - w(\boldsymbol{e}; k_0)\}_{j \in [2d]}$$

*are linearly independent, where*

$$w(\boldsymbol{e}; k) := \left( \frac{\partial \log p^1_{\boldsymbol{\mathcal{E}},k}(\boldsymbol{e}^1)}{\partial \boldsymbol{e}^1}, \dots, \frac{\partial \log p^d_{\boldsymbol{\mathcal{E}},k}(\boldsymbol{e}^d)}{\partial \boldsymbol{e}^d}, \frac{\partial^2 \log p^1_{\boldsymbol{\mathcal{E}},k}(\boldsymbol{e}^1)}{\partial (\boldsymbol{e}^1)^2}, \dots, \frac{\partial^2 \log p^d_{\boldsymbol{\mathcal{E}},k}(\boldsymbol{e}^d)}{\partial (\boldsymbol{e}^d)^2} \right).$$

*Then, there exists an algorithm $\mathcal{A}$ such that given independent and identically generated data sets $\{\boldsymbol{Z}^k_i\}_{i=1}^{n_k} \overset{i.i.g.}{\leftarrow} \mathcal{M}_k (k \in [K])$, $\mathcal{A}\left(\{\boldsymbol{Z}^1_i\}_{i=1}^{n_1}, \dots, \{\boldsymbol{Z}^K_i\}_{i=1}^{n_K}\right)$ consistently estimates $\boldsymbol{F}^{-1}$ up to component-wise invertible transformations.*

Proposition 2.9 implies that it is possible (under additional assumptions) to estimate $\boldsymbol{F}$ by employing the data from multiple ICMs sharing the same $\boldsymbol{F}$, thereby providing a sufficient condition under which the RSF $\boldsymbol{F}$ of an SCM is estimable. That is, if we have multiple SCMs $\mathcal{M}_1, \dots, \mathcal{M}_K$ that share the same SFs (and hence the same RSFs), and if the models satisfy the additional technical assumptions in Proposition 2.9, then the RSF $\boldsymbol{F}$ is estimable.

*Remark* 2.12 (Methods of independent component analysis). On the other hand, it is well-known that correctly estimating $\boldsymbol{F}$ of an ICM $\mathcal{M}$ from a single sample $\{\boldsymbol{Z}_i\}_{i=1}^n \overset{\text{i.i.g.}}{\leftarrow} \mathcal{M}$ is impossible in general (Hyvärinen and Pajunen [120]).

*Remark* 2.13 (Estimation of SCMs). Proposition 2.9 indicates that, under certain conditions, we can estimate $\boldsymbol{F}$ that is partial knowledge of the SF $\boldsymbol{f}$. In general, even if we successfully estimate $\boldsymbol{F}$, it is not always possible to recover $\boldsymbol{f}$ from $\boldsymbol{F}$. For example, consider

$$\begin{cases} X = \mathbb{1}[e_1 > 0], \\ Y = X^2 + e_2, \end{cases} \quad \text{and} \quad \begin{cases} X = \mathbb{1}[e_1 > 0], \\ Y = X + e_2. \end{cases}$$

Although this is an artificial example, we see that two different SFs can yield the same RSF (imagine, e.g., $e_1$ is the body temperature, $X$ is the amount of fever reducer you take, $e_2$ is the base metabolism, and $Y$ is the appetite). However, in some special cases, e.g., if $\boldsymbol{f}$ is linear, such a recovery may be possible. See, e.g., Shimizu et al. [240].

In this dissertation, we exploit an estimated RSF, $\boldsymbol{F}$, as prior knowledge about an independence structure in the data-generating process. Intuitively, we can expect that an RSF estimated in one environment can be applied to another when there is no active intervention taking place; even if

we cannot estimate the RSF in one environment, we may be able to take advantage of the *stability* which is a salient characteristic of causality (Woodward [295, Chapter 6]).[29]

## 2.5 Problem Setup and Approach

In this section, we describe the general problem we tackle, namely the small-data learning problem. The general approach of the dissertation to the problem, namely *data augmentation*, is also explained.

### 2.5.1 Problem: Small-data Learning

The problem we study in this dissertation is the *small-data learning problem*. Despite the rapid progress in the methodology of machine learning, learning from small data remains an important challenge in various application fields. When data is limited in quantity, it is essential to incorporate appropriate prior knowledge about the property of the data distribution for learning an accurate predictor.

**Supervised learning problem.** Let us formulate the learning problem considered in this dissertation. Consider random variables $Z = (X, Y)$ taking values in $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is called the *input space* and $\mathcal{Y}$ the *label space*. Let $\mathbb{P}_{\mathcal{Z}}$ be a data distribution over $\mathcal{Z}$.

Suppose an i.i.d. sample $\mathcal{D} = \{Z_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathcal{Z}}$ is given. The data set $\mathcal{D}$ is called the *training data set*. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a set of predictors that take $X$ as input and output label $Y$. The set $\mathcal{H}$ is called a *hypothesis class*, and each element $h \in \mathcal{H}$ is called a *hypothesis*. Let $\ell : \mathcal{H} \times \boldsymbol{\mathcal{Z}} \to \mathbb{R}_{\geq 0}$ be a *loss function*. Our goal is to find a predictor $h \in \mathcal{H}$ for which the *risk*, or the *expected loss*, defined by $\mathcal{R}(h) := \mathbb{E}[\ell(h, Z)]$, is small, where $\mathbb{E}$ is the expectation with respect to $Z \sim \mathbb{P}_{\mathcal{Z}}$. The process of finding such a good $h$, in this context, is called *learning*.

**Empirical risk minimization.** The prototypical approach to this learning problem is *empirical risk minimization* (ERM; Vapnik [278]). In ERM, one selects the hypothesis $h \in \mathcal{H}$ that minimizes the *empirical risk* $\hat{\mathcal{R}}_{\mathrm{ERM}}(h) := \hat{\mathbb{E}}[\ell(h, Z)]$, where $\hat{\mathbb{E}}$ is the empirical average operator with respect to the data set $\mathcal{D}$ defined by $\hat{\mathbb{E}}(g) := \frac{1}{n}\sum_{i=1}^n g(Z_i)$.

**Small-data learning problem.** The *small-data learning* problem, in this context, refers to the case where $n$ is small. Typical machine learning methods based on ERM, roughly speaking, can be justified by the laws of large numbers (see, e.g., Vapnik [279, 2.3.1]);[30] given a large data set (i.e., large $n$), we can expect $\hat{\mathcal{R}}_{\mathrm{ERM}}(h)$ to be close to $\mathcal{R}(h)$ uniformly over the hypothesis class $\mathcal{H}$. Therefore if $n$ is large, the minimizer of $\hat{\mathcal{R}}_{\mathrm{ERM}}$ in $\mathcal{H}$ can be expected to have small $\mathcal{R}(h)$. On the other hand, when $n$ is small, we are often left with no ground to rely on.

Given a small data set, the learning machine may be unable to extract useful patterns for making accurate predictions. In this case, even if we successfully train the model to make accurate predictions within the training data set, the predictor may not perform well when it is evaluated by $\mathcal{R}(h)$: a situation referred to as *overfitting* ([279, p.124]). Thus, when the data is small, we need an additional source of information to complement the knowledge that can be extracted solely from the data. In the context of statistical machine learning, such additional information is referred to as

[29] Note, however, the degree to which the stability of a law can be believed may depend on the domain of interest. See, e.g., Woodward [296].

[30] While it is probably impossible to tell whether the statistical learning algorithms achieve good performance precisely because of the laws of large numbers, it remains an important design principle for developing the methods of statistical machine learning. In this dissertation, we also provide theoretical justifications of the proposed methods based on the statistical learning theory (e.g., [278, 280, 236, 184]).

*prior knowledge*. For example, *regularization* techniques (see, e.g., Shalev-Shwartz and Ben-David [236, Chapter 13])introduce prior knowledge that essentially restricts $\mathcal{H}$ to be "small", thereby reducing the risk of overfitting.

In this dissertation, we consider how causal knowledge captured by structural causal models could be used as prior knowledge in small-data learning problems. Intuitively, one salient characteristic of causal knowledge is re-usability. Causal knowledge is generally believed to be invariant unless we actively intervene in the data-generating process and that it is valid in similar systems that are different from the system in which we acquired the knowledge (see, e.g., Glennan [86] and Woodward [296]). Therefore, even when the data is scarce in a specific environment in which we wish to train a machine learning model, known or acquired causal knowledge may provide a reliable source of information.

### 2.5.2 Problem Taxonomy: Two Estimable Causal Model Layers

As we have seen in Section 2.3.3, SCMs and GCMs have a hierarchical relation, where the GCMs are a coarser description of the causal relations than the SCMs. Besides the two, there is another level of modeling in the SCF that has been discussed in the literature, namely the *physical causal models* that use differential equations to represent the causal mechanisms (Mooij et al. [186]). Adding to this hierarchy the *statistical models* that do not retain the causal knowledge, i.e., the models concerning only the observational distributions, we have a hierarchy of models in the SCF: the physical models, the SCMs, the GCMs, and the statistical models, from the finest level to the coarsest level (Peters et al. [208, Table 1.1]).

In this hierarchy, the GCMs and the SCMs form the two shallowest levels where some information of the models can be estimated from observational data, namely the causal graphs and the RSFs (e.g., [88, 137, 124]) under certain conditions. Therefore, in this dissertation, we discuss two cases, namely the case that the GCM is estimable (or known) and the case that the SCM is estimable. Specifically, we consider how the estimated causal models can be used to aid learning in small-data learning problems.

**When GCM is known or estimable.** In this dissertation, we consider exploiting the implication of a GCM on the data distribution $\mathbb{P}_{\boldsymbol{Z}}$. We consider a situation where the causal graph is either known thanks to domain knowledge or has been estimated from data. The question is how such a causal graph can be used for enhancing supervised learning.

**Problem 2.1** (Causal-Graph-Informed Learning Problem; Tentative Version)**.** *Let* $\mathcal{M} := \langle \mathcal{I}, \boldsymbol{\mathcal{Z}}, \mathcal{G}, \mathbb{P}_{\boldsymbol{Z}} \rangle$ *be a semi-Markovian GCM, and assume that* $\mathbb{P}_{\boldsymbol{Z}}$ *has a density function* $p_{\boldsymbol{Z}}$*. Then, Proposition 2.6 implies that, for any topological ordering* $\prec$ *of* $\mathcal{I}$*, the density function* $p_{\boldsymbol{Z}}$ *satisfies the topological ADMG factorization property (Definition 2.15) with respect to* $(\mathcal{G}, \prec)$*. Now, given* $\{\boldsymbol{Z}_i\}_{i=1}^n \overset{i.i.d.}{\sim} p_{\boldsymbol{Z}}$ *and an ADMG* $\hat{\mathcal{G}}$ *that is an estimator of* $\mathcal{G}$ *as well as a topological ordering of* $\mathcal{I}$ *with respect to* $\hat{\mathcal{G}}$*, find a predictor* $\hat{h} \in \mathcal{H}$ *for which the risk* $\mathcal{R}(\hat{h}) := \mathbb{E}[\ell(\hat{h}, \boldsymbol{Z})]$ *is small, where* $\mathbb{E}$ *denotes the expectation with respect to* $\boldsymbol{Z} \sim \mathbb{P}_{\boldsymbol{Z}}$*.*

In Chapter 3, we tackle a slightly generalized version of Problem 2.1 by assuming that the assertion of Proposition 2.6 is satisfied instead of assuming the existence of a GCM.

**When SCM is estimable.** Even if it is difficult to estimate an SCM (i.e., estimating $\boldsymbol{f}$) from the small data in the target problem domain, one may be able to estimate some information of an SCM (e.g., its RSF, $\boldsymbol{F}$) from the data of other relevant problem domains and apply the knowledge in the target domain, since causal knowledge is believed to be stable and invariant across a range of different environments unless actively intervened in ([86, 296, 119]). For example, in medical

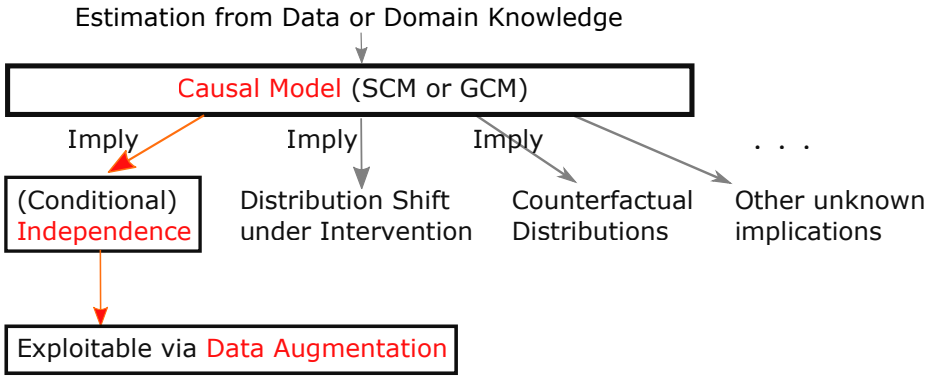



Figure 2.6: General idea of this dissertation: exploiting the structural causal models via data augmentation.

record analysis for disease risk prediction [304], it can be reasonable to assume that the pathological mechanism is common across regions or generations. Such a hidden structure, ideally, may be exploited to obtain accurate predictors for under-investigated regions or new generations, where the data may be scarce.

In light of Proposition 2.9, if we have a large amount of data from multiple other SCMs sharing the same SFs as the one underlying the target data distribution of our interest, it may be possible to estimate $\boldsymbol{F}$ and somehow use its estimator as prior knowledge to facilitate the learning from few data from the target distribution. This situation corresponds to the *domain adaptation* problem in statistical machine learning [19]: learn a good predictor for a target *domain* of interest given data from other domains. We can formulate the problem we consider in this dissertation as follows:

**Problem 2.2** (Causal Mechanism Transfer Problem; Tentative Version)**.** *Let $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_K$ be acyclic SCMs sharing the same SFs. Given $\{\boldsymbol{Z}_i^k\}_{i=1}^{n_k} \overset{i.i.g.}{\leftarrow} \mathcal{M}_k (k \in \{0\} \cup [K])$ where $n_0$ is small but $n_1, \ldots, n_K$ are large, find a predictor $\hat{h} \in \mathcal{H}$ for which the risk $\mathcal{R}(\hat{h}) := \mathbb{E}_0[\ell(\hat{h}, \boldsymbol{Z})]$ is small, where $\mathbb{E}_0$ denotes the expectation with respect to $\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}_0$.*

In Chapter 4, we tackle a slightly generalized version of Problem 2.2 by assuming the existence of ICMs instead of SCMs from which they are derived.

### 2.5.3 Approach: Data Augmentation

The general idea of this dissertation in approaching Problems 2.1 and 2.2 is to design *data augmentation* procedures that reflect the statistical independence relations implied by the estimated causal models (Figure 2.6).

**Data augmentation.** Data augmentation is a collective term to refer to the methodologies that synthesize data based on some original samples (e.g., Shorten and Khoshgoftaar [242]). By creating additional data and training a learning model using it, one can effectively introduce prior knowledge into the learning process. Typical examples can be found in the field of computer vision and natural language processing; for example, by applying some operations on the image data which do not change the meaning of the image, such as a small rotation or adding a small noise, the trained model is expected to learn what are relevant patterns in the input data. Data augmentation has the virtue of model-independence: it can be easily combined with virtually any machine learning method because the interface is the *data* itself, which is a central component in modern machine learning [242]. On the other hand, data augmentation typically introduces additional computation costs, which can be problematic when the original data set is large. However, since we are concerned

with the small-data learning problem, we expect that this problem will not be restrictive in the scope of this dissertation.

**Our approach.** In our approach to both Problems 2.1 and 2.2, the basic idea is based on the following observation: if there are independent random vectors, new random vectors created by scrambling the pairings are equally likely. For example, if $\mathcal{D} = \{(X_i, Y_i)\}_{i=1,2} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_Y$, then the new combinations $\tilde{\mathcal{D}} = \{(X_1, Y_2), (X_2, Y_1)\}$ follow the same distribution as $\mathcal{D}$. We use this and similar ideas to design the data augmentation methods in the subsequent chapters. Since we use $\mathcal{D} \cup \tilde{\mathcal{D}}$ as the new data set instead of $\tilde{\mathcal{D}}$ only, the statistical properties of the inference based on such data is expected to be different from the ones using only the original data $\mathcal{D}$. Therefore, we also provide theoretical analyses for each of the proposed methods to understand their characteristics as well as the statistical benefits they bring.

For Problem 2.1, we take advantage of the topological ADMG factorization structure. In the case of DAGs, this corresponds to exploiting the conditional independence relations (Definition 2.15 and Proposition A.1). We design a data augmentation method that reflects the factorization structure of the data distribution.

For Problem 2.2, we first estimate the RSF, which is the only commonality among different environments from which data sets are sampled. Then, taking advantage of the ICM structure (Definition 2.16), we use the RSF as a feature extractor that can extract independent components from the observed data.[31] We design a data augmentation method to reflect the independence structure of the independent component distribution.

## 2.6 Causal Machine Learning

In this section, we review the researches of *causal machine learning* (e.g., [208, 230, 233]): the intersection of statistical causal modeling and statistical machine learning. The review is intended to be provided in the context of the present dissertation. The readers who are familiar with this field may well skip this section.

The literature in the interaction between the causal frameworks and machine learning can be roughly divided into three categories, namely (i) causality *for* machine learning, (ii) causality *by* machine learning, and (iii) causality *in* machine learning. The present dissertation fits into the category of *causality for machine learning* in this trichotomy.

### 2.6.1 Causality for Machine Learning

Various attempts to integrate causal concepts into machine learning have been put forward. The distinct usage includes: (I) as a conceptual guide of methodological designs, (II) as a tool to analyze specific causal models that appear in specific application fields, (III) as a theoretical foundation of *invariance*-guided learning, (IV) as feature selection criteria, (V) as regularization and model selection, and (VI) as model architecture design.

**(I) As a guiding principle: intuition for the *relevance of information* and *sparsity of distribution shifts*.** The seminal paper, Schölkopf et al. [232], discussed how the difference in the types of data-generating process, namely *causal* ($X \to Y$) and *anti-causal* ($X \leftarrow Y$) learning problems, may explain the general difficulty of different problem setups of machine learning. The concept of the anti-causal scenario has been used to motivate some *unsupervised transfer learning*

[31] In this approach, *identifiability* plays a crucial role: the reason why we can believe that the estimated RSF is applicable to the data distribution of interest is that, under the identifiability assumptions, we can obtain a function that is close to the correct RSF.

methods [313, 311, 91, 90], where the *source domain data* are *labeled* (i.e., paired samples of $(X, Y)$) and the *target domain data* are *unlabeled* (i.e., samples of $X$ without the corresponding $Y$). Concretely, Zhang et al. [313] and Zhang et al. [311] and Gong et al. [91, 90] justified their parametric distribution shift assumptions or the parameter estimation procedure; their model selection criterion is based on the distribution matching of the marginal $p_X$ between the source domain and the target domain (which can be performed without access to labeled target domain data), and its justification argument is that the distribution of $X$ is likely to contain some information of $p(Y|X)$ in the anti-causal scenario ($X \leftarrow Y$). Also, the same argument for the anti-causal scenario is used to justify the modeling of the distribution shifts of $p(X|Y)$ and of $p(Y)$ separately (e.g., Zhang et al. [313]).

**(II) As a tool to analyze specific problem instances.** In some specific application fields where the causal graph can be drawn, specialized methodologies have been developed based on the knowledge encoded in the graph. Among the early examples is the *half-sibling regression* in the exoplanet search (Schölkopf et al. [231]), where the specific causal structure of the data acquisition was used to derive and justify the regression analysis method. Another example is the instance weight estimation for episodic reinforcement learning, where methods to perform *state simplification* based on the causal graphs have been proposed (Bottou et al. [32] and Peters et al. [208, Section 8.2]). Pitis et al. [211] proposed a method to enhance the sample efficiency in reinforcement learning by a procedure to exchange the realizations of the variables within the (conditionally) disconnected components in the causal graph of the *Markov decision process* of specific reinforcement learning instances.

**(III) As the foundation of *invariance* or *stability*.** Rojas-Carulla et al. [223] developed a *domain generalization* method where the exploited conjecture is that if the conditional distribution $p(Y|\boldsymbol{X}_S)$ is invariant among multiple source distributions, it may be invariant in the target distribution. In order to justify this assumption, the *stability* of causal mechanisms was used as a guiding principle. Arjovsky et al. [8] proposed *invariant risk minimization* (IRM) for the *out-of-distribution generalization* problem. The IRM approach tries to learn a feature extractor that makes the optimal predictor invariant across domains, and its theoretical validity was argued based on SCMs. This line of work is the most closely related to this dissertation in that we exploit the stability of causal mechanisms as conceptual support. On the other hand, our work is relatively distinct from this line of work in that our emphasis is on how the knowledge of stable causal mechanisms may facilitate learning in a small-data regime, whereas this line of work targets distribution-shift problems.

**(IV-i) Variable selection in a single-distribution setting.** When the causal graph is known or when it has been estimated, one of the classical ideas for leveraging the knowledge is *feature selection* (e.g., Yu et al. [307]). Concretely, the graphical knowledge can be used to select the set of variables that are informative for making predictions, namely, the *Markov blanket* or the *Markov boundary* [274].

**(IV-ii) Variable selection in a distribution-shift setting.** Another line of research is concerned with making predictions under distribution shift by leveraging feature selection based on causal background knowledge or causal discovery. Magliacane et al. [176] considered the case that a distribution shift is due to intervention in some variables, and they proposed a method to perform *domain adaptation* (e.g., [19]) by identifying a set of variables that is likely to perform well regardless of the intervention. Rojas-Carulla et al. [223] assumed that if the conditional distribution of the predicted variable given some subset of features is invariant across different distributions, this conditional distribution is the same in the *target distribution* for which one wants to make good

predictions. Then they leveraged it to find the set of variables for which the relation to the target variable does not change.

**(V) Regularization and model selection.** Kyono and van der Schaar [155] proposed a model selection criterion that can reflect the structure of a causal graph of a Markovian GCM. The goal of Kyono and van der Schaar [155] is *domain generalization* and *out-of-distribution prediction*, i.e., making good predictions under a distribution shift without access to any samples from the target distribution or making good predictions for the data that are outside the support of the training data distribution. To achieve it, given a DAG as prior knowledge and assuming its validity in the testing domain, Kyono and van der Schaar [155] proposed to first modify the graph so that the edges coming out of the target variable are removed. Then, to score the predictor model candidates, their method generates a data set whose predicted variables are replaced by the predictions of the model and computes the *Bayes Information Criterion* (BIC) that evaluates the fitness of the modified DAG structure to the generated data set. Another approach for using the background knowledge of a causal graph of a Markovian GCM is the *CASTLE regularization* (Kyono et al. [156]). CASTLE regularization introduces a regularization term to induce sparsity and acyclicity in the structure of a neural network that is the predictor hypothesis class. The method imposes a reconstruction loss using the internal layers of the predictor implemented by a neural network under a DAG constraint.

**(VI) As model architecture design.** Another natural approach to exploiting the prior knowledge of a causal graph, when it has no bi-directed edges, is to build a *Bayesian network* (BN) model according to the graphical structure (e.g., [174]) by specifying the conditional distributions appearing in the Markov factorization (Definition 2.7). This approach has the limitation that it inevitably restricts the modeling choice, e.g., the constructed predictor is a generative model as opposed to a discriminative model [236, Chapter 24]. For ADMGs, in some other special cases, canonical parametrization of the joint distributions conforming to the causal graphs has been proposed. In the multivariate binary case (i.e., $\mathcal{Z}^v = \{0, 1\}$ $(v \in \mathcal{I})$), Evans and Richardson [73] and Evans and Richardson [76] provided a smooth parametrization of the set of distributions that satisfy the equality constraints according to a given ADMG (Remark 2.10). Complementarily, for the case of $\mathcal{Z}^v = \mathbb{R}$ $(v \in \mathcal{I})$, Silva and Ghahramani [249] and Silva et al. [248] proposed parametrizations of certain families of distributions satisfying the equality constraints by using probit models and cumulative distribution networks, respectively, but they impose additional constraints induced by their parametric structure.

**This dissertation.** The present dissertation adds a distinct form of interaction between statistical causal models and statistical machine learning: using the knowledge encoded in statistical causal models to facilitate learning from small data in prediction problems. The structural causal framework was originally developed to enable causal inference, but as we have seen in Section 2.4, the causal structures can have tangible consequences in the observational distributions. Our approach falls into this category: we consider leveraging such additional structures captured by the models in the structural causal framework to enhance statistical machine learning. Thanks to the nature of data augmentation that it uses *data* as the interface to other components of a machine learning system, the approach tends to result in generic methods that are independent of the modeling choices, such as the predictor hypothesis class. Due to this characteristic, our approach tends to be conceptually orthogonal to other approaches listed in this section, i.e., the proposed methods can be easily combined with other approaches to integrating causal concepts into machine learning systems.

### 2.6.2 Causality by Machine Learning

This category is mainly concerned with (I) aiding the estimation of various causal quantities and (II) performing causal inference using machine learning methods. Another interesting line of work also considered (III) automated construction of the variables in the causal models by developing representation learning techniques.

**(I-i) Estimating causal parameters.** This direction of research aims to perform the tasks of causal inference, e.g., estimating the *conditional average treatment effect* or estimating the *structural parameters*, with the help of various function models used in machine learning, such as *decision trees* and *deep neural networks* [236]. In order to optimize the estimation methods in this context, modifications to the algorithms have been proposed [285, 9], along with how to appropriately modify the estimation procedure to achieve a high sample efficiency of parameter estimation when machine learning models are used to estimate *nuisance parameters* (double/debiased machine learning; Chernozhukov et al. [43]).

**(I-ii) Causal discovery.** Various methods of causal discovery have been developed in computer science, e.g., constraint-based methods (Glymour et al. [88]) and score-based methods (Glymour et al. [88] and Huang et al. [114]). See Glymour et al. [88] for a review. The developments in the field of *independent component analysis* have yielded a series of methodologies for causal discovery (e.g., Shimizu et al. [239], Peters et al. [209], Peters and Schölkopf [210], and Monti et al. [185] and Hyvärinen et al. [124]). A series of other methods have been proposed based on the restrictions of function classes of the SFs (e.g., [113, 312, 238]; see also Wiedermann and von Eye [291, Chapter II]). More approaches have been proposed from other perspectives such as information geometry (Janzing et al. [132]), algorithmic complexity (Janzing and Schölkopf [133]), and finding statistical patterns in the joint distributions (Mooij et al. [188] and Mooij et al. [187]).

**(II) Optimizing interventions.** Another line of work considered the problem of selecting the optimal intervention to maximize the expected reward by formulating it as a multi-armed bandit problem with an underlying causal structure (e.g., Lattimore et al. [157] and Lee and Bareinboim [164]).

**(III) Causal feature learning.** Another interesting research direction is *causal feature learning*, whose goal is to train a feature extractor that can extract the *macroscopic* causally-interpretable variables from *microscopic* variables (Chalupka et al. [38, 39, 37]).

### 2.6.3 Causality in Machine Learning

The last category considers the causality-based analysis of machine learning systems seen as causal systems. Such a viewpoint has been fruited in the methods for (I) explanation and algorithmic recourse and (II) promoting fairness by detecting and correcting the bias. Also, (III) machine learning systems that interact with the environment have obvious connections to the (causal) concept of interventions.

**(I) Explanation and algorithmic recourse.** In the philosophy of sciences, the intimate connections between causation and explanation have been discussed (e.g., Lipton [168] and Woodward [295]), and in its simplest form, a causal model of explanation maintains that to explain some phenomenon is to give some information about its causes. A field that is highly relevant to this perspective is that of *explainable artificial intelligence* (XAI) whose goal is to improve the interpretability and the accountability of artificial intelligent systems, and in particular, black-box models

(e.g., Adadi and Berrada [2]). Concretely, *counterfactual explanation* methods aim at providing an interpretation of a prediction based on hypothetical outputs of the model when a small *actionable* change is applied to the input (see, e.g., Verma et al. [281]). A task related to explanation is that of *algorithmic recourse*, the task of providing actionable recommendations to individuals for obtaining a more favorable prediction (e.g., Karimi et al. [138, 139]).

**(II) Fairness.** Causality has been an important notion in the studies of *fairness* and *bias* of the data or the trained predictors. Various methods for detecting, measuring, and correcting (un)fairness have been developed (e.g., [153, 300, 44, 301, 299, 284, 46]).

**(III) Policy evaluation and optimization in reinforcement learning and recommendation systems.** Bareinboim et al. [14] explored the connection between causal models with unobserved confounders and reinforcement learning. Zhang and Bareinboim [310] proposed off-policy evaluation methods for multi-armed bandit problems by identifying the causal structure tied with the problem setup. *Recommendation systems* powered by machine learning also interact with the environment. The primary goal of such systems is to improve their user experience by (softly) *making* the users select certain items (e.g., [30, 287]) instead of predicting the user responses in a natural environment. Based on such a view, the policy evaluation methods that take into account the difference between the distributions with and without the recommendation system have been developed (e.g., Bonner and Vasile [30]). These lines of work typically employ the POF to formulate the problems since their focus is on the evaluation or optimization of interventions, and the detailed mechanisms are not of primary interest in this context.

## 2.7 Conclusion

Statistical causal models consider additional structure in the random variables that are not necessarily reflected in their joint distribution. In this chapter, we introduced two representative model classes in the structural causal framework, namely the SCMs and the GCMs, which reside in two different layers in the hierarchy of the framework. SCMs devise the concept of SFs to capture deterministic relations satisfied by the random variables, and GCMs use coarser graphical representations to capture the dependency structures among the random variables. By taking into account such additional structures, they distinguish the random variables even if they follow the same probability distribution. In this dissertation, we treat causal models as the tools to capture the informative and stable structure of the data-generating processes, which can allow us to go beyond the standard statistical machine learning methodologies whose algorithms are designed based on the concept of joint probability distributions.

From the next chapter, let us see how concretely the knowledge of the underlying causal system, either provided a priori by domain knowledge or estimated a posteriori from data, can be used to aid the supervised machine learning methodology.

# Chapter 3

# When Graphical Causal Model is Known or Estimable: Causal-graph Data Augmentation

As we have seen in Chapter 2, causal graphs (CGs) play various important roles in the structural causal framework. One important aspect of the CGs is that they imply certain constraints that should be satisfied by the joint distributions, such as conditional independence relations. Thus, if a CG is known or has been previously estimated in a relevant problem domain, and if we can design an appropriate method, such prior knowledge of the data distribution could be used to enhance statistical machine learning as an information source to complement the data. In this chapter, we design a data augmentation method to directly take advantage of the constraints implied by the CG in training a prediction model.

## 3.1 Overview

*Causal graphs* (CGs; [204]) are compact representations of the knowledge of data-generating processes. Such a CG is sometimes provided by domain experts in some problem instances, e.g., in biology [227] or sociology [240]. Otherwise, it may also be learned from data using the statistical causal discovery methods developed over the last decades [253, 204, 45, 239, 210, 208]. Once a CG is obtained, it can be used to infer the conditional independence (CI) relations that the data distribution should satisfy [204].

### 3.1.1 Motivation

The CI relations encoded in the CG could be strong prior knowledge for predictive tasks in machine learning, e.g., regression or classification, especially in the *small-data* regime where data alone may be insufficient to witness the CI relations [253, Section 5.2.2]. However, it is not trivial how the CI relations should be directly incorporated into general supervised learning methods.

In previous research, methods that leverage the causality for feature selection have been proposed (see, e.g., Yu et al. [307] for a review). However, most of them are based on the notion of the *Markov blanket* or the *Markov boundary* [274]. As a result, they only take into account partial information of all that is encoded in a CG, since a CG often entails more constraints on the data distribution than the specifications of Markov blankets or a Markov boundary [217]. Another approach to exploiting the prior knowledge of a CG is to build a *Bayesian network* (BN) model according to

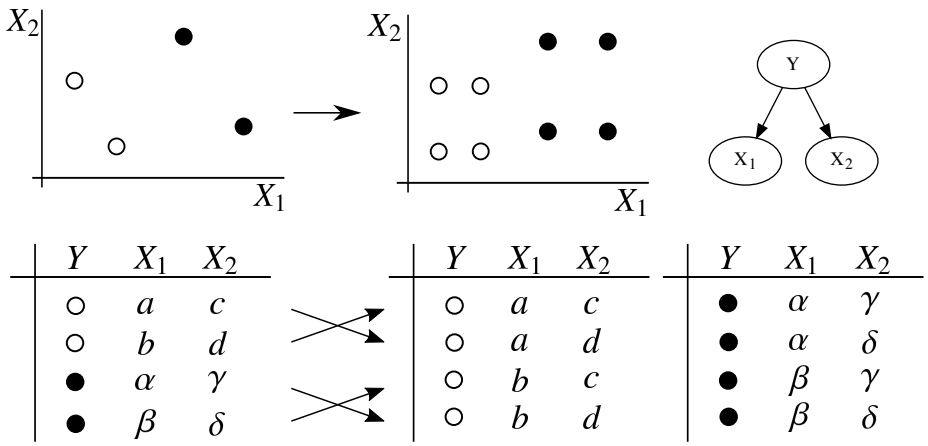



Figure 3.1: Visualization of the basic idea of the proposed method for the trivariate case $X_1 \leftarrow Y \rightarrow X_2$. In this case, the CI $X_1 \perp\!\!\!\perp X_2 \mid Y$ is known to hold. One way to use this knowledge via data augmentation is to group the data according to $Y$ and then to shuffle $X_1$ and $X_2$ within each group. Our method extends this idea to more general graphs.

the CG structure (e.g., [174]). However, constructing the predictors by employing BNs as the framework entails a specific modeling choice, e.g., it constructs a *generative* model as opposed to a *discriminative* model [236, Chapter 24], precluding the choice of some flexible and effective models such as tree-based predictors [81] and neural networks [92] that may be preferred in the application area of one's interest.

### 3.1.2 Idea

In this chapter, we propose a model-agnostic method to incorporate the CI relations implied by CGs directly into supervised learning via data augmentation. To illustrate our idea, let us consider the following trivariate case.

**Illustrative example: trivariate case (Figure 3.1).** Suppose we want to predict a binary variable $Y$ from $(X_1, X_2)$. If the random variables have the underlying CG $X_1 \leftarrow Y \rightarrow X_2$, then the CI $X_1 \perp\!\!\!\perp X_2 \mid Y$ is known to hold [204]. If we know this relation, a natural idea of data augmentation is to stratify the sample by $Y$ and then to take all combinations of $X_1$ and $X_2$ within each stratum.

In this trivariate example, it is straightforward to derive such a plausible data augmentation procedure to incorporate the CI relations since the relation $X_1 \perp\!\!\!\perp X_2 \mid Y$ involves all three variables. On the other hand, deriving such a procedure for general graphs is not straightforward as they may encode a multitude of CI relations each of which may involve only a subset of all variables.

### 3.1.3 Contributions

Our contributions can be summarized as follows.

1. We propose a method to augment data based on the prior knowledge expressed as CGs, assuming that an estimated CG is available.

2. We theoretically justify the proposed method via an excess risk bound based on the Rademacher complexity [15]. The bound indicates that the proposed method suppresses overfitting at the cost of introducing additional complexity and bias into the problem.

3. We empirically show that the proposed method yields consistent performance improvements especially in the small-data regime, through experiments using real-world data with CGs obtained from the domain knowledge.

## 3.2 Problem Setup and Main Assumption

In this section, we formally state the problem setup, the goal, and the main assumption to be exploited in our proposed method.

**Basic notation.** For $N, M \in \mathbb{N}$ with $N \leq M$, define $[N : M] := \{N, N+1, \dots, M\}$. To simplify the notation, we let $[0] = \emptyset$, $\mathbb{R}^0 := \{0\}$, $\boldsymbol{x}^{\emptyset} = 0$, and $[N]^0 = \{0\}$.

### 3.2.1 Base Problem: Supervised Learning

Throughout the chapter, we fix $d \in \mathbb{N}$, and consider $\boldsymbol{\mathcal{Z}} = \prod_{j=1}^{d} \boldsymbol{\mathcal{Z}}^j$, where each $\boldsymbol{\mathcal{Z}}^j$ is a measurable subset of $\overline{\mathcal{Z}}^j$ that is $\mathbb{R}$, $\mathbb{N}$, or a finite set. Let $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ be the probability distribution of a random vector $\boldsymbol{Z} := (Z^1, \dots, Z^d)$ taking values in $\boldsymbol{\mathcal{Z}}$, and assume that it has the density $p_{\boldsymbol{\mathcal{Z}}}$. One of the variables, e.g., $Z^{j^*}$ ($j^* \in [d]$), is the target variable which we want to predict. Let $\mathcal{X} = \prod_{j \in [d] \setminus j^*} \overline{\mathcal{Z}}^j$ and $\mathcal{Y} = \overline{\mathcal{Z}}^{j^*}$. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $\ell : \mathcal{H} \times \left( \prod_{j=1}^{d} \overline{\mathcal{Z}}^j \right) \to \mathbb{R}$ be a loss function.

We consider the supervised learning setting; that is, given the training data $\mathcal{D} = \{\boldsymbol{Z}_i\}_{i=1}^n$ that is an i.i.d. sample from $p_{\boldsymbol{\mathcal{Z}}}$, our goal is to find a predictor $\hat{h} \in \mathcal{H}$ with a small risk $\mathcal{R}(\hat{h}) = \mathbb{E}[\ell(\hat{h}, \boldsymbol{Z})]$.

### 3.2.2 Main Assumption

Instead of assuming the existence of an SCM behind the data distribution as we did in the tentative version of the problem description (Problem 2.1), we tackle its slight generalization, where we simply assume that the density satisfies the topological ADMG factorization property (Definition 2.15).

With this slight generalization, the problem setup technically no longer requires the causal interpretation of the ADMGs. However, the causal modeling perspective can be useful in obtaining the graphs from domain experts, i.e., one may be able to draw the graphs by considering the (non-parametric) structural equations [204].

### 3.2.3 Problem Statement

Combining the above, following is the formal statement of our problem setup in this chapter.

**Problem 3.1.** *Let $\mathcal{G} = \langle [d], \mathfrak{D}, \mathfrak{B} \rangle$ be an ADMG and $\prec$ be a topological ordering over $[d]$ with respect to $\mathcal{G}$. Assume that the probability distribution of the data $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ has a density function $p_{\boldsymbol{\mathcal{Z}}}$ and that $p_{\boldsymbol{\mathcal{Z}}}$ satisfies the topological ADMG factorization property with respect to $(\mathcal{G}, \prec)$, i.e.,*

$$p_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{Z}) = \prod_{j=1}^{d} p_{j|\mathrm{mp}(j;\prec)}(\boldsymbol{Z}^{(j)} | \boldsymbol{Z}^{\mathrm{mp}_{\mathcal{G}}(j;\prec)}), \tag{3.1}$$

*where $p_{j|\mathrm{mp}(j;\prec)}$ denotes the conditional density. Now, given $\mathcal{D} = \{\boldsymbol{Z}_i\}_{i=1}^n \overset{i.i.d.}{\sim} p_{\boldsymbol{\mathcal{Z}}}$ and an ADMG $\hat{\mathcal{G}}$ that is an estimator of $\mathcal{G}$,[1] find a predictor $\hat{h} \in \mathcal{H}$ for which the risk $\mathcal{R}(\hat{h}) := \mathbb{E}[\ell(\hat{h}, \boldsymbol{Z})]$ is small, where $\mathbb{E}$ denotes the expectation with respect to $\boldsymbol{Z} \sim \mathbb{P}_{\boldsymbol{\mathcal{Z}}}$.*

[1] Without loss of generality, throughout the chapter, we assume that $[d]$ induces a topological ordering of $\hat{\mathcal{G}}$, i.e., if $1 \leq i < j \leq d$, there is no directed path from $j$ to $i$ in $\hat{\mathcal{G}}$.

## 3.3 Proposed Method

In this section, we explain the proposed data augmentation method to directly incorporate the prior knowledge of an ADMG into supervised learning.

### 3.3.1 Overview of the Method

The method generalizes the intuitive data augmentation method described in the trivariate DAG example in Section 3.1, making it applicable to general ADMGs. The idea is to consider a *nested conditional resampling*; instead of trying to generate all elements of the new data vector at once, we successively resample each variable from the *conditional empirical distribution* [260, 112] conditioning on its Markov pillow. Then, our proposed method, *causal-graph data augmentation*, is obtained by considering all possible resampling paths simultaneously. We later confirm that the proposed method indeed generalizes the previous procedure considered in the trivariate case of Figure 3.1.

### 3.3.2 Derivation of the Proposed Method

Recall, given Equation (3.1), we can express the risk functional as

$$\mathcal{R}(h) = \int_{\mathcal{Z}} \ell(h, \boldsymbol{Z}) \prod_{j=1}^{d} \underbrace{p_{j|\mathrm{mp}(j;\prec)}(Z^j | \boldsymbol{Z}^{\mathrm{mp}_{\mathcal{G}}(j;\prec)})}_{(*)} \mathrm{d}\boldsymbol{Z}.$$

Then, to formulate the nested conditional resampling procedure, we select a kernel function $K^j : \overline{\mathcal{Z}}^{\mathrm{mp}(j)} \to \mathbb{R}_{\geq 0}$ for each $j \in [d]$.[2] Using this kernel function in the spirit of kernel-type function estimators [191, 290, 69], we approximate each conditional density $(*)$ by using the training data $\mathcal{D}$ as

$$\hat{p}_{j|\mathrm{mp}(j)}(Z^j | \boldsymbol{Z}^{\mathrm{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\boldsymbol{Z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\boldsymbol{Z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_k^{\mathrm{mp}(j)})},$$

where $\delta_{\boldsymbol{z}}$ denotes Dirac's delta function centered at $\boldsymbol{z}$ (e.g., [317, Section E.4.1]), and the right-hand side is defined to be zero when the denominator is zero. The resulting approximation to the risk functional $\mathcal{R}(h)$ is

$$\hat{\mathcal{R}}_{\mathrm{aug}}(h) := \int_{\mathcal{Z}} \ell(h, \boldsymbol{Z}) \prod_{j=1}^{d} \hat{p}_{j|\mathrm{mp}(j)}(Z^j | \boldsymbol{Z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{Z}. \tag{3.2}$$

Here, the right-hand side can be interpreted as representing a nested conditional resampling procedure, in which we sequentially select $i_1, \ldots, i_d \in [n]$. Indeed, since each $\hat{p}_{j|\mathrm{mp}(j)}$ places its mass on $\{Z_i^j\}_{i=1}^n$, the integration for $Z^j$ amounts to substituting $Z^j = Z_{i_j}^j$ and summing over the choices $i_j \in [n]$ with appropriate weights. The weight placed on $Z_i^j$ by $\hat{p}_{j|\mathrm{mp}(j)}$, namely $\frac{K^j(\boldsymbol{Z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\boldsymbol{Z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_k^{\mathrm{mp}(j)})}$, depends on $\boldsymbol{Z}^{\mathrm{mp}(j)}$, and it can be computed from $(Z_{i_1}^1, \ldots, Z_{i_{j-1}}^{j-1})$ that have already been selected at the time we select $Z_{i_j}^j$ since $\mathrm{mp}(j) \subset [j-1]$.

[2] Since $\mathcal{G}$ is unknown in practice, we use $\mathrm{mp}(j) := \mathrm{mp}_{\hat{\mathcal{G}}}(j; [d])$ designated by $\hat{\mathcal{G}}$ instead of $\mathrm{mp}_{\mathcal{G}}(j; \prec)$. Also, for notational simplicity, we define $K^j := 1$ where $j$ is such that $\mathrm{mp}(j) = \emptyset$.

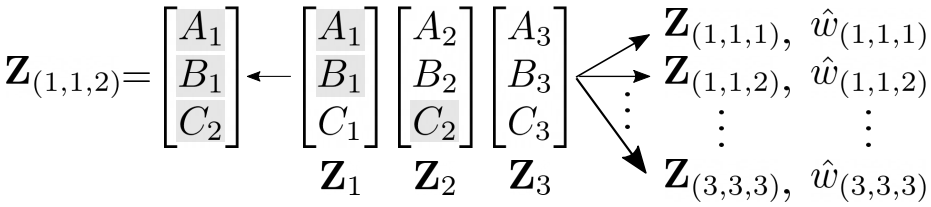



Figure 3.2: Illustration of the proposed data augmentation procedure. Equation (3.3) can be computed from the generated $\boldsymbol{Z_i}$ and the weights $\hat{w}_{\boldsymbol{i}}$.

### 3.3.3 Proposed Method: Causal-graph Data Augmentation

By recursively resolving the integrals in Equation (3.2), we reach at the *instance-weighted risk estimator*:

$$\hat{\mathcal{R}}_{\text{aug}}(h) = \sum_{\boldsymbol{i}\in[n]^d} \hat{w}_{\boldsymbol{i}} \cdot \ell(h, \boldsymbol{Z_i}), \tag{3.3}$$

where

$$\hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{d} \frac{K^j(\boldsymbol{Z}^{\text{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \boldsymbol{Z}^{\text{mp}(j)}_{i_j})}{\sum_{k=1}^{n} K^j(\boldsymbol{Z}^{\text{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \boldsymbol{Z}^{\text{mp}(j)}_{k})}, \tag{3.4}$$
$$\boldsymbol{Z_i} = (Z^1_{i_1}, \ldots, Z^d_{i_d}), \quad \boldsymbol{Z}_{\boldsymbol{i}_{1:j-1}} = (Z^1_{i_1}, \ldots, Z^{j-1}_{i_{j-1}}),$$

for $\boldsymbol{i} = (i_1, \ldots, i_d) \in [n]^d$ and $\boldsymbol{i}_{1:j-1} = (i_1, \ldots, i_{j-1})$, and the right-hand side of Equation (3.4) is defined to be zero when the denominator is zero.[3]

Equation (3.3) can be computed by the following data-augmentation procedure (Figure 3.2). First, $D$ training data points are selected with replacement (specified by $\boldsymbol{i} = (i_1, \ldots, i_d) \in [n]^D$). Then, $\boldsymbol{Z_i}$ is constructed by copying the $j$-th element $Z^j_{i_j}$ from $\boldsymbol{Z}_{i_j}$ ($j \in [d]$). Equation (3.3) can be computed by performing this procedure for all combinations of the indices $\boldsymbol{i} \in [n]^D$. In the proposed data augmentation method, which we call *causal-graph data augmentation*, we consider $\mathcal{D}_{\text{aug}} := \{\boldsymbol{Z_i}\}_{\boldsymbol{i}\in[n]^d}$ to be a weighted training data whose weights are $\mathcal{W}_{\text{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n]^d}$, and we perform supervised learning using $\mathcal{D}_{\text{aug}}$ and $\mathcal{W}_{\text{aug}}$, where any standard method that can incorporate instance weights may be employed. As a practical device, to account for the possibility that $\hat{\mathcal{G}}$ is only an inaccurate approximation of $\mathcal{G}$, we propose to use a convex combination of the *empirical risk estimator* $\hat{\mathcal{R}}_{\text{emp}}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h, \boldsymbol{Z}_i)$ and the *augmented empirical risk estimator* $\hat{\mathcal{R}}_{\text{aug}}(h)$, that is to use

$$\hat{h} \in \operatorname*{arg\,min}_{h\in\mathcal{H}} \{(1-\lambda)\hat{\mathcal{R}}_{\text{emp}}(h) + \lambda\hat{\mathcal{R}}_{\text{aug}}(h) + \Omega(h)\}$$

as the predictor, where $\lambda \in [0, 1]$ is a hyper-parameter and $\Omega$ is a regularization term for $h \in \mathcal{H}$. In the experiments in Section 3.5, we used a fixed parameter $\lambda = .5$ and observed that it performs reasonably well for all data sets.

**Relation to the trivariate case (Section 3.1).** The causal-graph data augmentation generalizes the idea described in the trivariate example $X_1 \leftarrow Y \rightarrow X_2$ in Section 3.1. In fact, in the trivariate example of Figure 3.1, $\mathcal{W}_{\text{aug}}$ places equal weights on the augmented data, essentially yielding the same augmented data set as that in Figure 3.1.

[3] Here, we use the convention $\boldsymbol{Z}^{\text{mp}(1)}_{\boldsymbol{i}_{1:0}} := 0$ to be consistent with the notation.

### 3.3.4 Implementation Details

To reduce the computation cost of calculating the weights $\mathcal{W}_{\mathrm{aug}}$, we exploit the recursive structure in Equation (3.4) that can be represented by a probability tree [34], where we sequentially select the values $i_1, \ldots, i_d \in [n]$ (Figure 3.3). To see this, recursively define

$$\hat{w}_{\boldsymbol{i}_{1:0}} = 1, \quad \hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}} \ (j \in [d], \boldsymbol{i}_{1:j-1} \in [n]^{j-1}),$$

where

$$\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} := \frac{K^j(\boldsymbol{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \boldsymbol{Z}^{\mathrm{mp}(j)}_{i_j})}{\sum_{k=1}^n K^j(\boldsymbol{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \boldsymbol{Z}^{\mathrm{mp}(j)}_{k})},$$

and the right-hand side is defined to be zero when the denominator is zero. Then, we have $\hat{w}_{\boldsymbol{i}} = \hat{w}_{\boldsymbol{i}_{1:d}}$.

With this recursive structure in mind, we construct the probability tree as follows: we index the root node by 0 and the nodes at depth $j \in [d]$ by $\boldsymbol{i}_{1:j}$ in a standard manner, assign the weight $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}$ to each edge $(\boldsymbol{i}_{1:j-1}, \boldsymbol{i}_{1:j})$, and assign to each node $\boldsymbol{i}_{1:j}$ the product of the weights of the edges on the path from the root to $\boldsymbol{i}_{1:j}$. Then, by recursively computing the weights of the nodes on this weighted tree, we can obtain $\mathcal{W}_{\mathrm{aug}}$ (Figure 3.3). Algorithm 2 summarizes the overall procedure of the proposed method.

To reduce the computation cost, we specify a threshold $\theta \in (0, 1)$, and we prune the branches once the node weight becomes lower than $\theta$ along the course of the recursive computation. Since the edge weights satisfy $\sum_{i_j=1}^n \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \in \{0, 1\}$ and $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \geq 0$ for each $\boldsymbol{i}_{1:j-1}$, the node weight $\hat{w}_{\boldsymbol{i}_{1:j}}$ is monotonically decreasing in $j$. Therefore, the above pruning procedure only discards the nodes for which $\hat{w}_{\boldsymbol{i}} < \theta$. The worst-case computational complexity of Algorithm 2 is $\mathcal{O}\left(n^d\right)$ (see Appendix B.4), and it is important in future work to explore how to effectively reduce the computation complexity. Apart from the pruning procedure, one may well consider employing heuristic top candidate search methods such as *beam search* [26] or stochastic optimization methods such as *stochastic gradient descent* [92, Section 5.9] to reduce the computation time by taking advantage of the probability-tree structure.

If all variables are categorical, i.e., $M_j := \left|\overline{\mathcal{Z}}^j\right| < \infty$, and if we employ $K^j(\boldsymbol{x} - \boldsymbol{y}) := \mathbb{1}[\boldsymbol{x} = \boldsymbol{y}]$, the worst-case computational complexity can be reduced by a careful implementation since $\hat{\mathcal{R}}_{\mathrm{aug}}(h)$ essentially becomes a sum of $\prod_{j=1}^d M_j$ terms. Indeed, in this case, the conditional empirical density is

$$\hat{p}_{j|\mathrm{mp}(j)}(Z^j|\boldsymbol{Z}^{\mathrm{mp}(j)}) = \sum_{r=1}^{M_j} \mathbb{1}\left[Z^j = r\right] \frac{\hat{m}_j(r, \boldsymbol{Z}^{\mathrm{mp}(j)})}{\sum_{r'=1}^{M_j} \hat{m}_j(r', \boldsymbol{Z}^{\mathrm{mp}(j)})},$$

where the right-hand side is defined to be zero if the denominator is zero, and $\hat{m}_j(r, \boldsymbol{Z}^{\mathrm{mp}(j)}) := \left|\{i : Z^j_i = r, \boldsymbol{Z}^{\mathrm{mp}(j)}_i = \boldsymbol{Z}^{\mathrm{mp}(j)}\}\right|$ (see Appendix B.2 for a derivation). Thus, after calculating $\hat{m}_j(r, \boldsymbol{Z}^{\mathrm{mp}(j)})$ $(j \in [d], r \in [M_j], \boldsymbol{Z}^{\mathrm{mp}(j)} \in \prod_{k \in \mathrm{mp}(j)})$ in $\mathcal{O}(n)$ computation, we can obtain the augmented data and the weights in $\mathcal{O}\left(\prod_{j=1}^d M_j\right)$ computation.

## 3.4 Theoretical Analysis

In this section, we provide a theoretical justification of the proposed method in the form of an excess risk bound, under the assumption that the CG is perfectly estimated. The goal here is to elucidate how the proposed data augmentation procedure facilitates statistical learning from a theoretical perspective. We focus on the case that $\overline{\mathcal{Z}}^j = \mathbb{R}$ for all $j \in [d]$. Select some $\tilde{K}^j$ and $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^d) \in \mathbb{R}^d_{>0}$, and define $K^j(u) := \frac{1}{|\det \mathbf{H}_j|} \tilde{K}^j(\mathbf{H}_j^{-1} u)$, where $\mathbf{H}_j := \mathrm{diag}(\boldsymbol{h}^{\mathrm{mp}(j)})$ is a

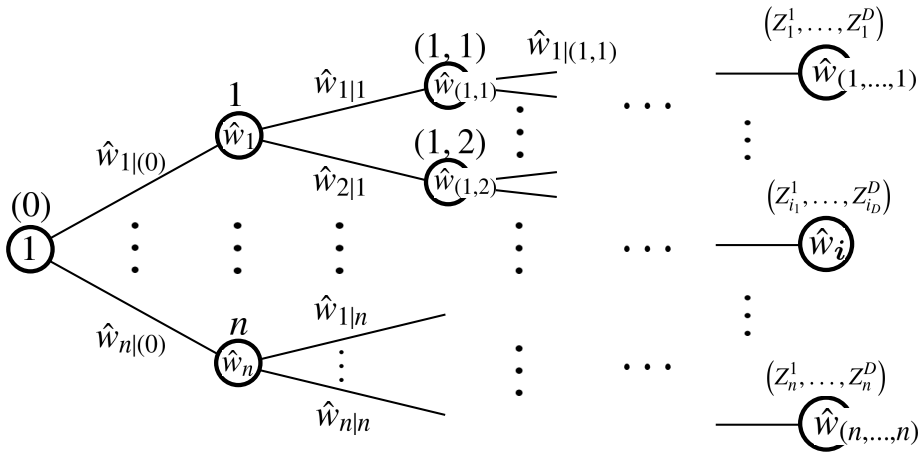



Figure 3.3: Probability tree to compute the weights of the augmented instances. At each depth $j$, the index $i_j$ is selected and the weight is updated as $\hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$.

**Algorithm 2** Proposed method: Causal-Graph Data Augmentation

**Input:** Training data $\mathcal{D}$, ADMG $\hat{\mathcal{G}}$, coefficient $\lambda \in [0,1]$, regularization functional $\Omega$, pruning threshold $\theta \in [0,1)$, hypothesis class $\mathcal{H}$, kernel functions $\{K^j\}_{j=1}^d$, loss function $\ell$.

**function** FillProbTree($\mathcal{D}, \hat{\mathcal{G}}, \theta, \{K^j\}_{j=1}^d$) ▷ see Figure 3.3
  **for** $j \in [d]$ ▷ $j$ is the variable index
    **for** $\boldsymbol{i}_{1:j-1} \in [n]^{j-1}$ ▷ current node (depth $j$)
      **for** $i_j \in [n]$ ▷ next node (depth $j+1$)
        $\hat{w}_{\boldsymbol{i}_{1:j-1}} \leftarrow \hat{w}_{\boldsymbol{i}_{1:j-1}} \mathbb{1}\left[\hat{w}_{\boldsymbol{i}_{1:j-1}} \geq \theta\right]$ ▷ pruning
        $\hat{w}_{\boldsymbol{i}_{1:j}} \leftarrow \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}}$
  **return** $\mathcal{W}_{\text{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^d}$
Let $\mathcal{W}_{\text{aug}} =$ FillProbTree($\mathcal{D}, \hat{\mathcal{G}}, \theta, \{K^j\}_{j=1}^d$).
Let $\hat{\mathcal{R}}_{\text{aug}}(h) := \sum_{\boldsymbol{i} \in [n]^d} \hat{w}_{\boldsymbol{i}} \cdot \ell(h, \boldsymbol{Z}_{\boldsymbol{i}})$.
Let $\tilde{\mathcal{R}}_\lambda(h) := (1-\lambda)\hat{\mathcal{R}}_{\text{emp}}(h) + \lambda\hat{\mathcal{R}}_{\text{aug}}(h) + \Omega(h)$.

**Output:** Trained predictor $\hat{h} \in \arg\min_{h \in \mathcal{H}} \tilde{\mathcal{R}}_\lambda(h)$.

diagonal matrix with elements $\boldsymbol{h}^{\text{mp}(j)}$.

For function classes, we quantify their complexities using the Rademacher complexity.

**Definition 3.1** (Rademacher complexity)**.** *Let $q$ denote a probability distribution on some measurable space $\mathcal{X}$. For a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, define*

$$\text{Rad}_{m,q}(\mathcal{F}) := \mathbb{E}_q \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i f(X_i) \right| \right],$$

*where $\{\sigma_i\}_{i=1}^m$ are independent uniform $\{\pm 1\}$-valued random variables, and $\{X_i\}_{i=1}^m \overset{i.i.d.}{\sim} q$.*

To state our result, let us define the set of marginalized functions and that of the shifted kernel

functions as

$$\mathcal{L}_{\mathcal{H}}^{j} := \left\{ \ell_{h,j}(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{j-1}, \cdot) : h \in \mathcal{H}, (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{j-1}) \in \boldsymbol{\mathcal{Z}}^{[1:j-1]} \right\},$$

$$\left( \ell_{h,j} : \begin{pmatrix} \boldsymbol{z}^1 \\ \vdots \\ \boldsymbol{z}^j \end{pmatrix} \mapsto \int_{\boldsymbol{\mathcal{Z}}^{[j+1:d]}} \ell(h, \boldsymbol{z}) \left( \prod_{k=j+1}^{d} p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k | \boldsymbol{z}^{\mathrm{mp}(k)}) \right) \mathrm{d}\boldsymbol{z}^{[j+1:d]} \right),$$

$$\mathcal{K}_{\mathbf{H}}^{j} := \left\{ K^j(\boldsymbol{z}^{\mathrm{mp}(j)} - (\cdot)) : \boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)} \right\}.$$

The following theorem provides a theoretical justification of the proposed method.

**Theorem 3.1** (Excess risk bound). *Let* $\hat{h} \in \arg\min_{h \in \mathcal{H}} \{\hat{\mathcal{R}}_{\mathrm{aug}}(h)\}$ *as well as* $h^* \in \arg\min_{h \in \mathcal{H}} \{\mathcal{R}(h)\}$, *assuming both exist. Assume* $\hat{\mathcal{G}} = \mathcal{G}$ *and also assume that* $\boldsymbol{\mathcal{Z}}^j \subset \mathbb{R}$ *is compact. Let* $p_{\mathrm{mp}(j)}$ *and* $p_{j,\mathrm{mp}(j)}$ *denote the marginal density of* $\boldsymbol{Z}^{\mathrm{mp}(j)}$ *and the joint density of* $(\boldsymbol{Z}^{(j)}, \boldsymbol{Z}^{\mathrm{mp}(j)})$, *respectively, and assume* $p_{\mathrm{mp}(j)}$ *and* $p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot)$ $(\boldsymbol{z}^j \in \boldsymbol{\mathcal{Z}}^j)$ *have extensions to the entire* $\mathbb{R}^{|\mathrm{mp}(j)|}$ *belonging to* $\Sigma(\beta, L)$, *where* $\Sigma(\beta, L)$ *denotes the Hölder class of functions,* $\beta > 1$, *and* $L > 0$. *Define*

$$R_h := \sum_{j=1}^{d} \left( \max_{j' \in \mathrm{mp}(j)} \boldsymbol{h}^{j'} \right)^{\beta}, \quad R_K := \sum_{j=1}^{d} |\det \mathbf{H}_j| \, \mathrm{Rad}_{n,p}(\mathcal{K}_{\mathbf{H}}^{j}),$$

$$R_{\mathcal{H},K} := \sum_{j=1}^{d} |\det \mathbf{H}_j| \, \mathrm{Rad}_{n,p}\left( \mathcal{L}_{\mathcal{H}}^{j} \otimes \mathcal{K}_{\mathbf{H}}^{j} \right).$$

*Under additional assumptions on the boundedness and smoothness of the kernels and the underlying densities (see Theorem B.1 in Appendix B.3.2), there exist* $C_1, C_p, C_2, C_3, C_4 > 0$ *depending on the boundedness and the smoothness of* $p, \ell, \{\tilde{K}^j\}_{j=1}^{d}$, *and* $h$, *such that for any* $\delta \in (0, 1)$, *we have with probability at least* $1 - \delta$,

$$\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \leq \underbrace{C_1 R_h + C_p}_{\textit{Kernel Bias}} + \underbrace{C_2 R_K}_{\textit{Kernel Complexity}} + \underbrace{C_3 R_{\mathcal{H},K}}_{\textit{Hypothesis Complexity}} + \underbrace{C_4 \sqrt{\frac{\log(4d/\delta)}{2n}}}_{\textit{Uncertainty}}.$$

A proof is provided in Appendix B.3.2. Note that the existence of a smooth extension is satisfied by, e.g., a truncated version of a smooth density on $\mathbb{R}^{|\mathrm{mp}(j)|}$.

**Implications.** Theorem 3.1 implies that the proposed method contributes to statistical learning by reducing the apparent complexity of the hypothesis class at the cost of introducing the additional complexity and bias arising from the kernel approximations. In the interest of space, we provide a formal assessment of this complexity reduction effect in Proposition B.2 in Appendix B.3.3 under some additional Lipschitz-continuity assumptions. In the derivation of Proposition B.2 indicating the complexity reduction effect, the fact that $\mathcal{L}_{\mathcal{H}}^{j}$ consists of univariate functions is critical. In Section 3.5, we empirically confirm that the complexity reduction effect is worth the newly introduced bias and complexity due to the kernel approximation in practice.

**Scope of the analysis.** The present theoretical guarantee only covers the case of $\hat{\mathcal{G}} = \mathcal{G}$, i.e., the case where the topological ADMG factorization property holds with respect to $\hat{\mathcal{G}}$. The robustness of the proposed method to the assumption is an important area of research in future work.

## 3.5 Experimental Evaluation

In this section, we report the results of the real-world data experiments to demonstrate the effectiveness of the proposed method in improving the prediction accuracy. We also show the results of the synthetic-data experiments to investigate the robustness of the proposed method to the estimation error of the CG.

### 3.5.1 Real-world Data Experiment

Here, we describe the setup of the experiment using the real-world data and report the results. The goal of this experiment is to confirm that the proposed method contributes to the performance of the trained predictor, especially in the small-data regime. To investigate the performance improvement, we make a comparison between the two cases: training with and without the proposed device, using the same hypothesis class and the same training algorithm. To analyze the performance improvement in relation to the sample size, we vary the fraction of the data used for training the predictor and compare the performances of the proposed method and that of the baseline without a device. For further details omitted here for the space limitation, please refer to Appendix B.1.

**Data sets.** We employ 6 data sets for the experiment, namely *Sachs* [227], *GSS* [240], *Boston Housing* [98], *Auto MPG* [212], *White Wine* [52], and *Red Wine* [52]. Table 3.1 summarizes these data sets. The *Sachs* data and the *GSS* data are accompanied by the ADMGs obtained from domain experts (Figure 3.4(b) and Figure 3.4(a), respectively), and hence we use them in the experiment. For the other data sets, we first perform *DirectLiNGAM* [240] on the entire data set to obtain the estimated CGs, simulating a situation that we have background knowledge from domain experts. Since DirectLiNGAM produces DAGs, the CGs used in this experiment are DAGs except for the case of *GSS* data set which is accompanied by an ADMG produced by domain experts (Figure 3.4(b)).

**Predictor model class.** We employ the gradient boosted regression trees (GBRTs; [81, 41]) as the predictor model class. The hypothesis class consists of the convex combinations of binary regression trees with at most $M$ leaves:

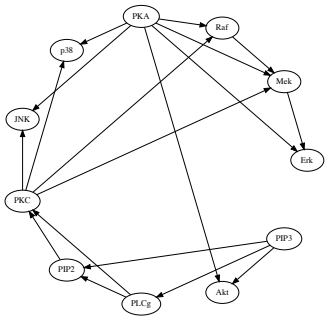
$$\mathcal{H}_{M,K} := \left\{ \sum_{k=1}^{K} \alpha^k w_k^{h_k(\cdot)} : \alpha \in \Delta_K, T_k \in [M], w_k \in \mathbb{R}^{T_k}, h_k \in \mathcal{T}_{T_k} \right\},$$

where $M, K \in \mathbb{N}$, $\mathcal{T}_T$ represents the set of binary tree structures mapping $\mathcal{X}$ to $[T]$, and $\Delta_K$ is the $(K-1)$-dimensional probability simplex. The loss function is the squared error $\ell(h, \boldsymbol{Z}) = (Y - h(\mathbf{X}))^2$ where $Y = Z^{j^*}$ and $\mathbf{X} = \boldsymbol{Z}^{[d]\setminus j^*}$, and the regularization function is $\Omega(h) = \sum_{k=1}^{K} \frac{\rho}{2} \|w_k\|^2$ $(\rho > 0)$. We fix $M = 64$ and search the number of boosting rounds $K$ in $\{10, 50, 250, 1250\}$ and the $\ell_2$-regularization coefficient $\rho$ in $\{1, 10, 100, 1000\}$. The hyper-parameters are selected by the grid-search based on 3-fold weighted cross-validation. Note that, for the proposed method, we perform cross-validation on the union of the original training data and the augmented data with the weights adjusted by $\lambda$, namely $\mathcal{D} \coprod \mathcal{D}_{\text{aug}}$ with weights $(1-\lambda)\mathcal{W}_{\text{orig}} \coprod \lambda \mathcal{W}_{\text{aug}}$ where $\mathcal{W}_{\text{orig}} = (\frac{1}{n}, \ldots, \frac{1}{n})$.

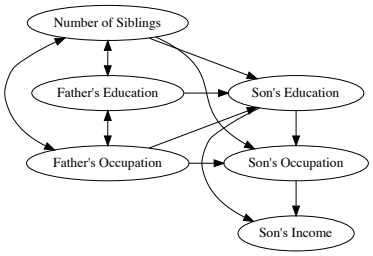
**Configurations of the proposed method.** We select $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^d) \in \mathbb{R}^d_{>0}$ and use the product kernel $K^j(\boldsymbol{x} - \boldsymbol{y}) := \prod_{j' \in \text{mp}(j)} \frac{1}{\boldsymbol{h}^{j'}} K^j_{j'} \left( \frac{\boldsymbol{x}^{j'} - \boldsymbol{y}^{j'}}{\boldsymbol{h}^{j'}} \right)$ for the proposed method. For each $j' \in \text{mp}(j)$, if the variable is continuous (i.e., $\overline{\mathcal{Z}}^{j'} = \mathbb{R}$), we use the Gaussian kernel $K^j_{j'}(x - y) := (2\pi)^{-1/2} \exp\left(-\frac{(x-y)^2}{2}\right)$. Otherwise, i.e., if the variable is discrete, we use the identity kernel $K^j_{j'}(x - y) := \mathbb{1}[x = y]$ and $\boldsymbol{h}^{j'} = 1$. For the Gaussian kernels, we select the *kernel bandwidth* $\boldsymbol{h}^{j'}$ based on

Table 3.1: Summary of Data Sets (*Name*: name of the data set, *#Var*: number of variables in the data set, *#Obs*: number of observations, *Graph*: CG used for the proposed method, *Consensus*: consensus network (Figure 3.4(b)), *Domain*: domain knowledge of the status attainment model (Figure 3.4(a)), *LiNGAM*: CG is estimated by performing DirectLiNGAM on the entire data set).

| Name | #Var | #Obs | Graph |
|---|---|---|---|
| *Sachs* | 11 | 853 | Consensus |
| *GSS* | 6 | 1380 | Domain |
| *Boston Housing* | 14 | 506 | LiNGAM |
| *Auto MPG* | 7 | 392 | LiNGAM |
| *White Wine* | 12 | 4898 | LiNGAM |
| *Red Wine* | 12 | 1599 | LiNGAM |




(a) Reference graph for Sachs data.




(b) Reference graph for GSS data.

Figure 3.4: Reference CGs for the data sets used in our experiments. (a) Consensus graph (Sachs et al. [227]). (b) Domain-knowledge graph based on the status attainment model (Duncan et al. [66]).

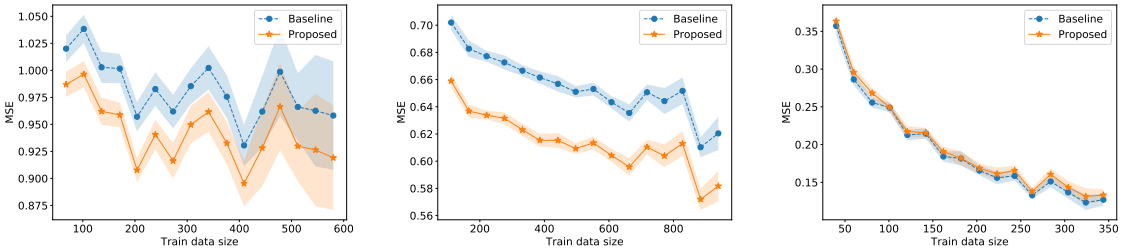
*Silverman's rule-of-thumb* [250, pp.45–47]. In the experiment, we fix $\lambda = .5$ throughout all runs and find that it yields reasonable performances in all data sets.

**Compared methods.** We compare the performances of the proposed method and the naive baseline method without a device:

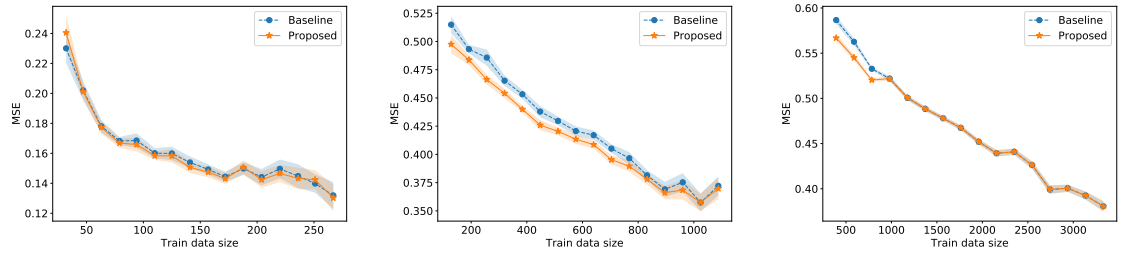
$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \{\hat{\mathcal{R}}_{\text{emp}}(h) + \Omega(h)\}.$$

In Section 3.5.1 where we report the results, the two methods are referred to as *Proposed* and *Baseline*, respectively.

**Evaluation procedure.** The prediction accuracy is measured by the mean squared error (MSE). For each data set, we randomly subsample a fraction of the data as the training set and use the rest as the testing set. The fraction of the training set is varied in $\{.1, .15, \ldots, .85\}$. For each training set fraction, random train-test splits are performed 20 times. Subsequently, for each split, *Proposed* and *Baseline* are trained on the training set, and then evaluated on the testing set. We report the average performances as well as the standard errors over the 20 runs for each training set fraction.




(a) Sachs data. (b) GSS data. (c) Boston Housing data.




(d) Auto MPG data. (e) Red Wine data. (f) White Wine data.

Figure 3.5: Illustration of the experimental results. In all figures, the horizontal axis is the varied size of the training data before augmentation, and the vertical axis is the performance metric (MSE; the lower the better). The markers and the lines indicate the average over the 20 independent runs, and the shades are drawn for the width of the standard errors both above and below the lines. The proposed method shows a consistent improvement over the naive baseline based on the empirical risk minimization with the same hypothesis class, particularly in the small-data regime.

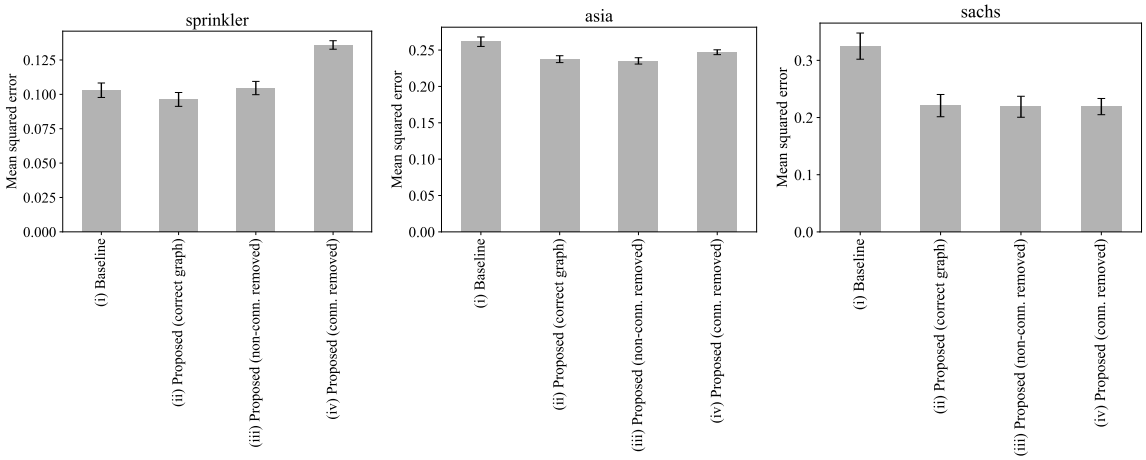
**Results.** Figure 3.5 shows the experimental result. We observe a consistent performance improvement in most of the data sets. For the data sets for which the domain knowledge CG is provided (i.e., *Sachs* and *GSS*), we can see clear relative improvement of 3–7% on average, especially in the small-data regime where the training set fraction is approximately 10–40%. In the other data sets without the background knowledge, relatively little improvement is observed except in the small-data regions of *Red Wine* and *White Wine*, where up to 4% relative improvement on average is observed. The lack of relative improvement in the majority of these cases emphasizes the importance of having accurate domain knowledge in the proposed approach, and it motivates the development of effective causal discovery methods. In the *White Wine* data, the proposed method coincides with the baseline in the larger-data region as the augmentation did not effectively take place due to the adaptive bandwidth that is narrowed according to the sample size. For supplementary figures visualizing the average relative improvements, see Appendix B.1.5.

### 3.5.2 Synthetic-data Experiment

Here, we report the experimental results to evaluate the robustness of the proposed method to the misspecification of the CG. We used synthetic data for which the ground-truth CG is known, and we apply edge alterations. We used three artificial data generation models: *sprinkler*, *asia*, and *sachs* (see Table 3.2). Each model consisted of a ground-truth CG and a set of conditional probability tables (CPDs) specifying the generative model, both of which were fixed throughout the experiments for all models. As the predictor model class, we used the GBRTs [81, 41]. See Appendix B.2 for further details. For each run of the experiment, a total of 100 data points were independently and identically generated. We used 30 data points from this set for training, while we used the remaining 70 for testing. After each data generation, we fit the predictor with or without the proposed method

Table 3.2: Summary of Synthetic Data Sets (*Name*: name of the data set, *#Var*: number of variables in the data set, *#Edge*: number of edges.

| Name | #Var | #Edge |
|---|---|---|
| *sprinkler* | 4 | 4 |
| *asia* | 8 | 8 |
| *sachs* | 11 | 17 |




(a) *sprinkler* data. (b) *asia* data. (c) *sachs* data.

Figure 3.6: Illustration of the results of the synthetic-data experiments: (i) the baseline without using the proposed method, (ii) the proposed method applied with the correct CG, (iii) the proposed method applied with a wrong CG where an edge is removed among those not connected to the target variable, and (iv) the proposed method applied with a wrong CG where an edge is removed among those connected to the target variable. The error bars indicate the *two-sigma* intervals, i.e., $[\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}]$ where $\hat{\mu}$ is the mean and $\hat{\sigma}$ is the standard error.

and measured the prediction accuracy. Concretely, we evaluated the following four cases: (i) the predictor is trained without employing the proposed method, (ii) the proposed method is employed with the correct underlying CG, (iii) the proposed method is applied using an altered CG where one edge is removed among those not attached to the prediction target variable, (iv) the proposed method is applied using an altered CG where one edge is removed among those attached to the prediction target variable.

For cases (iii) and (iv), since multiple candidates for the removal exist, we measured the performance for each possible edge removal and calculated the average performance. We repeated the experiment 100 times for each data generation model and reported the summarized results.

Figure 3.6 shows the results. In the cases (iii) and (iv), i.e., where an edge was removed, performance degradation was generally observed compared to the ideal case that the proposed method is applied with the correct ground-truth CG. In all three synthetic data sets, the degree of degradation was more prominent when the removed edge was directly connected to the predicted variable than when it was not connected to the predicted variable. For instance, in the *sprinkler* data, 9% excess error was observed in case (iii) relatively to case (ii), whereas 41% excess relative error was observed in case (iv). Another observation is that the effect of edge removal is milder in larger graphs: the error of case (iv) relative to case (ii) was approximately 41%, 4%, and 0%, respectively, in the *sprinkler*, *asia*, and *sachs* data sets. This may be attributed to the redundancy

of information, i.e., in larger graphs, even if one edge is wrongly removed, the other edges may typically retain a majority of the dependency structure.

## 3.6 Related Work and Discussion

In this section, we explain the context of the content of the chapter in relation to existing work.

### 3.6.1 GCMs and Predictive Modeling

**Variable selection in a single-distribution setting.** The background knowledge encoded in a CG can be used for variable selection by identifying a *Markov boundary* of the target variable. Here, $\mathrm{mb}(j) \subset [d]$ is called a *Markov blanket* of $j$ if $Z^j$ is conditionally independent of all the other variables given $\boldsymbol{Z}^{\mathrm{mb}(j)}$. If, moreover, $\mathrm{mb}(j)$ is minimal, i.e., if none of its proper subsets are Markov blankets, it is called a *Markov boundary* (MB). Under certain assumptions, the MB of a target variable is known to be the minimal set of variables with optimal predictive performance [274]. For a recent comprehensive review on MB estimation, see Yu et al. [307]. The present work is orthogonal to this line of work. In fact, the CGs can encode more information than a specification of the Markov boundary of the predicted variable; for example, consider the CG $X_1 \leftarrow Y \rightarrow X_2$ where $Y$ is the target variable and $(X_1, X_2)$ are the predictors. In this case, the Markov boundary of $Y$ is $(X_1, X_2)$, and hence the variable selection does not reduce the number of the predictors. On the other hand, the proposed method still leverages the factorization structure of the data distribution entailing the CG. In practice, the two approaches can be combined straightforwardly. In our experiments, we do not perform variable selection using the data regarding the possibility that the obtained CGs are inaccurate.

**Variable selection in distribution-shift setting.** Another line of research is concerned with making predictions under distribution shift and leverage feature selection based on causal background knowledge or causal discovery. Magliacane et al. [176] considered the case that a distribution shift is due to intervention in some variables, and they proposed a method to perform domain adaptation by identifying a set of variables that is likely to perform well regardless of the intervention. Rojas-Carulla et al. [223] assume that if the conditional distribution of the predicted variable given some subset of features is invariant across different distributions, then this conditional distribution is the same in the *target distribution* for which one wants to make good predictions, and leveraged it to find the set of variables for which the relation to the target variable does not change. The present work is complementary to this line of work since our goal is to make good predictions in a single fixed distribution.

**Regularization and model selection.** Kyono and van der Schaar [155] proposed a model selection criterion that can reflect the structure of a CG. The goal of Kyono and van der Schaar [155] is *domain generalization* and *out-of-distribution prediction*, i.e., making good predictions under a distribution shift without access to any samples from the target distribution or making good predictions for the data that is outside the support of the training data distribution. To achieve it, given a DAG as prior knowledge, Kyono and van der Schaar [155] first modify it so that the edges coming out of the target variable are removed. Then, to score the predictor model candidates, it generates a data set whose predicted variables are replaced by the predictions of the model and computes the *Bayes Information Criterion* (BIC) that evaluates the fitness of the modified DAG structure to the generated data set. Another approach for using the background knowledge of a CG is the *CASTLE regularization* [156]. CASTLE regularization regularizes a neural network while performing the CG discovery as an auxiliary task. The method imposes a reconstruction loss using

the internal layers of the predictor implemented by neural networks under a DAG constraint. The present work is orthogonal to these researches and can be straightforwardly combined in practice. Also note that our method has a theoretical justification while Kyono and van der Schaar [155] provided no theoretical justifications.

**Inference under specific CGs.** Under some specific problem settings with known specific underlying CGs, methods to take advantage of the prior knowledge have been developed. For example, in the instance weight estimation for episodic reinforcement learning, methods to perform *state simplification* based on the CGs have been proposed [32, 208, Section 8.2]. Schölkopf et al. [231] considered removing systematic errors using *half-sibling regression* inspired by the CG of the observation mechanism found in the *exoplanet search*. Pitis et al. [211] proposed a method to enhance the sample efficiency in reinforcement learning (RL) by a procedure to exchange the realizations of the variables within the (conditionally) disconnected components in the CG of the *Markov decision process* of specific RL instances. This line of work and the present work are complementary in that our approach is widely applicable to general ADMGs whereas these analyses have the potential to exploit the characteristics of the specific problem setups.

**Causal bootstrapping.** Recently, Little and Badawy [169] proposed *causal bootstrapping*, a weighted bootstrap-type algorithm that is relevant to our method. While, methodologically, both the present work and Little and Badawy [169] can be seen to be based on kernel-type function estimators [260, 112, 69] and CGs [204], the two works are complementary in that the problem setups differ. Causal bootstrapping of Little and Badawy [169] aims at mitigating the performance degradation due to a distribution shift arising from an intervention, and it uses kernel-type function estimators to simulate sampling from an interventional distribution. On the other hand, we investigate the performance improvement yielded from using the background knowledge of a CG in a scenario without a distribution shift.

**Constructing probabilistic graphical models.** Evans and Richardson [73] provided a smooth parametrization of the set of distributions that are *Markov with respect to* an ADMG $\mathcal{G}$ in the binary case: $\overline{\mathcal{Z}}^j = \{0, 1\}$ $(j \in [d])$. Complementarily, for the case of $\overline{\mathcal{Z}}^j = \mathbb{R}$ $(j \in [d])$, Silva et al. [248] proposed the construction of flexible probability models that are Markov with respect to a given ADMG. Similarly, in the case that the ADMG has no bi-directed edges, constructing a Bayesian network by specifying the conditional distributions appearing in the Markov factorization (Equation (3.1)) is one natural way to exploit this prior knowledge [174]. This approach has the limitation that it inevitably restricts the modeling choice, e.g., the constructed predictor is a generative model as opposed to a discriminative model [236, Chapter 24], whereas our approach has the virtue of being model-agnostic.

### 3.6.2 Causal Discovery and Transfer Learning

Our method provides a channel through which an estimated CG can be used for enhancing the predictive modeling. In this sense, the proposed method can serve as a transfer learning method under a *transfer assumption* of *common CG*, i.e., an assumption that one is given many samples from another distribution sharing the same CG with the distribution for which we want to make the predictions. Under such an assumption, one may first estimate the ADMG using causal discovery methods to estimate the *Markov equivalence class* of ADMGs expressed as a *partial ancestral graph* (PAG) [309], e.g., the *fast causal inference* (FCI) algorithm [254, 309], enumerate the ADMGs in the equivalence class (e.g., by the *Pag2admg* algorithm; [261]), select a plausible candidate ADMG that is concordant with the domain knowledge, and apply the proposed method. Such an assumption of

a common causal mechanism has been exploited in recent work of causal discovery [302, 82, 185] and transfer learning [206, 176, 267], and it is based on a common belief that a causal mechanism remains invariant unless explicitly intervened in (Hünermund and Bareinboim [119]).

### 3.6.3 GCMs and Efficient Estimation

Our method could also be seen as a method to perform sample-efficient inference given a CG. In the existing work, the knowledge of a CG has been used for deriving efficient estimators for *identifiable causal estimands* [204] such as the *interventional distributions* [135, 134] or the *average causal effect* [24]. For instance, Jung et al. [135] and Jung et al. [134] derived expressions of efficient estimators of the identifiable interventional distributions given an ADMG and a PAG, respectively, by leveraging the knowledge of the CG in the *double/debiased machine learning* [43] framework. Another line of research provided graphical criteria for selecting the *efficient adjustment sets*, the set of covariates to be adjusted for producing a valid estimator of a causal effect with the minimal asymptotic variance [101, 225, 293, 251]. Our goal differs from the goals of these lines of research; we are interested in improving the sample efficiency of training the predictor whereas they aimed to improve the sample efficiency of causal inference. Nevertheless, it is an interesting direction of future research to elucidate whether the proposed method is optimally efficient in estimating the risk functional given the CG.

### 3.6.4 Cyclic CGs

In this chapter, we focused on the case that the CG has no cycles. In the case of acyclic SCMs (i.e., those whose CGs are ADMGs; Definition 2.4), the topological ADMG factorization (Problem 3.1) is known to hold without additional assumptions. In contrast, in the case of cyclic SCMs (i.e., those that are not acyclic), an analogous property (the *recursive factorization property* in [80, Definition 3.6.1(4)]) holds only under certain sufficient conditions on the SCMs ([80, Section 3.6]). See Forré and Mooij [80, Section 3.6] for details.

### 3.6.5 Users' Burden of Inputting CGs

To apply the proposed method, one has to specify the estimated CG. If the user can collaborate with domain experts to draw a CG based on its semantics, i.e., that the arrows indicate direct influence relations (e.g., [265]), they can apply the proposed method easily. In such a case, the ecosystem to facilitate drawing and communicating CGs may potentially reduce the user's burden, e.g., web-based software such as *DAGitty* [268].

On the other hand, in the application domains where the CG is not readily available, the user may resort to the statistical causal discovery methods to estimate the CG from data [253, 88, 114, 239, 209, 210, 187, 133, 132, 25]. The user may apply causal discovery methods to the data in relevant domains that share the CG and use the graph with the proposed method as explained in Section 3.6.2. For example, ideally, the CG representing a pathological mechanism may potentially be estimated from the data of some demographic group and applied in the data of another group. When partial knowledge of some (direct/indirect) causal relations is available, one can incorporate it into the estimation procedure [31, 261].

In principle, when multiple candidate CGs are available, since the goal is to obtain an accurate predictor, one may well treat the CG estimator in Algorithm 2 as a hyper-parameter and use cross-validation to choose one from the candidates. One can optionally assign a weight coefficient $\lambda \in [0, 1]$ to control the impact of the augmented data on the learning process, as explained in Section 3.3.3.

## 3.7 Conclusion

In this chapter, we proposed a general method for exploiting the causal graphical prior knowledge in predictive modeling. We theoretically provided an excess risk bound indicating that the proposed method has a complexity reduction effect that mitigates overfitting while it introduces additional complexity and bias arising from the kernel approximations. Through the experiments using real-world data, we demonstrated that the proposed method consistently improves the predictive performance especially in the small-data regime, which implies that the complexity reduction effect is worth the newly introduced bias and complexity in practice. Important areas in future work include incorporating the other equality constraints than the topological ADMG factorization that are imposed by an ADMG, and handling more relaxed assumptions such as those expressed as PAGs. In Chapter 6, we discuss further possibilities of future research directions.

# Chapter 4

# When Structural Causal Model is Estimable: Causal Mechanism Transfer

In the previous chapter, we considered the case that a causal graph is estimable or known thanks to domain knowledge, and discussed how to exploit the graphical knowledge in training a predictor in supervised machine learning. In this chapter, we turn to the case that a structural causal model (or more precisely, its reduced-form structural function, RSF) is estimable from the data of relevant domains, and discuss how the estimated (partial) knowledge of the causal system can be exploited in training a predictor.

The problem setup corresponds to a *domain adaptation* problem, a scenario where we have access to auxiliary data from different but relevant domains (the *source* domains) in addition to a sample from the *target* distribution for which we want to eventually find a good predictor. A key question in domain adaptation is "what is the *transfer assumption* to specify the relation between the auxiliary data distributions and the target distribution?" Our assumption, namely the estimability of the RSF from source domain data, gives rise to the novel transfer assumption of a *shared causal mechanism*, i.e., that the distributions are derived from structural causal models whose structural equations are identical.

## 4.1 Overview

Learning from a limited amount of data is a long-standing yet actively studied problem of machine learning. Domain adaptation (DA) [19] tackles this problem by leveraging auxiliary data sampled from related but different domains. In particular, we consider *few-shot supervised* DA for regression problems, where only a few labeled target domain data and many labeled source domain data are available.

### 4.1.1 Motivation

A key component of DA methods is the *transfer assumption* (TA) to relate the source and the target distributions. Many of the previously explored TAs have relied on certain direct distributional similarities, e.g., identical conditionals [241] or small distributional discrepancies [20]. However, these TAs may preclude the possibility of adaptation from apparently very different distributions. Many others assume parametric forms of the distribution shift [313] or the distribution family

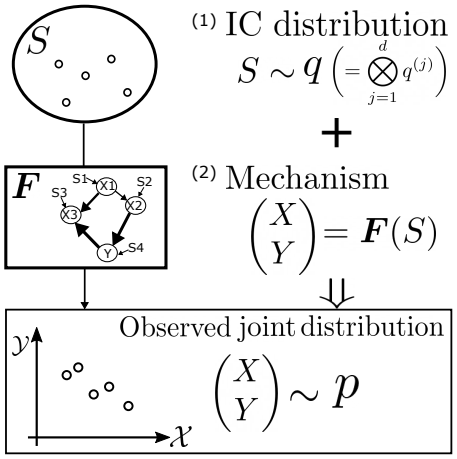



Figure 4.1: Nonparametric generative model of nonlinear independent component analysis. Our meta-distributional transfer assumption is built on the model, where there exists an invertible function $\boldsymbol{F}$ representing the mechanism to generate the labeled data $(X, Y)$ from the independent components (ICs), $S$, sampled from $q$. As a result, each pair $(\boldsymbol{F}, q)$ defines a joint distribution $p$.

[259] which can highly limit the considered set of distributions. (we further review related work in Section 4.6.1). To alleviate the intrinsic limitation of previous TAs due to relying on apparent distribution similarities or parametric assumptions, we focus on a meta-distributional scenario where there exists a common generative *mechanism* behind the data distributions (Figures 4.1 and 4.2). Such a common mechanism may be more conceivable in applications involving structured table data such as medical records [304]. For example, in medical record analysis for disease risk prediction, it can be reasonable to assume that there is a pathological mechanism that is common across regions or generations, but the data distributions may vary due to the difference in cultures or lifestyles. Such a hidden structure (pathological mechanism, in this case), once estimated, may provide portable knowledge to enable DA, allowing one to obtain accurate predictors for under-investigated regions or new generations.

### 4.1.2 Idea

Concretely, our assumption relies on the generative model of nonlinear independent component analysis (nonlinear ICA; Figure 4.1), where the observed labeled data are generated by first sampling latent independent components (ICs) $S$ and later transforming them by a nonlinear invertible *mixing function* denoted by $\boldsymbol{F}$ [124]. Under this generative model, our TA is that $\boldsymbol{F}$ representing the mechanism is identical across domains (Figure 4.2). This TA allows us to formally relate the domain distributions and develop a novel DA method without assuming their apparent similarities or making parametric assumptions.

**Example: Structural equation models** A salient example of generative models expressed as Equation (4.1) is *structural equation models* (SEMs; [204, 208]), which are used to describe the data-generating mechanism involving the causality of random variables [204]. More precisely, the generative model of Equation (4.1) corresponds to the *reduced form* [215] of a *Markovian* SEM [204], i.e., a form where the structural equations to determine $Z$ from $(Z, S)$ are solved so that $Z$ is expressed as a function of $S$. Such a conversion is always possible because a Markovian SEM induces an *acyclic* causal graph [204], and hence the structural equations can be solved by elimination of variables. This interpretation of reduced-form SEMs as Equation (4.1) has been exploited in methods of *causal discovery*, e.g., in the linear non-Gaussian additive-noise models and their successors [137, 239, 185]. In the case of SEMs, the key assumption of this paper translates into the invariance of the causal mechanisms (expressed by the structural equations) across domains, which enables an intuitive assessment of the assumption based on prior knowledge. For instance, if all domains have the same causal mechanism and are in the same intervention state (including an intervention-free

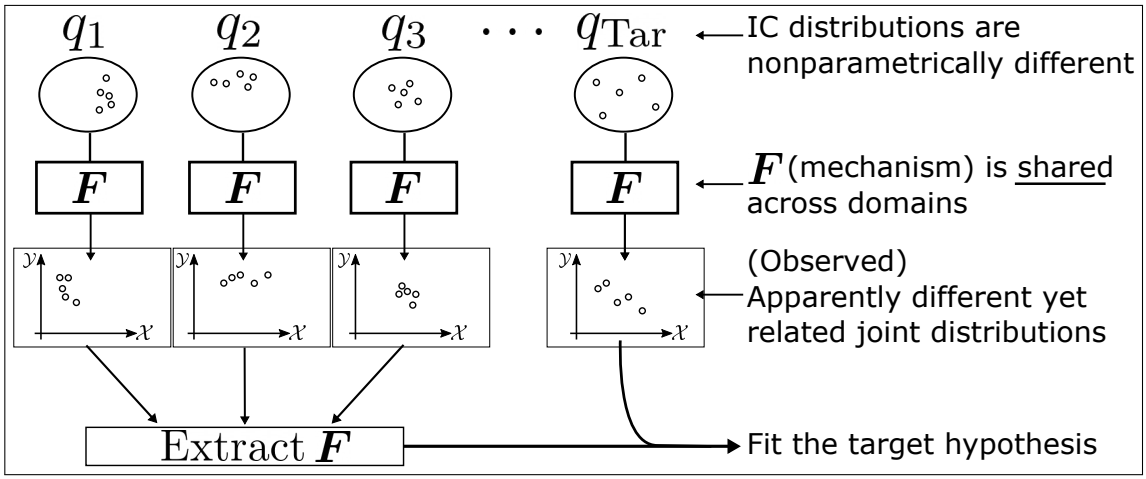



Figure 4.2: Our assumption of common generative mechanism. By capturing the common data generation mechanism, we enable domain adaptation among seemingly very different distributions without relying on parametric assumptions.

case), the modeling choice is deemed plausible. Note that we do not estimate the original structural equations in the proposed method (Section 4.3.2) but we only require estimating the reduced form, which is an easier problem compared to causal discovery.

### 4.1.3 Contributions

Our key contributions can be summarized in three points as follows.

1. We formulate the flexible yet intuitively accessible TA of shared generative mechanism and develop a few-shot regression DA method (Section 4.3.2). The idea is as follows. First, from the source domain data, we estimate the mixing function $\boldsymbol{F}$ by nonlinear ICA [124] because $\boldsymbol{F}$ is the only assumed relation of the domains. Then, to transfer the knowledge, we perform data augmentation using the estimated $\boldsymbol{F}$ on the target domain data using the independence of the IC distributions. In the end, the augmented data is used to fit a target predictor (Figure 4.3).

2. We theoretically justify the augmentation procedure by invoking the theory of generalized U-statistics [162]. The theory shows that the proposed data augmentation procedure yields the uniformly minimum variance unbiased risk estimator in an ideal case. We also provide an excess risk bound [184] to cover a more realistic case (Section 4.4).

3. We experimentally demonstrate the effectiveness of the proposed algorithm (Section 4.5). The real-world data we use is taken from the field of *econometrics*, for which structural equation models have been applied in previous studies [93].

A salient example of the generative model we consider is the structural equations of causal modeling (Section 4.2). In this context, our method can be seen as the first attempt to fully leverage the structural causal models for DA (Section 4.6.2).

## 4.2 Problem Setup and Main Assumption

In this section, we describe the problem setup and the notation. To summarize, our problem setup is *homogeneous*, *multi-source*, and *few-shot supervised* domain adapting regression. That is, respectively, all data distributions are defined on the same data space, there are multiple source

domains, and a limited number of labeled data is available from the target distribution (and we do *not* assume the availability of unlabeled data). In this chapter, we use the terms *domain* and *distribution* interchangeably.

**Basic notation.** Throughout the chapter, we fix $d(\in \mathbb{N}) > 1$ and suppose that the input space $\mathcal{X}$ is a measurable subset of $\mathbb{R}^{d-1}$ and the label space $\mathcal{Y}$ is a measurable subset of $\mathbb{R}$. As a result, the overall data space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is a measurable subset of $\mathbb{R}^d$. We generally denote a labeled data point by $Z = (X, Y)$. We denote by $\mathcal{Q}$ the set of independent distributions on $\mathbb{R}^d$ with absolutely continuous marginals. For a distribution $p$, we denote its induced expectation operator by $\mathbb{E}_p$. Table C.1 in Appendix C provides a summary of notation.

### 4.2.1 Base Problem: Few-shot Domain Adapting Regression

Let $p_{\mathrm{Tar}}$ be a distribution (the *target distribution*) over $\mathcal{Z}$, and let $\mathcal{H} \subset \{h : \mathbb{R}^{d-1} \to \mathbb{R}\}$ be a hypothesis class. Let $\ell : \mathcal{H} \times \mathbb{R}^d \to [0, B_\ell]$ be a loss function where $B_\ell > 0$ is a constant. Our goal is to find a predictor $h \in \mathcal{H}$ which performs well for $p_{\mathrm{Tar}}$, i.e., the target risk $\mathcal{R}(h) := \mathbb{E}_{p_{\mathrm{Tar}}} \ell(h, \boldsymbol{Z})$ is small. We denote $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{R}(h)$. To this goal, we are given an i.i.d. sample $\mathcal{D}_{\mathrm{Tar}} := \{\boldsymbol{Z}_i\}_{i=1}^{n_{\mathrm{Tar}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{Tar}}$. In a fully supervised setting where $n_{\mathrm{Tar}}$ is large, a standard procedure is to select $h$ by empirical risk minimization (ERM), i.e., $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{\mathcal{R}}(h)$, where $\hat{\mathcal{R}}(h) := \frac{1}{n_{\mathrm{Tar}}} \sum_{i=1}^{n_{\mathrm{Tar}}} \ell(h, \boldsymbol{Z}_i)$. However, when $n_{\mathrm{Tar}}$ is not sufficiently large, $\hat{\mathcal{R}}(h)$ may not accurately estimate $\mathcal{R}(h)$, resulting in a high generalization error of $\hat{h}$.

### 4.2.2 Main Assumption

To compensate for the scarcity of data from the target distribution, let us assume that we have data from $K$ distinct *source distributions* $\{p_k\}_{k=1}^K$ over $\mathcal{Z}$, that is, we have independent i.i.d. samples $\mathcal{D}_k := \{\boldsymbol{Z}_{k,i}^{\mathrm{Src}}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} p_k (k \in [K], n_k \in N)$ whose relations to $p_{\mathrm{Tar}}$ are described shortly. We assume $n_{\mathrm{Tar}}, n_k \geq d$ for simplicity.

In this chapter, the key transfer assumption is that all domains follow nonlinear ICA models with identical mixing functions (Figure 4.2). To be precise, we assume that there exists a set of IC distributions $q_{\mathrm{Tar}}, q_k \in \mathcal{Q} (k \in [K])$ , and a smooth invertible function $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^d$ (the *transformation* or *mixing*) such that $\boldsymbol{Z}_{k,i}^{\mathrm{Src}} \sim p_k$ is generated by first sampling $S_{k,i}^{\mathrm{Src}} \sim q_k$ and later transforming it by

$$\boldsymbol{Z}_{k,i}^{\mathrm{Src}} = \boldsymbol{F}(S_{k,i}^{\mathrm{Src}}), \tag{4.1}$$

and similarly $\boldsymbol{Z}_i = \boldsymbol{F}(S_i), S_i \sim q_{\mathrm{Tar}}$ for $p_{\mathrm{Tar}}$. The above assumption allows us to formally relate $p_k$ and $p_{\mathrm{Tar}}$. It also allows us to estimate $\boldsymbol{F}$ when sufficient identification conditions required by the theory of nonlinear ICA are met. Due to space limitation, we provide a brief review of the nonlinear ICA method used in this paper and the known theoretical conditions in Appendix C.1. Having multiple source domains is assumed here for the identifiability of $\boldsymbol{F}$: it comes from the currently known identification condition of nonlinear ICA [124]. Note that complex changes in $q$ are allowed, and hence the assumption of invariant $\boldsymbol{F}$ can accommodate intricate shifts in the apparent distribution $p$. We discuss this further in Section 4.6.3 by taking a simple example.

### 4.2.3 Problem Statement

Instead of assuming the existence of SCMs as we did in the tentative version of the problem description (Problem 2.2), we tackle its slight generalization where the data are generated by ICMs (Definition 2.16).

**Problem 4.1.** *Let $K \in \mathbb{N}$ and let*

$$\begin{aligned}\mathcal{M}_1 &= \langle [d], [d], \mathbb{R}^d, \mathbb{R}^d, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},1} \rangle,\\ &\vdots\\ \mathcal{M}_K &= \langle [d], [d], \mathbb{R}^d, \mathbb{R}^d, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},K} \rangle,\\ \mathcal{M}_{\mathrm{Tar}} &= \langle [d], [d], \mathbb{R}^d, \mathbb{R}^d, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{\mathcal{E}},\mathrm{Tar}} \rangle\end{aligned}$$

*be ICMs. Let the source domain data sets $\mathcal{D}_k := \{\boldsymbol{Z}_{k,i}^{\mathrm{Src}}\}_{i=1}^{n_k} \overset{i.i.g.}{\leftarrow} \mathcal{M}_k (k \in [K])$ and the target domain training data set $\mathcal{D}_{\mathrm{Tar}} := \{\boldsymbol{Z}_i\}_{i=1}^{n_{\mathrm{Tar}}} \overset{i.i.g.}{\leftarrow} \mathcal{M}_{\mathrm{Tar}}$ be given, where $\{n_k\}_{k\in[K]}$ are large and $n_{\mathrm{Tar}}$ is small. Assume that $\mathcal{M}_1, \ldots, \mathcal{M}_K$ satisfy some identifiability condition (such as the conditions of Proposition 2.9) so that there exists an algorithm* ICA*, and* $\mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$ *is a consistent estimator of $\boldsymbol{F}$. Find a predictor $\hat{h} \in \mathcal{H}$ for which the risk $\mathcal{R}(\hat{h}) := \mathbb{E}_{\mathrm{Tar}}[\ell(\hat{h}, \boldsymbol{Z})]$ is small, where $\mathbb{E}_{\mathrm{Tar}}$ denotes the expectation with respect to $\boldsymbol{Z} \overset{gen}{\leftarrow} \mathcal{M}_{\mathrm{Tar}}$.*

## 4.3 Proposed Method

In this section, we detail the proposed method, causal mechanism transfer (Algorithm 3).

### 4.3.1 Overview of the Method

The method proceeds in three steps: *estimation*, *inflation*, and *synthesis*, which are visually summarized in Figure 4.3. Each step is elaborated upon in Section 4.3.2.

The *estimation* step estimates the common transformation $\boldsymbol{F}$ from the source domain data by applying nonlinear ICA, namely via *generalized contrastive learning* (GCL; [124]), since it is the sole connection that we posed as the transfer assumption. Then, the estimated $\boldsymbol{F}^{-1}$ and $\boldsymbol{F}$ are used in the *inflation* step and the *synthesis* step, respectively, to perform data augmentation. The *inflation* step applies the estimated $\boldsymbol{F}^{-1}$ to estimate the ICs from the target domain data, and it generates many fictional *candidate* ICs using the knowledge that sets of independent IC vectors are equally likely even if we scramble the combinations. Finally, the *synthesis* step applies the estimated $\boldsymbol{F}$ to all of the generated candidate ICs to obtain the augmented data for the target domain. The generated *pseudo training data* is used for training the predictor.

### 4.3.2 Proposed Method: Causal Mechanism Transfer

**Step 1: Estimate $\boldsymbol{F}$ using the source domain data.** First, we estimate the common generative mechanism $\boldsymbol{F}$, which is the sole connection between the source domains and the target domain. The estimation can be realized by performing nonlinear ICA using the source domain data, namely via *generalized contrastive learning* (GCL; [124]). GCL uses auxiliary information for training a certain binary classification function, $r_{\hat{\boldsymbol{F}},\boldsymbol{\psi}}$, equipped with a parametrized feature extractor $\hat{\boldsymbol{F}} : \mathbb{R}^d \to \mathbb{R}^d$ and a set of functions $\psi = \{\psi_j\}_{j=1}^d$, where each $\psi_j$ is a function from $\mathbb{R} \times \mathcal{U}$ to $\mathbb{R}$, and $\mathcal{U}$ is some measurable space of the auxiliary labels. The auxiliary information we use in our problem setup is the domain indices, and hence $\mathcal{U} = [K]$. The classification function to be trained in GCL is $r_{\hat{\boldsymbol{F}},\boldsymbol{\psi}}(z, u) := \sum_{j=1}^d \psi_j(\hat{\boldsymbol{F}}^{-1}(z)_j, u)$ consisting of $(\hat{\boldsymbol{F}}, \psi)$, and the classification task of GCL is to classify $(\boldsymbol{Z}_k^{\mathrm{Src}}, k)$ as positive and $(\boldsymbol{Z}_k^{\mathrm{Src}}, k')(k' \neq k)$ as negative when $\boldsymbol{Z}_k^{\mathrm{Src}} \in \mathcal{D}_k$. This yields the following domain-contrastive learning criterion to estimate $\boldsymbol{F}$:

$$\underset{\hat{\boldsymbol{F}}\in\mathcal{F},\psi_j\in\Psi(j\in[d])}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \left( \phi\left( r_{\hat{\boldsymbol{F}},\boldsymbol{\psi}}(\boldsymbol{Z}_{k,i}^{\mathrm{Src}}, k) \right) + \mathbb{E}_{k'\neq k} \phi\left( -r_{\hat{\boldsymbol{F}},\boldsymbol{\psi}}(\boldsymbol{Z}_{k,i}^{\mathrm{Src}}, k') \right) \right),$$

where $\mathcal{F}$ and $\Psi$ are sets of parametrized functions, $\mathbb{E}_{k' \neq k}$ denotes the expectation with respect to $k' \sim \mathrm{Unif}([K] \setminus k)$ ("Unif" denotes the uniform distribution), and $\phi$ is the logistic loss $\phi(m) := \log(1 + \exp(-m))$. The trained feature extractor $\hat{\boldsymbol{F}}$ is used as an estimator of $\boldsymbol{F}$. In experiments, $\mathcal{F}$ is implemented by invertible neural networks [145], $\Psi$ by multi-layer perceptron [99], and $\mathbb{E}_{k' \neq k}$ is replaced by a random sampling renewed for every mini-batch.

**Step 2: Extract and inflate the target ICs using $\hat{\boldsymbol{F}}$.** Second, the method uses the estimated $\hat{\boldsymbol{F}}$ to perform data augmentation of the target domain data based on the knowledge transferred from the source domains. The second step extracts and inflates the target domain ICs using the estimated $\hat{\boldsymbol{F}}$. We first extract the ICs of the target domain data by applying the inverse of $\hat{\boldsymbol{F}}$ as

$$\hat{s}_i = \hat{\boldsymbol{F}}^{-1}(\boldsymbol{Z}_i).$$

After the extraction, we inflate the set of IC values by taking all dimension-wise combinations of the estimated IC:

$$\bar{s}_{\boldsymbol{i}} = (\hat{s}_{i_1}^{(1)}, \ldots, \hat{s}_{i_d}^{(d)}), \quad \boldsymbol{i} = (i_1, \ldots, i_d) \in [n_{\mathrm{Tar}}]^d,$$

to obtain new plausible IC values $\bar{s}_{\boldsymbol{i}}$. The intuitive motivation of this procedure stems from the independence of the IC distributions. Theoretical justifications are provided in Section 4.4.

**Step 3: Synthesize target data from the inflated ICs.** The third step estimates the target risk $\mathcal{R}$ by the empirical distribution of the augmented data:

$$\check{\mathcal{R}}(h) := \frac{1}{n_{\mathrm{Tar}}^d} \sum_{\boldsymbol{i} \in [n_{\mathrm{Tar}}]^d} \ell(h, \hat{\boldsymbol{F}}(\bar{s}_{\boldsymbol{i}})), \tag{4.2}$$

and performs empirical risk minimization. In experiments, we used a regularization term $\Omega(\cdot)$ to control the complexity of $\mathcal{H}$ and select

$$\check{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \left\{ \check{\mathcal{R}}(h) + \Omega(h) \right\}. \tag{4.3}$$

The generated hypothesis $\check{h}$ is then used to make predictions in the target domain. In our experiments, we used $\Omega(h) = \lambda \|h\|^2$, where $\lambda > 0$ and the norm is that of the reproducing kernel Hilbert space (RKHS) which we take the subset $\mathcal{H}$ from. Note that we may well subsample only a subset of combinations in Equation (4.2) to mitigate the computation costs similarly to Clémençon et al. [48] and Papa et al. [197]. In practice, one may well use a weighted average of the risk estimator $\check{\mathcal{R}}$ and the empirical risk $\hat{\mathcal{R}}$ for training to mitigate the effect of the estimation error in $\hat{\boldsymbol{F}}$. That is, one may well replace $\check{\mathcal{R}}$ by $(1-\rho)\hat{\mathcal{R}} + \rho\check{\mathcal{R}}$ where $\rho \in [0, 1]$ in Equation (4.3). In our experiment, we solely used $\check{\mathcal{R}}$ (i.e., $\rho = 1$).

### 4.3.3 Implementation Details Based on Invertible Neural Networks

In Steps 2–3 of the proposed method (Section 4.3.2), we require access to $\hat{\boldsymbol{F}}^{-1}$ and $\hat{\boldsymbol{F}}$ of the estimator of $\boldsymbol{F}$, respectively. To ensure the existence of an inverse map as well as its tractability, we implement the function class $\mathcal{F}$ by *invertible neural networks* (INNs; [145]). As a recently emerged modeling technique, INNs did not have a theoretical guarantee on their representation power. In particular, whether they possess the *universal approximation property* for a wide range of invertible functions had not been elucidated. We dedicate Chapter 5 to introducing the theoretical results obtained

**Algorithm 3** Proposed method: Causal Mechanism Transfer

**Input:** Source domain data sets $\{\mathcal{D}_k\}_{k\in[K]}$, target domain data set $\mathcal{D}_{\mathrm{Tar}}$, nonlinear ICA algorithm ICA, and a learning algorithm $\mathcal{A}_{\mathcal{H}}$ to fit the hypothesis class $\mathcal{H}$ of predictors.

1: $\hat{\boldsymbol{F}} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$ ▷ Step 1. Estimate the shared transformation.

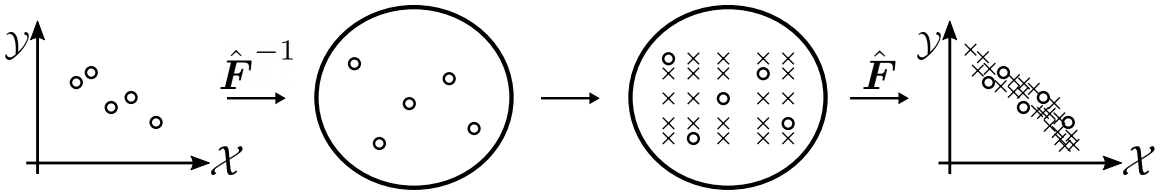
2: $\hat{s}_i \leftarrow \hat{\boldsymbol{F}}^{-1}(\boldsymbol{Z}_i), \quad (i = 1, \ldots, n_{\mathrm{Tar}})$ ▷ Step 2. Extract and

3: $\{\bar{s}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n_{\mathrm{Tar}}]^d} \leftarrow \mathrm{AllCombinations}(\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}})$ ▷ shuffle target ICs

4: $\bar{z}_{\boldsymbol{i}} \leftarrow \hat{\boldsymbol{F}}(\bar{s}_{\boldsymbol{i}})$ ▷ Step 3. Synthesize target data

5: $\check{h} \leftarrow \mathcal{A}_{\mathcal{H}}(\{\bar{z}_{\boldsymbol{i}}\}_{\boldsymbol{i}})$ ▷ and fit the predictor

**Output:** $\check{h}$: the predictor in the target domain.




(a) Labeled target data (b) Find IC (c) Shuffle (d) Pseudo target data

Figure 4.3: Schematic illustration of proposed few-shot domain adaptation method after estimating the common mechanism $\boldsymbol{F}$. With the estimated $\hat{\boldsymbol{F}}$, the method augments the small target domain sample in a few steps to enhance statistical efficiency: (a) The algorithm is given labeled target domain data. (b) From labeled target domain data, extract the ICs. (c) By shuffling the values, synthesize likely values of IC. (d) From the synthesized IC, generate pseudo target data. The generated data is used to fit a predictor for the target domain.

in this dissertation, which provides an additional layer of theoretical justification to the method proposed in this chapter.

## 4.4 Theoretical Analysis

In this section, we state two theorems to investigate the statistical properties of the method proposed in Section 4.3.2 and provide plausibility beyond the intuition that we take advantage of the independence of the IC distributions.

### 4.4.1 Complete-estimation Case: Minimum Variance Property

First, we consider the case that $\boldsymbol{F}$ has been estimated perfectly. While this is an idealistic case, the analysis provides us with the intuition that the proposed method helps the learner in terms of the *variance* of the risk estimator.

**Theorem 4.1** (Minimum variance property of $\check{\mathcal{R}}$)**.** *Assume that $\hat{\boldsymbol{F}} = \boldsymbol{F}$. Then, for each $h \in \mathcal{H}$, the proposed risk estimator $\check{\mathcal{R}}(h)$ is the uniformly minimum variance unbiased estimator of $\mathcal{R}(h)$, i.e., for any unbiased estimator $\tilde{\mathcal{R}}(h)$ of $\mathcal{R}(h)$,*

$$\forall q \in \mathcal{Q}, \quad \mathrm{Var}(\check{\mathcal{R}}(h)) \leq \mathrm{Var}(\tilde{\mathcal{R}}(h))$$

*as well as $\mathbb{E}_{p_{\mathrm{Tar}}}\check{\mathcal{R}}(h) = \mathcal{R}(h)$ holds.*

The proof of Theorem 4.1 is immediate once we rewrite $\mathcal{R}(h)$ as a $d$-variate regular statistical

$$
\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ d-1 \\ d \end{array}
\begin{array}{c}
\begin{array}{cccccc} \hat{S}_1 & \hat{S}_2 & \cdots & \hat{S}_{n-1} & \hat{S}_n \end{array} \\
\left[\begin{array}{ccccc}
\hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{1,n-1} & \hat{s}_{1n} \\
\hat{s}_{21} & \hat{s}_{22} & \cdots & \hat{s}_{2,n-1} & \hat{s}_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\hat{s}_{d-1,1} & \hat{s}_{d-1,2} & \cdots & \hat{s}_{d-1,n-1} & \hat{s}_{d-1,n} \\
\hat{s}_{d1} & \hat{s}_{d2} & \cdots & \hat{s}_{d,n-1} & \hat{s}_{dn}
\end{array}\right]
\end{array}
\begin{array}{c} \\ \longrightarrow \\ \longrightarrow \\ \longrightarrow \\ \longrightarrow \\ \longrightarrow \end{array}
\begin{pmatrix} \hat{s}_{1,n-1} \\ \hat{s}_{22} \\ \vdots \\ \hat{s}_{d-1,1} \\ \hat{s}_{d2} \end{pmatrix}
$$

Figure 4.4: Illustration of the proposed data augmentation procedure.

functional and $\check{\mathcal{R}}(h)$ as its corresponding generalized U-statistic [162]. Details can be found in Appendix C.5.

**Implications.** Theorem 4.1 implies that the proposed risk estimator can have superior statistical efficiency in terms of the variance over the ordinary empirical risk.

### 4.4.2 Incomplete-estimation Case: Excess Risk Bound

In real situations, one has to estimate $\boldsymbol{F}$. The following theorem characterizes the statistical gain and loss arising from the estimation error $\boldsymbol{F} - \hat{\boldsymbol{F}}$. The intuition is that the increased number of data points suppresses the possibility of overfitting because the hypothesis has to fit the majority of the inflated data, but the estimator $\hat{\boldsymbol{F}}$ has to be accurate so that fitting the inflated data is meaningful (Figure 4.5). Theorem 4.2 quantifies this consideration:

**Theorem 4.2** (Excess risk bound)**.** *Let $\check{h}$ be a minimizer of Equation (4.2). Under appropriate assumptions (see Theorem C.1 in Appendix C.4), for arbitrary $\delta, \delta' \in (0,1)$, we have with probability at least $1-(\delta+\delta')$,*

$$
\begin{aligned}
\mathcal{R}(\check{h}) - \mathcal{R}(h^*) \leq & \underbrace{C \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}}}_{\textit{Approximation error}} + \underbrace{4d\mathfrak{R}(\mathcal{H}) + 2dB_\ell \sqrt{\frac{\log 2/\delta}{2n}}}_{\textit{Estimation error}} \\
& + \underbrace{\kappa_1(\delta', n) + dB_\ell B_q \kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})}_{\textit{Higher order terms}}.
\end{aligned}
$$

*Here, $\|\cdot\|_{W^{1,1}}$ is the $(1,1)$-Sobolev norm, and we define the effective Rademacher complexity $\mathfrak{R}(\mathcal{H})$ by*

$$
\mathfrak{R}(\mathcal{H}) := \frac{1}{n} \mathbb{E}_{\hat{S}} \mathbb{E}_\sigma \left[ \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{n} \sigma_i \mathbb{E}_{S'_2, \ldots, S'_d} [\tilde{\ell}(\hat{s}_i, S'_2, \ldots, S'_d)] \right| \right], \tag{4.4}
$$

*where $\{\sigma_i\}_{i=1}^n$ are independent sign variables, $\mathbb{E}_{\hat{S}}$ is the expectation with respect to $\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}}$, the dummy variables $S'_2, \ldots, S'_d$ are i.i.d. copies of $\hat{s}_1$, and $\tilde{\ell}$ is defined by using the degree-$d$ symmetric group $\mathfrak{S}_d$ as*

$$
\tilde{\ell}(s_1, \ldots, s_d) := \frac{1}{d!} \sum_{\pi \in \mathfrak{S}_d} \ell(h, \hat{\boldsymbol{F}}(s^{(1)}_{\pi(1)}, \ldots, s^{(d)}_{\pi(d)})),
$$

*and $\kappa_1(\delta', n)$ and $\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$ are higher order terms. The constants $B_q$ and $B_\ell$ depend only on $q$*

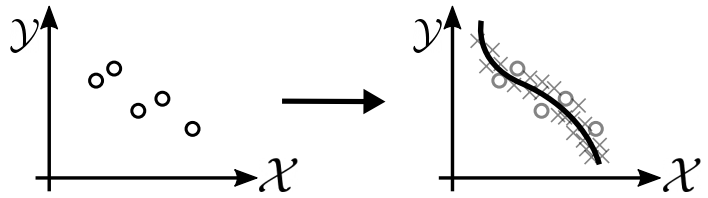



Figure 4.5: Fitting the data inflated by the proposed method. If the inflated data appear at the appropriate locations, the increment of the data has the effect of apparent complexity reduction since one can fit a complex predictor with less fear of overfitting. On the other hand, if the estimation of $\hat{\boldsymbol{F}}$ is poor, the fitting may be biased.

*and $\ell$, respectively, while $C$ depends only on $\boldsymbol{F}, q, \ell$, and $d$.*

Details of the statement and the proof can be found in Appendix C.4. The Sobolev norm [3] emerges from the evaluation of the difference between the estimated IC distribution and the ground-truth IC distribution. Note that the theorem is agnostic to how $\hat{\boldsymbol{F}}$ is obtained, hence it applies to more general problem setup as long as $\boldsymbol{F}$ can be estimated.

**Implications.** In Theorem 4.2, the utility of the proposed method appears in the effective complexity measure. The complexity is defined by a set of functions that are marginalized over all but one argument, resulting in mitigated dependence on the input dimensionality from exponential to linear (Remark C.3 in Appendix C.4).

### 4.4.3 Representation Power of Invertible Neural Networks

The feasibility of the proposed method (Algorithm 3) relies on the availability of a invertible model that is sufficiently flexible to approximate $\boldsymbol{F}$. Our proposal is to employ *invertible neural networks* (INNs) such as Glow architecture (Kingma and Dhariwal [145]). INNs have the virtue of having invertibility by design, and they provide explicit formula for calculating the inverse map. However, as a recently emerged modeling technique, no theoretical results had been known (prior to this project) to guarantee that INNs had sufficient flexibility to approximate complex invertible maps. This was a critical concern that could have determined whether the proposed method could be confidently applied to a broad range of application domains.

In this project, we have shown that certain classes of INNs, which includes Glow, have the *universal approximation property* for a rather large class of smooth invertible maps. This is a result that provides an additional layer of theoretical justification to the work of this chapter. We elaborate on this theoretical analysis in Chapter 5 because the scope of the results is not limited to providing a theoretical justification to the work of this chapter, and because the theoretical result is (intricate and) interesting by itself.

## 4.5 Experimental Evaluation

In this section, we provide the results of proof-of-concept experiments to demonstrate the effectiveness of the proposed approach. Note that the primary purpose of the experiments is to confirm whether the proposed method can properly perform DA in real-world data, and it is not to determine which DA method and TA are the most suited for the specific dataset.

### 4.5.1 Design of the Experiment

**Dataset.** We used the gasoline consumption data [93, p.284, Example 9.5], which is a panel data of gasoline usage in 18 of the OECD countries over 19 years. We considered each country as a domain, and we disregarded the time-series structure and considered the data as i.i.d. samples for each country in this proof-of-concept experiment. The dataset contains four variables, all of which were log-transformed: the motor gasoline consumption per car (the predicted variable), per-capita income, the motor gasoline price, and the stock of cars per capita (the predictor variables) [12]. For further details of the data, see Appendix C.2. We used the dataset because there are very few public datasets for domain adapting regression tasks [50] especially for multi-source DA, and also because the dataset has been used in econometric analyses involving SEMs [11], conforming to our approach.

**Compared methods.** We compared the following DA methods, all of which can be applied to regression problems. Unless explicitly specified, the predictor class $\mathcal{H}$ is chosen to be kernel ridge regression (KRR; see, e.g., [234]) with the same hyperparameter candidates as the proposed method (Section 4.5.2). Further details are described in Appendix C.2.5.

- Naive baselines (*SrcOnly*, *TarOnly*, and *S&TV*): *SrcOnly* (resp. *TarOnly*) trains a predictor on the source domain data (resp. target training data) without any device. *SrcOnly* can be effective if the source domains and the target domain have highly similar distributions. The *S&TV* baseline trains on both source and target domain data, but the LOOCV score is computed only from the target domain data.

- *TrAdaBoost*: Two-stage TrAdaBoost.R2; a boosting method tailored for few-shot regression transfer proposed in Pardoe and Stone [200]. It is an iterative method with early-stopping [200], for which we use the leave-one-out cross-validation score on the target domain data as the criterion. As suggested in Pardoe and Stone [200], we set the maximum number of outer loop iterations at 30. The base predictor is the decision tree regressor with the maximum depth 6 [99]. Note that although TrAdaBoost does not have a clarified transfer assumption, we compared the performance for reference.

- *IW*: Importance-weighted KRR using RuLSIF [305]. The method directly estimates a relative joint density ratio function $\frac{p_{\mathrm{Tar}}(x,y)}{\alpha p_{\mathrm{Tar}}(x,y)+(1-\alpha)p_{\mathrm{Src}}(x,y)}$ for $\alpha \in [0,1)$, where $p_{\mathrm{Src}}$ is a hypothetical source distribution created by pooling all source domain data. Following Yamada et al. [305], we experimented on $\alpha \in \{0, 0.5, 0.95\}$. The results are similar across these values, and we reported the results of 0.5 which performed the best among the three. The regularization coefficient $\lambda'$ was selected from $\lambda' \in 2^{\{-10,\dots,10\}}$ using importance-weighted cross-validation [263].

- *GDM*: Generalized discrepancy minimization [51]. This method performs instance-weighted training on the source domain data with the weights that minimize the *generalized discrepancy* (via quadratic programming). We selected the hyper-parameters $\lambda_r$ from $2^{\{-10,\dots,10\}}$ as suggested in Cortes et al. [51]. The selection criterion is the performance of the trained predictor on the target training labels as the method trains on the source domain data and the target unlabeled data.

- *Copula*: The non-parametric regular-vine copula method [172]. This method presumes to use a specific joint density estimator called regular-vine (R-vine) copulas. Adaptation is realized in two steps: the first step estimates which components of the constructed R-vine model are different by performing two-sample tests based on maximum mean discrepancy [172], and the

second step re-estimates the components in which a change is detected using only the target domain data.

- *LOO* (reference score): The leave-one-out cross-validated error estimate was also calculated for reference. It is the average prediction error for a single held-out test point when the predictor is trained on the rest of the target domain data.

**Evaluation procedure.** The prediction accuracy was measured by the mean squared error (MSE). For each train-test split, we randomly selected one-third (6 points) of the target domain dataset as the training set and use the rest as the test set. All experiments were repeated 10 times with different train-test splits of target domain data.

### 4.5.2 Details of the Experiment

**Estimation of $\boldsymbol{F}$ (Step 1).** We modeled $\mathcal{F}$ (i.e., the class of $\hat{\boldsymbol{F}}$) by an 8-layer Glow neural network (Appendix C.2.2). We modeled $\Psi$ (i.e., the class of $\{\psi_j\}_{j=1}^d$) by a 1-hidden-layer neural network with a varied number of hidden units, $K$ output units, and the rectified linear unit activation [161]. We used its $k$-th output ($k \in [K]$) as the value for $\psi_j(\cdot, k)$. For training, we used the Adam optimizer [144] with fixed parameters $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$, fixed initial learning rate $10^{-3}$, and the maximum number of epochs 300. The other fixed hyperparameters of $\hat{\boldsymbol{F}}$ and its training process are described in Appendix C.2.

**Augmentation of target data (Step 3).** For each evaluation step, we took all combinations (with replacement) of the estimated ICs to synthesize target domain data. After we synthesized the data, we filtered them by applying a novelty detection technique with respect to the union of source domain data. Namely, we used the one-class support vector machine [235] with the fixed parameter $\nu = 0.1$ and radial basis function (RBF; see, e.g., [234]) kernel $k(x, y) = \exp(-\|x - y\|^2/\gamma)$ with $\gamma = d$. This is because the estimated transform $\hat{\boldsymbol{F}}$ is not expected to be trained well outside the union of the supports of the source distributions. After performing the filtration, we combined the original target training data with the augmented data to ensure the original data points to be always included.

**Predictor hypothesis class $\mathcal{H}$.** As the predictor model, we used the KRR with RBF kernel. The bandwidth $\gamma$ was chosen by the median heuristic similarly to Yamada et al. [305] for simplicity. Note that the choice of the predictor model is for the sake of comparison with the other methods tailored for KRR [51], and that an arbitrary predictor hypothesis class and learning algorithm can be easily combined with the proposed approach.

**Hyperparameter selection.** We performed grid-search for hyperparameter selection. The number of hidden units for $\psi$ was chosen from $\{10, 20\}$ and the coefficient of weight-decay from $10^{\{-2,-1\}}$. The $\ell^2$ regularization coefficient $\lambda$ of KRR was chosen from $\lambda \in 2^{\{-10,\dots,10\}}$ following Cortes et al. [51]. To perform hyperparameter selection as well as early-stopping, we recorded the leave-one-out cross-validation (LOOCV) mean-squared error on the target training data every 20 epochs and selected its minimizer. The leave-one-out score was computed using the well-known closed-form formula instead of training the predictor for each split (e.g., [219]). Note that we only used the original target domain data as the held-out set and not the synthesized data. In practice, if the target domain data is extremely few, one may well use *percentile-cv* [193] to mitigate overfitting of hyperparameter selection.

**Computation environment.** All experiments were conducted on an Intel Xeon(R) 2.60 GHz CPU with 132 GB memory. They were implemented in Python using the PyTorch library [201] or the R language [214].

### 4.5.3 Experimental Results

In Table 4.1, we report the MSE scores normalized by that of *LOO* to facilitate the comparison, similarly to Cortes and Mohri [50]. In many of the target domain choices, the naive baselines (*SrcOnly* and *S&TV*) suffered from negative transfer, i.e., higher average MSEs than *TarOnly* (in 12 out of 18 domains). On the other hand, the proposed method performed better than *TarOnly* or was more resistant to negative transfer than the other compared methods. The performances of *GDM*, *Copula*, and *IW* were often inferior even compared to the baseline performance of *S&TV*. For *GDM* and *IW*, this can be attributed to the fact that these methods presume the availability of abundant (unlabeled) target domain data, which was unavailable in the current problem setup. For *Copula*, the performance inferior to the naive baselines was possibly due to the restriction of the predictor model to its accompanied probability model [172]. *TrAdaBoost* worked reasonably well for many but not all domains. For some domains, it suffered from negative transfer similarly to others, possibly because of the very small number of training data points. Note that the transfer assumption of TrAdaBoost has not been stated [200], and it is not clear when the method is reliable. The domains on which the baselines perform better than the proposed method can be explained by the following two cases: (i) easier domains allow naive baselines to perform well and (ii) some domains may have deviated $\boldsymbol{F}$. Case (i) implies that estimating $\boldsymbol{F}$ is unnecessary, and hence the proposed method can be suboptimal (more likely for JPN, NLD, NOR, and SWE in Table 4.1, where SrcOnly or S&TV improved upon TrgOnly). On the other hand, case (ii) implies that an approximation error was induced as in Theorem 4.2 (more likely for IRL and ITA in Table 4.1). In this case, others also perform poorly, implying the difficulty of the problem instance. In either case, in practice, one may well perform cross-validation to fall back into the baselines.

### 4.5.4 Synthetic-data Experiment

Here, we experimentally evaluate the robustness of the proposed method to the misspecification of the invertibility assumption of the mixing function $\boldsymbol{F}$. The invertibility assumption is exploited in this paper in two ways: (i) as part of the identifiability condition of $\boldsymbol{F}$ [124] and (ii) as a structure that allows us to take advantage of the statistical independence in $S_i$. To do so, we use synthetic data for which the ground-truth mixing function is known, and we tweak the data-generating process.

**Data generation.** The synthetic data sets are generated as follows. First, we fix $r \in [d]$, and we randomly generate an invertible neural network $\tilde{\boldsymbol{F}}$ and a rank-$r$ matrix $B \in \mathbb{R}^{d \times d}$. Then, we define $\boldsymbol{F} = \frac{1}{a}\tilde{\boldsymbol{F}} \circ B$, where $a > 0$ is a scaling parameter to standardize the scale of the generated data. After randomly generating $\boldsymbol{F}$, we used it to generate both the source domain data and the target domain data by applying $\boldsymbol{F}$ to the ICs. See Appendix C.3 for the details of the sampling procedure.

When $r = d$, i.e., $B$ is a regular matrix, the mixing map $\boldsymbol{F} = \frac{1}{a}\tilde{\boldsymbol{F}} \circ B$ is invertible. On the other hand, when $r \neq d$, i.e., $B$ is a singular matrix, $\boldsymbol{F}$ is no longer invertible, and hence the invertibility assumption of the data generating process is violated. By varying $r \in [d]$, we control the level of *ill-conditioning*: $r = d$ means that there is no violation of the invertibility assumption, and a smaller $r$ makes the estimation of the original ICs more difficult.

**Evaluation.** We varied $r$ in $[d]$ and conducted 30 independent runs of the experiment for each $r$. We also performed a sampling in the spirit of *incomplete U-statsitics* [197, 220, 48] instead of

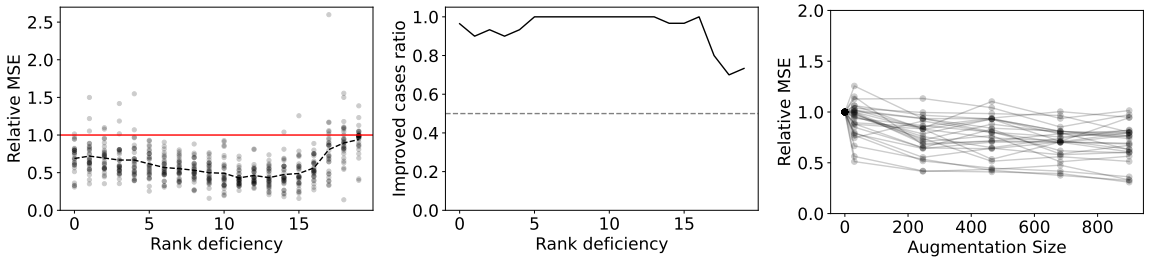
Table 4.1: Results of the real-world data experiments for different choices of the target domain. The evaluation score was MSE normalized by that of *LOO* (the lower the better). All experiments were repeated 10 times with different train-test splits of target domain data, and the average performance is reported with the standard errors in the brackets. The target column indicates abbreviated country names. Bold-face indicates the best score (Prop: proposed method, TrAda: *TrAdaBoost*, the numbers in the brackets of IW indicate the value of $\alpha$). The proposed method often improved upon the baseline *TarOnly* or was relatively more resistant to negative transfer, with notable improvements in *DEU*, *GBR*, and *USA*.

| Target | (LOO) | TarOnly | **Prop** | SrcOnly | S&TV | TrAda | GDM | Copula | IW(.5) |
|---|---|---|---|---|---|---|---|---|---|
| AUT | 1 | 5.88 (1.60) | **5.39 (1.86)** | 9.67 (0.57) | 9.84 (0.62) | 5.78 (2.15) | 31.56 (1.39) | 27.33 (0.77) | 34.06 (0.67) |
| BEL | 1 | 10.70 (7.50) | **7.94 (2.19)** | 8.19 (0.68) | 9.48 (0.91) | 8.10 (1.88) | 89.10 (4.12) | 119.86 (2.64) | 105.68 (3.13) |
| CAN | 1 | 5.16 (1.36) | **3.84 (0.98)** | 157.74 (8.83) | 156.65 (10.69) | 51.94 (30.06) | 516.90 (4.45) | 406.91 (1.59) | 571.33 (1.60) |
| DNK | 1 | 3.26 (0.61) | **3.23 (0.63)** | 30.79 (0.93) | 28.12 (1.67) | 25.60 (13.11) | 16.84 (0.85) | 14.46 (0.79) | 21.83 (0.93) |
| FRA | 1 | 2.79 (1.10) | **1.92 (0.66)** | 4.67 (0.41) | 3.05 (0.11) | 52.65 (25.83) | 91.69 (1.34) | 156.29 (1.96) | 113.5 (1.15) |
| DEU | 1 | 16.99 (8.04) | **6.71 (1.23)** | 229.65 (9.13) | 210.59 (14.99) | 341.03 (157.80) | 739.29 (11.81) | 929.03 (4.85) | 807.88 (4.14) |
| GRC | 1 | 3.80 (2.21) | **3.55 (1.79)** | 5.30 (0.90) | 5.75 (0.68) | 11.78 (2.36) | 26.90 (1.89) | 23.05 (0.53) | 39.56 (1.70) |
| IRL | 1 | **3.05 (0.34)** | 4.35 (1.25) | 135.57 (5.64) | 12.34 (0.58) | 23.40 (17.50) | 3.84 (0.22) | 26.60 (0.59) | 5.79 (0.12) |
| ITA | 1 | **13.00 (4.15)** | 14.05 (4.81) | 35.29 (1.83) | 39.27 (2.52) | 87.34 (24.05) | 226.95 (11.14) | 343.10 (10.04) | 237.15 (6.46) |
| JPN | 1 | 10.55 (4.67) | 12.32 (4.95) | **8.10 (1.05)** | 8.38 (1.07) | 18.81 (4.59) | 95.58 (7.89) | 71.02 (5.08) | 129.3 (10.47) |
| NLD | 1 | 3.75 (0.80) | 3.87 (0.79) | **0.99 (0.06)** | 0.99 (0.05) | 9.45 (1.43) | 28.35 (1.62) | 29.53 (1.58) | 33.38 (1.63) |
| NOR | 1 | 2.70 (0.51) | 2.82 (0.73) | 1.86 (0.29) | **1.63 (0.11)** | 24.25 (12.50) | 23.36 (0.88) | 31.37 (1.17) | 27.09 (0.76) |
| ESP | 1 | 5.18 (1.05) | 6.09 (1.53) | 5.17 (1.14) | **4.29 (0.72)** | 14.85 (4.20) | 33.16 (6.99) | 152.59 (6.19) | 56.54 (2.16) |
| SWE | 1 | 6.44 (2.66) | 5.47 (2.63) | 2.48 (0.23) | **2.02 (0.21)** | 2.18 (0.25) | 15.53 (2.59) | 2706.85 (17.91) | 113.55 (1.72) |
| CHE | 1 | 3.51 (0.46) | **2.90 (0.37)** | 43.59 (1.77) | 7.48 (0.49) | 38.32 (9.03) | 8.43 (0.24) | 29.71 (0.53) | 9.33 (0.22) |
| TUR | 1 | 1.65 (0.47) | 1.06 (0.15) | 1.22 (0.18) | **0.91 (0.09)** | 2.19 (0.34) | 64.26 (5.71) | 142.84 (2.04) | 139.29 (2.41) |
| GBR | 1 | 5.95 (1.86) | **2.66 (0.57)** | 15.92 (1.02) | 10.05 (1.47) | 7.57 (5.10) | 50.04 (1.75) | 68.70 (1.25) | 69.19 (0.87) |
| USA | 1 | 4.98 (1.96) | **1.60 (0.42)** | 21.53 (3.30) | 12.28 (2.52) | 2.06 (0.47) | 308.69 (5.20) | 244.90 (1.82) | 393.45 (1.68) |
| #Best | - | 2 | 10 | 2 | 4 | 0 | 0 | 0 | 0 |

generating the augmented data for all possible $n_{\mathrm{Tar}}^d$ combinations. For the predictor model class, we employed the gradient boosted regression trees [81, 41]. For the proposed method, after estimating $\boldsymbol{F}$, we generated an augmented data set by sampling $n_{\mathrm{aug}}$ points, and we trained a predictor using the augmented data set in addition to the original target-domain data $\mathcal{D}_{\mathrm{Tar}}$. We then measured the MSE, denoted by $\mathrm{MSE}_{\mathrm{prop}}$, on the target-domain testing data set $\mathcal{D}_{\mathrm{test}}$. As the baseline of comparison, we also measured the performance, denoted by $\mathrm{MSE}_{\mathrm{base}}$, of a predictor trained only on the original target-domain data $\mathcal{D}_{\mathrm{Tar}}$, and we reported the relative MSE, i.e., $\mathrm{relMSE} := \frac{\mathrm{MSE}_{\mathrm{prop}}}{\mathrm{MSE}_{\mathrm{base}}}$.

**Results.** Figure 4.6 illustrates the experiment results with $d = 20$, $K = 42$, $n_k = 512$, and $n_{\mathrm{Tar}} = 30$. In Appendix C.3.4, we additionally report the results for $n_{\mathrm{Tar}} \in \{10, 20\}$, in both of which cases the results were similar.

Figure 4.6(a) reports relMSE in relation to the rank deficiency $d - r$. The rank deficiency $d - r$ was varied in $\{0, \ldots, d-1\}$ while the size of the data augmentation was fixed at $n_{\mathrm{aug}} = n_{\mathrm{Tar}}^2$. Figure 4.6(b) visualizes the ratio of cases in which relMSE $< 1$. In these figures, we can observe that the proposed method has some robustness to the violation of the invertibility assumption, and that it maintains a performance improvement (i.e., relMSE $< 1$) as the rank deficiency $d-r$ increases up to 15. When the rank deficiency is above 15, we observe a higher ratio of degraded performance.




(a) relMSE in relation to rank deficiency $d - r$.

(b) Ratio of improved cases (relMSE $< 1$) in relation to rank deficiency $d - r$.

(c) relMSE in relation to augmentation size $n_{\mathrm{aug}}$.

Figure 4.6: Results of the synthetic-data experiments ($n_{\mathrm{Tar}} = 30$). (a) The plotted points represent independent runs of the experiment. The dotted curve indicates the mean, and the horizontal line is drawn at relMSE $= 1$. That is, if a point is below this line, it indicates that the proposed method yielded a smaller error than the baseline. (b) The dotted horizontal line is at 0.5. That is, if a point is above this line, it indicates that a majority of cases had a performance improvement due to the proposed method. For (a) and (b), we fixed the augmented data size at $n_{\mathrm{aug}} = n_{\mathrm{Tar}}^2$. For (c), we fixed the rank deficiency at $d - r = 0$.

These results imply that the proposed method may be reliable to some extent unless the violation of the invertibility is severe, i.e., the considered problem is such that the estimation of the mixing map is highly ill-conditioned.

Figure 4.6(c) shows relMSE in relation to the augmentation size of the proposed method. For each independent run of the experiment, after estimating $\boldsymbol{F}$, we sampled the augmented data sets of varying sizes $n_{\mathrm{aug}}$, and for each augmented data set, we trained a predictor and evaluated its relMSE. The augmentation size was varied in $n_{\mathrm{aug}} \in \{(1-\alpha)n_{\mathrm{Tar}} + \alpha n_{\mathrm{Tar}}^2\}_{\alpha \in \{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}}$, i.e, from $n_{\mathrm{Tar}}$ to $n_{\mathrm{Tar}}^2$, while the rank deficiency was fixed at $d - r = 0$. We can observe that sampling $n_{\mathrm{aug}} = n_{\mathrm{Tar}}$ points already yields a similar performance improvement as $n_{\mathrm{aug}} = n_{\mathrm{Tar}}^2$. This conforms with the claims of the existing work on incomplete U-statistics [220, 197, 48] that sampling $\mathcal{O}(n_{\mathrm{Tar}})$ terms with a uniform distribution is an effective strategy for learning with an incomplete U-statistic.

## 4.6 Related Work and Discussion

In this section, we review some existing TAs for DA to clarify the relative position of this chapter. We also clarify the relation to the literature of causality-related transfer learning.

### 4.6.1 Existing Transfer Assumptions

Here, we review some of the existing work and TAs. See Table 4.2 for a summary.

**(i) Parametric assumptions.** Some TAs assume parametric distribution families, e.g., Gaussian mixture model in covariate shift [259]. Some others assume parametric distribution shift, i.e., parametric representations of the target distribution given the source distributions. Examples include location-scale transform of class conditionals [313, 91], linearly dependent class conditionals [311], and low-dimensional representation of the class conditionals after kernel embedding [257]. In some applications, e.g., remote sensing, some parametric assumptions have proven useful [313].

**(ii) Invariant conditionals and marginals.** Some methods assume invariance of certain conditionals or marginals [213], e.g., $p(Y|X)$ in the covariate shift scenario [241], $p(Y|\mathcal{T}(X))$ for an

Table 4.2: Comparison of TAs for DA (*Parametric*: parametric distribution family or distribution shift, *Invariant dist.*: invariant distribution components such as conditionals, marginals, or copulas. *Disc./IPM*: small discrepancy or integral probability metric, *Param-transfer*: existence of transferable parameter, *Mechanism*: invariant mechanism). AD: adaptation among Apparently Different distributions is accommodated. NP: Non-Parametrically flexible. The numbers indicate the paragraphs of Section 4.6.1.

| TA | AD | NP | Suited application example |
|---|---|---|---|
| (i) Parametric | ✓ | - | Remote sensing |
| (ii) Invariant dist. | - | ✓ | Brain computer interface |
| (iii) Disc./IPM | - | ✓ | Computer vision |
| (iv) Param-transfer | ✓ | ✓ | Computer vision |
| (Ours) Mechanism | ✓ | ✓ | Medical records |

appropriate feature transformation $\mathcal{T}$ in transfer component analysis [196], $p(Y|\mathcal{T}(X))$ for a feature selector $\mathcal{T}$ [223, 176], $p(X|Y)$ in the target shift (TarS) scenario [313, 194], and few components of regular-vine copulas and marginals in Lopez-paz et al. [172]. For example, the covariate shift scenario has been shown to fit well to brain computer interface data [263].

**(iii) Small discrepancy or integral probability metric.** Another line of work relies on certain distributional similarities, e.g., integral probability metric [53] or hypothesis-class dependent discrepancies [20, 27, 19, 152, 314, 51]. These methods assume the existence of the *ideal joint hypothesis* [19], corresponding to a relaxation of the covariate shift assumption. These TA are suited for unsupervised or semi-supervised DA in computer vision applications [53].

**(iv) Transferable parameter.** Some others consider parameter transfer [151], where the TA is the existence of a parameterized feature extractor that performs well in the target domain for linear-in-parameter hypotheses and its learnability from the source domain data. For example, such a TA has been known to be useful in natural language processing or image recognition [163, 151].

### 4.6.2 Causality for Transfer Learning

Our method can be seen as the first attempt to fully leverage structural causal models for DA. Most of the causality-inspired DA methods express their assumptions in the level of *graphical causal models* (GCMs), which only has much coarser information than *structural causal models* (SCMs) [208, Table 1.1] exploited in this paper. Compared to previous work, our method takes one step further to assume and exploit the invariance of SCMs. Specifically, many studies assume the GCM $X \leftarrow Y$ (the *anticausal* scenario) following the seminal meta-analysis of Schölkopf et al. [232] and use it to motivate their parametric distribution shift assumptions or the parameter estimation procedure [313, 311, 91, 90]. Although such assumptions on the GCM have the virtue of being more robust to misspecification, they tend to require parametric assumptions to obtain theoretical justifications. On the other hand, our assumption enjoys a theoretical guarantee without relying on parametric assumptions.

One notable work in the existing literature is Magliacane et al. [176] that considered the domain adaptation among *different intervention states*, a problem setup that complements ours that considers an intervention-free (or identical intervention across domains) case. To model intervention states, Magliacane et al. [176] also formulated the problem setup using SCMs, similarly to the present paper. Therefore, we clarify a few key differences between Magliacane et al. [176] and our

work here. In terms of the methodology, Magliacane et al. [176] takes a variable selection approach to select a set of predictor variables with an invariant conditional distribution across different intervention states. On the other hand, our method estimates the SEMs (in the reduced form) and applies a data augmentation procedure to transfer the knowledge. To the best of our knowledge, the present paper is the first to propose a way to directly use the estimated SEMs for domain adaptation, and the fine-grained use of the estimated SEMs enables us to derive an excess risk bound. In terms of the plausible applications, their problem setup may be more suitable for application fields with interventional experiments such as genomics, whereas ours may be more suited for fields where observational studies are more common such as health record analysis [304] or economics [253]. In Appendix C.6, we provide a more detailed comparison.

### 4.6.3 Plausibility of the Assumptions

**Checking the validity of the assumption.** As is often the case in DA, the scarcity of data disables data-driven testing of the TAs, and we need domain knowledge to judge the validity. For our TA, the intuitive interpretation as invariance of causal models (Section 4.2) can be used.

**Invariant causal mechanisms.** The invariance of causal mechanisms has been exploited in recent work of causal discovery such as Xu et al. [302] and Monti et al. [185], or under the name of the *multi-environment setting* in Ghassami et al. [82]. Moreover, the SEMs are normally assumed to remain invariant unless explicitly intervened in [119]. However, the invariance assumption presumes that the intervention states do not vary across domains (allowing for the intervention-free case), which can be limiting for some applications where different interventions are likely to be present, e.g., different treatment policies being put in place in different hospitals. Nevertheless, the present work can already be of practical interest if it is combined with the effort to find suitable data or situations. For instance, one may find medical records in group hospitals where the same treatment criteria is put in place or local surveys in the same district enforcing identical regulations. In future work, relaxing the requirement to facilitate the data-gathering process is an important area. For such future extensions, the present theoretical analyses can also serve as a landmark to establish what can be guaranteed in the basic case without mechanism alterations.

**Fully observed variables.** As the first algorithm in the approach to fully exploit SCMs for DA, we also consider the case where all variables are observable. Although it is often assumed in a causal inference problem that there are some unobserved confounding variables, we leave further extension to such a case for future work.

**Required number of source domains.** A potential drawback of the proposed method is that it requires a number of source domains in order to satisfy the identification condition of the nonlinear ICA, namely GCL in this paper (Appendix C.1). The requirement solely arises from the identification condition of the ICA method and therefore has the possibility to be made less stringent by the future development of nonlinear ICA methods. Moreover, if one can accept other identification conditions, one-sample ICA methods (e.g., linear ICA) can also be used in the proposed approach in a straightforward manner, and our theoretical analyses still hold regardless of the method chosen.

**Flexibility of the model.** The relation between $X$ and $Y$ can drastically change while $\boldsymbol{F}$ is invariant. For example, even in a simple additive noise model $(X, Y) = \boldsymbol{F}(S_1, S_2) = (S_1, S_1 + S_2)$, the conditional $p(Y|X)$ can shift drastically if the distribution of the independent noise $S_2$ changes in a complex manner, e.g., becoming multimodal from unimodal.

## 4.7 Conclusion

In this chapter, we proposed a novel few-shot supervised DA method for regression problems based on the assumption of shared generative mechanism. Through theoretical and experimental analysis, we demonstrated the effectiveness of the proposed approach. By considering the latent common structure behind the domain distributions, the proposed method successfully induces positive transfer even when a naive usage of the source domain data can suffer from negative transfer. Our future work includes making an experimental comparison with extensively more datasets and methods as well as an extension to the case where the underlying mechanism are not exactly identical but similar. In Chapter 6, we discuss further possibilities of future research directions.

# Chapter 5

# Theoretical Analysis of the Representation Power of Invertible Neural Networks

In this chapter, we elaborate on a theoretical justification of the method introduced in Chapter 4. The method introduced in Chapter 4 relies on *invertible neural networks* (INNs), which have recently emerged from the field of generative modeling. However, as a recently emerged modeling technique, no theoretical results had been known (prior to this project) to guarantee that INNs had sufficient flexibility to approximate complex invertible maps. This was a critical concern which could determine whether the proposed method could be applied in a broad range of application domains. To address this point, this chapter presents the theoretical results we obtained to guarantee the representation power of INNs. The theoretical guarantees are based on the notion of *universal approximation property* (or *universality*), which roughly states that, any target function of interest can be approximated to any precision by a model class on any (bounded) input region of interest.

## 5.1 Overview

Invertible neural networks based on coupling flows (CF-INNs) are neural network architectures with invertibility by design [198, 148]. Endowed with the analytic-form invertibility and the tractability of the Jacobian, CF-INNs have demonstrated their usefulness in various machine learning tasks such as generative modeling [63, 145, 195, 143, 315], probabilistic inference [16, 288, 173], solving inverse problems [7], and feature extraction and manipulation [145, 192, 130, 267].

### 5.1.1 Motivation

The attractive properties of CF-INNs come at the cost of potential restrictions on the set of functions that they can approximate because they rely on carefully designed network layers. To circumvent the potential drawback, a variety of layer designs have been proposed to construct CF-INNs with high representation power, e.g., the affine coupling flow [62, 63, 145, 199, 146], the neural autoregressive flow [115, 59, 108], and the polynomial flow [131], each demonstrating enhanced empirical performance.

Despite the diversity of layer designs [198, 148], the theoretical understanding of the representation power of CF-INNs has been limited. Indeed, the most basic property as a function approximator, namely the *universal approximation property* (or *universality* for short) [56, 111], has not

been elucidated for CF-INNs. The universality can be crucial when CF-INNs are used to learn an invertible transformation (e.g., feature extraction [192] or independent component analysis [267]) because, informally speaking, lack of universality implies that there exists an invertible transformation, even among well-behaved ones, that CF-INN can never approximate, and it would render the model class unreliable for the task of function approximation.

### 5.1.2 Idea

To elucidate the universality of CF-INNs, we first prove a theorem to show the equivalence of the universality for certain diffeomorphism classes, which allows us to reduce the approximation of a general diffeomorphism to that of a much simpler one. By leveraging this problem reduction, we show that CF-INNs based on *affine coupling flows* (ACFs; see Section 5.2), one of the least expressive flow designs, are in fact universal approximators for a general class of diffeomorphisms. The result can be interpreted as a convenient (sufficient) condition to check the universality of a CF-INN: if the flow design can represent ACFs as special cases, then it is universal.

The difficulty in proving the universality of CF-INNs lies in two complications, and following are the approach to overcoming them in this chapter.

**(i) Only function composition can be used to make complex approximators (e.g., linear combination is not allowed).** We overcome this complication by essentially decomposing a general diffeomorphism into much simpler ones, by using a structural theorem of differential geometry that elucidates the structure of a certain diffeomorphism group. Our equivalence theorem provides an interface to implicitly take advantage of this technique.

**(ii) The flow layers tend to be inflexible due to the parametric restrictions.** As an extreme example, ACFs can only apply a uniform transformation along the transformed dimension, i.e., the parameter of the transformation cannot depend on the variable which undergoes the transformation. For ACFs, the reduction of the problem allows us to find an approximator with a clear outlook by approximating a step function.

### 5.1.3 Contributions

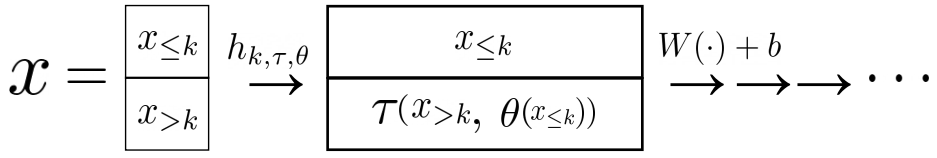
Our contributions are summarized as follows.

1. We present a theorem to show the equivalence of universal approximation properties for certain classes of functions. The result enables the reduction of the task of proving the universality for general diffeomorphisms to that for much simpler coordinate-wise ones.

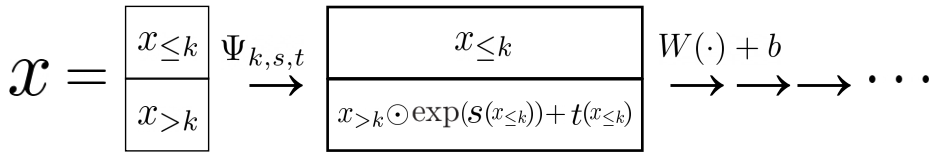
2. We leverage the result to show that some flow architectures, in particular even ACFs, can be used to construct a CF-INN with the universality for approximating a fairly general class of diffeomorphisms. This result can be seen as a convenient criterion to check the universality of a CF-INN: if the flow designs can reproduce ACF as a special case, it is universal.

3. As a corollary, we give an affirmative answer to a previously unsolved problem, namely the *distributional universality* [115, 131] of ACF-based CF-INNs.

Our result is an interesting application of a deep theorem in differential geometry to investigate the representation power of a neural network architecture.




(a) CF layer




(b) ACF layer

Figure 5.1: Illustration of coupling-based flow layers. CF layers are invertible by design; one can easily find the input vector given the output vector. Indeed, the first $k$ elements $\boldsymbol{x}_{\leq k}$ are the same as those of the output vector. Using this, one can recompute $\theta(\boldsymbol{x}_{\leq k})$. The last elements $\boldsymbol{x}_{>k}$ can be found by using the invertibility of $\tau(\cdot, \theta(\boldsymbol{x}_{\leq k}))$ which is assumed to be invertible given the parameter $\theta(\boldsymbol{x}_{\leq k})$. ACFs are examples of CFs using simple affine transformations as $\tau$.

**Theoretical implications to causal mechanism transfer.** The result of this chapter adds another layer to the theoretical guarantee of *causal mechanism transfer* (Algorithm 3), whose feasibility relies on the availability of a flexible model of invertible maps equipped with a tractable inverse.

## 5.2 Problem Setup

In this section, we prepare the formulation of the models and the notion of approximation called *universality*. The definitions will be used for stating the main results in Section 5.3.

### 5.2.1 Definitions of Models

In this section, we describe the models analyzed in this study, the notion of universality, and the goal of this chapter. Throughout the chapter, we fix $d \in \mathbb{N}$ and assume $d \geq 2$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ and $k \in [d-1]$, we define $\boldsymbol{x}_{\leq k}$ as the vector $(x_1, \ldots, x_k)^\top \in \mathbb{R}^k$ and $\boldsymbol{x}_{>k}$ the vector $(x_{k+1}, \ldots, x_d)^\top \in \mathbb{R}^{d-k}$.

The following are the important building blocks of CF-INNs. Figure 5.1 illustrates the definitions of CFs and ACFs.

**Coupling flows.** We define a coupling flow (CF; [198]) $h_{k,\tau,\theta}$ by $h_{k,\tau,\theta}(\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k}) = (\boldsymbol{x}_{\leq k}, \tau(\boldsymbol{x}_{>k}, \theta(\boldsymbol{x}_{\leq k}))$, where $k \in [d-1]$, $\theta \colon \mathbb{R}^k \to \mathbb{R}^l$ and $\tau : \mathbb{R}^{d-k} \times \mathbb{R}^l \to \mathbb{R}^{d-k}$ are maps, and $\tau(\cdot, \theta(\boldsymbol{y}))$ is an invertible map for any $\boldsymbol{y} \in \mathbb{R}^k$.

**Affine coupling flows.** One of the most standard types of CFs is *affine coupling flows* (ACFs; [63, 145, 146, 199]). We define an affine coupling flow $\Psi_{k,s,t} \colon \mathbb{R}^d \to \mathbb{R}^d$ by

$$\Psi_{k,s,t}(\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k}) = (\boldsymbol{x}_{\leq k}, \boldsymbol{x}_{>k} \odot \exp(s(\boldsymbol{x}_{\leq k})) + t(\boldsymbol{x}_{\leq k})),$$

where $k \in [d-1]$, $\odot$ is the Hadamard product, exp is applied in an element-wise manner, and $s, t : \mathbb{R}^k \to \mathbb{R}^{d-k}$ are maps typically parametrized by neural networks.

**Single-coordinate affine coupling flows.** Let $\mathcal{H}$ be a set of functions from $\mathbb{R}^{d-1}$ to $\mathbb{R}$. We define *$\mathcal{H}$-single-coordinate affine coupling flows* by $\mathcal{H}$-ACF $:= \{\Psi_{d-1,s,t} : s,t \in \mathcal{H}\}$, which is a subclass of ACFs. It is the least expressive flow design appearing in this dissertation, but we show in Section 5.3.2 that it can form a CF-INN with universality. We specify the requirements on $\mathcal{H}$ later.

**Invertible linear flows.** We define the set of all affine transforms by $\mathrm{Aff} := \{\boldsymbol{x} \mapsto A\boldsymbol{x} + b : A \in \mathrm{GL}, b \in \mathbb{R}^d\}$, where GL denotes the set of all regular matrices on $\mathbb{R}^d$.

We consider the invertible neural network architectures constructed by composing flow layers:

**Definition 5.1** (CF-INNs)**.** *Let $\mathcal{G}$ be a set consisting of invertible maps. We define the set of invertible neural networks based on $\mathcal{G}$ as*

$$\mathrm{INN}_{\mathcal{G}} := \{W_1 \circ g_1 \circ \cdots \circ W_n \circ g_n : \ n \in \mathbb{N}, g_i \in \mathcal{G}, W_i \in \mathrm{Aff}\}.$$

*When $\mathcal{G}$ can represent the addition of a constant vector, we can obtain the same set of maps by replacing* Aff *with* GL*, which has been adopted by previous studies such as Kingma and Dhariwal [145]. Moreover, it is possible to use only the symmetric group $\mathfrak{S}_d$ that is the permutations of variables, instead of* Aff*, when $\mathcal{G}$ contains $\mathcal{H}$-*ACF*. For details, see Appendix D.8.*

### 5.2.2 Notions of Universality

Here, we clarify the notion of universality in this chapter. First, we prepare some notation. Let $p \in [1, \infty)$ and $m, n \in \mathbb{N}$. For a measurable map $f : \mathbb{R}^m \to \mathbb{R}^n$ and a subset $K \subset \mathbb{R}^m$, we define

$$\|f\|_{p,K} := \left( \int_K \|f(x)\|^p \, dx \right)^{1/p},$$

where $\|\cdot\|$ is the Euclidean norm of $\mathbb{R}^n$. We also define $\|f\|_{\sup,K} := \sup_{x \in K} \|f(x)\|$.

**Definition 5.2** ($L^p$-/sup-universality)**.** *Let $\mathcal{M}$ be a model which is a set of measurable maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. Let $p \in [1, \infty)$, and let $\mathcal{F}$ be a set of measurable maps $f : U_f \to \mathbb{R}^n$, where $U_f$ is a measurable subset of $\mathbb{R}^m$ which may depend on $f$. We say that $\mathcal{M}$ is an $L^p$-*universal approximator *or* has the *$L^p$-universal approximation property *for $\mathcal{F}$ if for any $f \in \mathcal{F}$, any $\varepsilon > 0$, and any compact subset $K \subset U_f$, there exists $g \in \mathcal{M}$ such that $\|f - g\|_{p,K} < \varepsilon$. We define the* sup-universality *analogously by replacing $\|\cdot\|_{p,K}$ with $\|\cdot\|_{\sup,K}$.*

We also define the notion of distributional universality. Distributional universality has been used as a notion of theoretical guarantee in the literature of normalizing flows, i.e., probability distribution models constructed using invertible neural networks [148].

**Definition 5.3** (Distributional universality)**.** *Let $\mathcal{M}$ be a model which is a set of measurable maps from $\mathbb{R}^m$ to $\mathbb{R}^n$. We say that a model $\mathcal{M}$ is a* distributional universal approximator *or* has the distributional universal approximation property *if, for any absolutely continuous*[1] *probability measure $\mu$ on $\mathbb{R}^m$ and any probability measure $\nu$ on $\mathbb{R}^n$, there exists a sequence $\{g_i\}_{i=1}^{\infty} \subset \mathcal{M}$ such that $(g_i)_* \mu$ converges to $\nu$ in distribution as $i \to \infty$, where $(g_i)_* \mu := \mu \circ g_i^{-1}$.*

If a model $\mathcal{M}$ has the distributional universal approximation property, then it implies that $\mathcal{M}$ approximately transforms a known distribution, for example, the uniform distribution on $[0,1]^m$, into any probability measure $\mu$ on $\mathbb{R}^n$, not only absolutely continuous but singular one. There

[1] In this dissertation, we say a measure on the Euclidean space is *absolutely continuous* when it is absolutely continuous with respect to the Lebesgue measure.

exists another convention that defines the distributional universality as a representation power for only absolutely continuous probability measures. However, since absolutely continuous probability measures are dense in the set of all the probability measures, that convention is equivalent to ours. We include a proof for this fact in Lemma D.3 in Appendix D.1.

The different notions of universality are interrelated. Most importantly, the $L^p$-universality for a certain function class implies the distributional universality (see Lemma 5.1). Moreover, if a model $\mathcal{M}$ is a sup-universal approximator for $\mathcal{F}$, it is also an $L^p$-universal approximator for $\mathcal{F}$ for any $p \in [1, \infty)$.

### 5.2.3 Goal

Our goal is to elucidate the representation power of the CF-INNs for some flow architectures $\mathcal{G}$ by proving the $L^p$-universality or sup-universality of $\mathrm{INN}_{\mathcal{G}}$ for a fairly large class of *diffeomorphisms*, i.e., smooth invertible functions. To prove universality, we need to construct a model $g \in \mathrm{INN}_{\mathcal{G}}$ that attains the approximation error $\varepsilon$ for given $f$ and $K$.

## 5.3 Main Results

In this section, we present the main results on the universality of CF-INNs. The first theorem provides a general proof technique to simplify the problem of approximating diffeomorphisms, and the second theorem builds on the first to show that the CF-INNs based on the affine coupling are $L^p$-universal approximators.

### 5.3.1 General Result: Universality of Invertible Models

Our first main theorem allows us to lift a universality result for a restricted set of diffeomorphisms to the universality for a fairly general class of diffeomorphisms by showing a certain equivalence of universalities. By using the result to reduce the approximation problem, we can essentially circumvent the major complication in proving the universality of CF-INNs, namely that only function composition can be leveraged to make complex approximators (e.g., a linear combination is not allowed).

First, we define the following classes of invertible functions. Our main theorem later reveals an equivalence of $L^p$-universality/sup-universality for these classes.

**Definition 5.4** ($C^2$-diffeomorphisms: $\mathcal{D}^2$)**.** *We define $\mathcal{D}^2$ as the set of all $C^2$-diffeomorphisms $f : U_f \to \mathrm{Im}(f) \subset \mathbb{R}^d$ , where $U_f \subset \mathbb{R}^d$ is an open set $C^2$-diffeomorphic to $\mathbb{R}^d$, which may depend on $f$.*

**Definition 5.5** (Triangular transformations: $\mathcal{T}^\infty$)**.** *We define $\mathcal{T}^\infty$ as the set of all $C^\infty$-*increasing triangular *maps from $\mathbb{R}^d$ to $\mathbb{R}^d$. Here, a map $\tau = (\tau_1, \ldots, \tau_d) : \mathbb{R}^d \to \mathbb{R}^d$ is increasing triangular if each $\tau_k(\boldsymbol{x})$ depends only on $\boldsymbol{x}_{\leq k}$ and is strictly increasing with respect to $x_k$.*

**Definition 5.6** (Single-coordinate transformations: $\mathcal{S}_{\mathrm{c}}^r$)**.** *We define $\mathcal{S}_{\mathrm{c}}^r$ as the set of all compactly-supported $C^r$-diffeomorphisms that alter only the last coordinate, i.e., those $\tau$ satisfying $\tau(\boldsymbol{x}) = (x_1, \ldots, x_{d-1}, \tau_d(\boldsymbol{x}))$. In this article, only $r = 0, 2, \infty$ appear, and we mainly focus on $\mathcal{S}_{\mathrm{c}}^\infty (\subset \mathcal{T}^\infty)$. Here, a bijection $\tau : \mathbb{R}^d \to \mathbb{R}^d$ is compactly supported if $\tau = \mathrm{Id}$ outside some compact set.*

Among the above classes of invertible functions, $\mathcal{D}^2$ is our main approximation target, and it is a fairly large class: it contains any $C^2$-diffeomorphism defined on the entire $\mathbb{R}^d$, an open convex set, or more generally a star-shaped open set. The class $\mathcal{T}^\infty$ relates to the distributional universality as

Table 5.1: CF-INN architectures analyzed in this chapter (*Model*: the considered CF-INN architecture. *Flow type*: the flow layer architecture. *This*: the universal approximation property that this chapter has shown. *Prev.*: previously claimed universal approximation property (*Dist.*: distributional universality). ) Our proof techniques are easy to apply to analyze the universality of various CF-INN architectures.

| Model | Flow type | This | Prev. |
|---|---|---|---|
| $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ | Affine coupling [63, 145, 146, 199] | $L^p$ | - |
| $\mathrm{INN}_{\mathrm{DSF}}$ | Deep sigmoidal flow [115] | sup | Dist. [115] |
| $\mathrm{INN}_{\mathrm{SoS}}$ | Sum-of-squares polynomial flow [131] | sup | Dist. [131] |

we will see in Lemma 5.1. The class $\mathcal{S}_{\mathrm{c}}^{\infty}$ is a much simpler class of diffeomorphisms that we use as a stepladder for showing the universality for $\mathcal{D}^2$.

Now we are ready to state the first main theorem. It reveals an equivalence among the universalities for $\mathcal{D}^2$, $\mathcal{T}^{\infty}$, and $\mathcal{S}_{\mathrm{c}}^{\infty}$, under mild regularity conditions. We can use the theorem to lift up the universality for $\mathcal{S}_{\mathrm{c}}^{\infty}$ to that for $\mathcal{D}^2$.

**Theorem 5.1** (Equivalence of Universality)**.** *Let $p \in [1, \infty)$ and let $\mathcal{G}$ be a set of invertible functions.*

*(A) If all elements of $\mathcal{G}$ are piecewise $C^1$-diffeomorphisms, then the $L^p$-universal approximation properties of $\mathrm{INN}_{\mathcal{G}}$ for $\mathcal{D}^2$, $\mathcal{T}^{\infty}$ and $\mathcal{S}_{\mathrm{c}}^{\infty}$ are all equivalent.*

*(B) If all elements of $\mathcal{G}$ are locally bounded, then the* sup*-universal approximation properties of $\mathrm{INN}_{\mathcal{G}}$ for $\mathcal{D}^2$, $\mathcal{T}^{\infty}$ and $\mathcal{S}_{\mathrm{c}}^{\infty}$ are all equivalent.*

The proof is provided in Appendix D.2. For the definitions of the piecewise $C^1$-diffeomorphisms and the locally bounded maps, see Appendix D.5. The regularity conditions in (A) and (B) assure that function composition within $\mathcal{G}$ is compatible with approximations (see Appendix D.6 for details), and they are usually satisfied, e.g., continuous maps are locally bounded.

If one of the two universality properties in Theorem 5.1 is satisfied, the model is also a distributional universal approximator. Let $p \in [1, \infty)$, and we have the following.

**Lemma 5.1.** *An $L^p$-universal approximator for $\mathcal{T}^{\infty}$ is a distributional universal approximator.*

Since sup-universality implies $L^p$-universality, Lemma 5.1 can be combined with both cases of (A) and (B) in Theorem 5.1. The proof is based on the existence of a triangular map connecting two absolutely continuous distributions [28]. See Appendix D.1 for details. Note that the previous studies [131, 115] have discussed the distributional universality of some flow architectures essentially via showing the sup-universality for $\mathcal{T}^{\infty}$. Lemma 5.1 clarifies that the weaker notion of $L^p$-universality is sufficient for the distributional universality, which can also apply to the case (A) in Theorem 5.1.

**Application to previously proposed CF-INN architectures.** Theorem 5.1 can upgrade a previously known sup-universality for $\mathcal{T}^{\infty}$ of a CF-INN architecture to that for $\mathcal{D}^2$. As examples, *deep sigmoidal flows* (DSF; a version of neural autoregressive flows [115]) and *sum-of-squares polynomial flows* (SoS; [131]) can both yield CF-INNs with the sup-universal approximation property for $\mathcal{D}^2$. We provide the proof in Appendix D.7. See Table 5.1 for a summary of the results. See Section 5.5.1 for a comparison with previous theoretical analyses on normalizing flows.

### 5.3.2 Application to Specific Architectures

Our second main theorem reveals the $L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ for $\mathcal{S}_{\mathrm{c}}^{0}$ (hence for $\mathcal{S}_{\mathrm{c}}^{\infty}$), which can be combined with Theorem 5.1 to show its $L^p$-universality for $\mathcal{D}^2$. We define $C_c^{\infty}(\mathbb{R}^{d-1})$ as the set of all compactly-supported $C^{\infty}$ maps from $\mathbb{R}^{d-1}$ to $\mathbb{R}$.

**Theorem 5.2** ($L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$)**.** *Let $p \in [1, \infty)$. Assume $\mathcal{H}$ is a* sup*-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$ and that it consists of piecewise $C^1$-functions. Then,* $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ *is an $L^p$-universal approximator for $\mathcal{S}_c^0$.*

We provide a proof in Appendix D.4. For the definition of piecewise $C^1$-functions, see Appendix D.5. Theorem 5.2 can be combined with Theorem 5.1 to show that $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ is an $L^p$-universal approximator for $\mathcal{D}^2$. Examples of $\mathcal{H}$ satisfying the condition of Theorem 5.2 include multi-layer perceptron models with the *rectifier linear unit* (ReLU) activation [161] and a linear-in-parameter model with smooth universal kernels [183]. The result can be interpreted as a convenient criterion to check the universality of a CF-INN: if the flow architecture $\mathcal{G}$ contains ACFs (or even just $\mathcal{H}$-ACF with sufficiently expressive $\mathcal{H}$) as special cases, then $\mathrm{INN}_{\mathcal{G}}$ is an $L^p$-universal approximator for $\mathcal{D}^2$.

By combining Theorem 5.1, Theorem 5.2, and Lemma 5.1, we can affirmatively answer a previously unsolved problem [198, p.14]: the distributional universality of CF-INN based on ACFs.

**Theorem 5.3** (Distributional universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$)**.** *Under the conditions of Theorem 5.2,* $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ *is a distributional universal approximator.*

**Implications of Theorem 5.2 and Theorem 5.3.** Theorem 5.2 implies that, if $\mathcal{G}$ contains $\mathcal{H}$-ACF as special cases, then $\mathrm{INN}_{\mathcal{G}}$ is an $L^p$-universal approximator for $\mathcal{D}^2$. In light of Theorem 5.3, it is also a distributional universal approximator, hence we can confirm the theoretical plausibility for using it for normalizing flows. Such examples of $\mathcal{G}$ include the *nonlinear squared flow* [316], *Flow++* [108], the *neural autoregressive flow* [115], and the *sum-of-squares polynomial flow* [131]. The result may not immediately apply to the typical *Glow* [145] models for image data that use the 1x1 invertible convolution layers and convolutional neuralnetworks for the coupling layers. However, the Glow architecture for non-image data [7, 267] can be interpreted as $\mathrm{INN}_{\mathcal{G}}$ with ACF layers, hence it is both an $L^p$-universal approximator for $\mathcal{D}^2$ and a distributional universal approximator.

## 5.4 Proof Outline

In this section, we outline the proof ideas of our main theorems to provide an intuition for the constructed approximator and derive reusable insight for future theoretical analyses.

### 5.4.1 Proof Outline for Theorem 5.1

Here, we outline the equivalence proof of Theorem 5.1. For details, see Appendix D.2. Since we have $\mathcal{S}_c^\infty \subset \mathcal{T}^\infty \subset \mathcal{D}^2$, it is sufficient to prove that the universal approximation properties for $S^\infty$ implies that for $\mathcal{D}^2$. Note that the proofs do not change for $L^p$-universality and sup-universality.

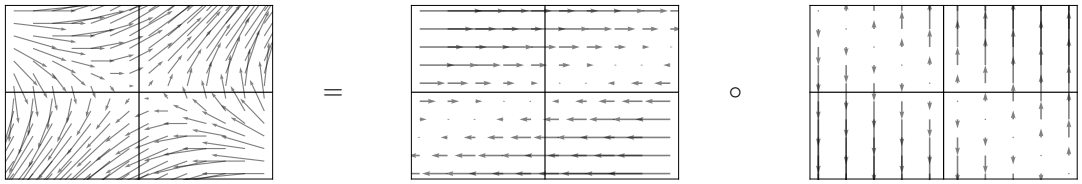
Therefore, we focus on describing the reduction from $\mathcal{D}^2$ to $\mathcal{S}_c^\infty$. Since the approximation of $\mathcal{S}_c^2$ can be reduced to that of $\mathcal{S}_c^\infty$ by a standard mollification argument (see Appendix D.2.2), we show a reduction from $\mathcal{D}^2$ to $\mathcal{S}_c^2$:

**Theorem 5.4.** *For any element $f \in \mathcal{D}^2$ and compact subset $K \subset U_f$, there exist $n \in \mathbb{N}$, $W_1, \dots, W_n \in \mathrm{Aff}$, and $\tau_1, \dots, \tau_n \in \mathcal{S}_c^2$ such that $f(x) = W_1 \circ \tau_1 \circ \cdots \circ W_n \circ \tau_n(x)$ for all $x \in K$.*

Behind the scenes, Theorem 5.4 reduces $\mathcal{D}^2$ to $\mathcal{S}_c^2$ in four steps:

$$\mathcal{D}^2 \rightsquigarrow \mathrm{Diff}_c^2 \rightsquigarrow \text{Flow endpoints} \rightsquigarrow \text{nearly-Id} \rightsquigarrow \mathcal{S}_c^2$$

Here, $A \rightsquigarrow B$ ($A$ *is reduced to* $B$) indicates that the universality for $A$ follows from that for $B$, and Id denotes the identity map. We explain each reduction step in the below.



$$f = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = f_2 \circ f_1 \qquad f_2\begin{pmatrix} x_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} ax_1 + b(\frac{y_2 - cx_1}{d}) \\ y_2 \end{pmatrix} \qquad f_1\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ cx_1 + dx_2 \end{pmatrix}$$

Figure 5.2: A nearly-Id transformation $f$ can be decomposed into coordinate-wise ones ($f_1$ and $f_2$: realized by $\mathcal{S}^2_c$ and permutations). The arrows indicate the transportation of the positions. A general nonlinear $f$ can be analogously decomposed by Proposition 5.3 when $f$ satisfies certain conditions.

*From $\mathcal{D}^2$ to $\mathrm{Diff}^2_c$.* We consider a special subset $\mathrm{Diff}^2_c \subset \mathcal{D}^2$, which is the group of *compactly-supported* $C^2$-diffeomorphisms on $\mathbb{R}^d$ whose group operation is functional composition. Here, a bijection $f : \mathbb{R}^d \to \mathbb{R}^d$ is compactly supported if $f = \mathrm{Id}$ outside some compact set. Proposition 5.1 below reduces the problem of the universality for $\mathcal{D}^2$ to that for $\mathrm{Diff}^2_c$.

**Proposition 5.1.** *For any $f \in \mathcal{D}^2$ and any compact subset $K \subset U_f$, there exist $h \in \mathrm{Diff}^2_c$, $W \in \mathrm{Aff}$, such that for all $\boldsymbol{x} \in K$, $f(\boldsymbol{x}) = W \circ h(\boldsymbol{x})$.*

*From $\mathrm{Diff}^2_c$ to flow endpoints.* In order to construct an approximation for the elements of $\mathcal{D}^2$, we devise its subset that we call the *flow endpoints*. A flow endpoint is an element of $\mathrm{Diff}^2_c$ which can be represented as $\phi(1)$ using an "additive" continuous map $\phi : [0, 1] \to \mathrm{Diff}^2_c$ with $\phi(0) = \mathrm{Id}$. Here, "additivity" means $\phi(s) \circ \phi(t) = \phi(s + t)$ for any $s, t \in [0, 1]$ with $s + t \in [0, 1]$. This additivity will be later used to decompose a flow endpoint into a composition of some mildly-behaved fragments of the flow map. Note that we equip $\mathrm{Diff}^2_c$ with the Whitney topology [95, Proposition 1.7.(9)] to define the continuity of the map $\phi$. The importance of the flow endpoints lies in the following lemma that we prove in Appendix D.3:

**Lemma 5.2.** *Any element in $\mathrm{Diff}^2_c$ can be represented as a finite composition of flow endpoints.*

Lemma 5.2 is essentially due to Fact 5.1, which is the following structure theorem in differential geometry attributed to Herman, Thurston [270], Epstein [71], and Mather [179, 180]:

**Fact 5.1.** *The group $\mathrm{Diff}^2_c$ is simple, i.e., any normal subgroup $H \subset \mathrm{Diff}^2_c$ is either $\{\mathrm{Id}\}$ or $\mathrm{Diff}^2_c$.*

*From flow endpoints to nearly-*Id. The flow endpoints in $\mathrm{Diff}^2_c$ can be decomposed into "nearly-Id" elements in $\mathrm{Diff}^2_c$ by leveraging its additivity property, as in the following proposition. Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm.

**Proposition 5.2.** *For any $f \in \mathrm{Diff}^2_c$, there exist finite elements $g_1, \ldots, g_r \in \mathrm{Diff}^2_c$ such that $f = g_1 \circ \cdots \circ g_r$ and $\sup_{x \in \mathbb{R}^d} \|Dg_i(x) - I\|_{\mathrm{op}} < 1$, where $Dg_i$ is the Jacobian of $g_i$.*

Proposition 5.2 leverages the continuity of the flows with respect to the Whitney topology of $\mathrm{Diff}^2_c$: $\phi(1/n)$ uniformly converges to the identity map both in its values and *its Jacobian* when $n \to \infty$. Thus, any flow endpoint $\phi(1)$ can be represented by an $n$-time composition of $\phi(1/n)$ each of which is close to identity (nearly-Id) when $n$ is sufficiently large.

*From nearly-*Id *to $\mathcal{S}^2_c$.* The nearly-Id elements, $g \in \mathrm{Diff}^2_c$ in Proposition 5.2, can be decomposed into elements of $\mathcal{S}^2_c$ and permutation matrices:

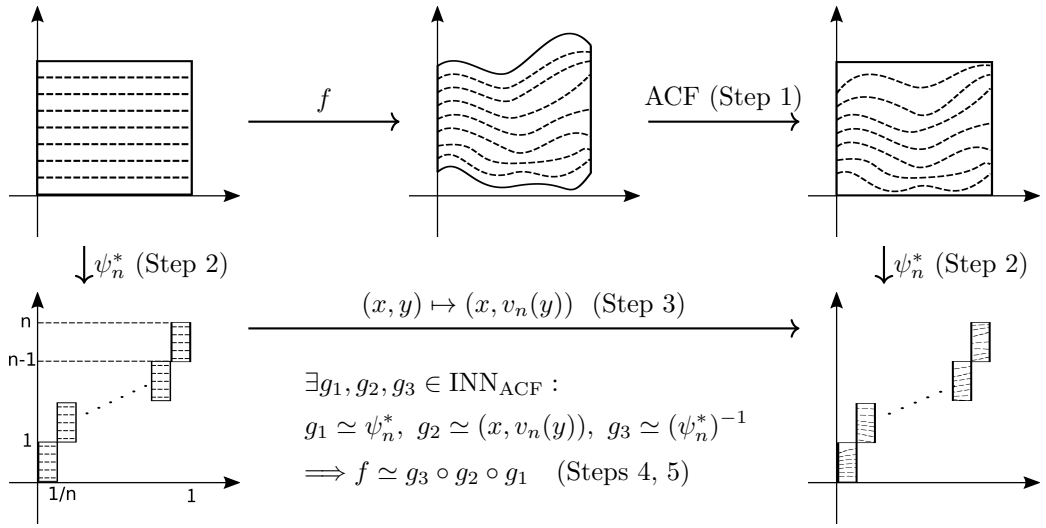



Figure 5.3: Illustration of the proof technique for the $L^p$-universal approximation property of $\mathrm{INN}_{\mathrm{ACF}}$ for $\mathcal{S}_{\mathrm{c}}^0$. The symbol $\simeq$ indicates approximation to arbitrary precision.

**Proposition 5.3.** *For any* $g \in \mathrm{Diff}_{\mathrm{c}}^2$ *with* $\sup_{x\in\mathbb{R}^d} \|Dg(x) - I\|_{\mathrm{op}} < 1$, *there exist* $d$ *elements* $\tau_1, \ldots, \tau_d \in \mathcal{S}_{\mathrm{c}}^2$ *and permutation matrices* $\sigma_1, \ldots, \sigma_d$ *such that*

$$g = \sigma_1 \circ \tau_1 \circ \cdots \circ \sigma_d \circ \tau_d.$$

The machinery of this decomposition is illustrated in Figure 5.2.

### 5.4.2 Proof Outline for Theorem 5.2

Here, we give the proof outline of Theorem 5.2. For details, see Appendix D.4. The main difficulty in constructing the approximator is the restricted functional form of ACFs. However, the problem reduction by Theorem 5.1 allows us to construct an approximator by approximating a step function.

For illustration, we only describe the case for $d = 2$ and $K \subset [0,1]^2$. For complete proof of Theorem 5.2, see Appendix D.4. Let $f(x, y) = (x, u(x, y))$ be the target function, where $u(\cdot, y)$ is a continuous function that is strictly increasing for each $y$ (i.e., $f \in \mathcal{S}_{\mathrm{c}}^0$). For the compact set $K \subset [0,1]^2 \subset \mathbb{R}^2$, we find $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ arbitrarily approximating $f$ on $K$ as follows (Figure 5.3).

Step 1. **Align the image into the square:** First, without loss of generality, we may assume that the image $f([0,1]^2)$ is again $[0,1]^2$. Indeed, we can align the image so that $u(x, 1) = 1$ and $u(x, 0) = 0$ for all $x \in [0,1]$ by using only an ACF $\Psi_{1,s,t}$ with continuous $s$ and $t$, which can be approximated by $\mathcal{H}$-ACF.

Step 2. **Slice the squares and stagger the pieces:** We consider an imaginary ACF $\psi_n^* := \Psi_{1,1,t_n}$ defined using a discontinuous step function $t_n := \sum_{k=0}^{n} k\mathbf{1}_{[k/n,(k+1)/n)}$. The map $\psi_n^*$ splits $[0,1]^2$ into pieces and staggers them so that a coordinate-wise independent transformation (e.g., $v_n$ in Step 3), which is uniform along the $x$-axis, can affect each piece separately.

Step 3. **Express $f$ by a coordinate-wise independent transformation:** We construct a continuous increasing function $v_n : \mathbb{R} \to \mathbb{R}$ such that for $y \in [k, k+1)$, $v_n(y) = u(k/n, y) + k$ $(k = 0, \ldots, n-1)$. A direct computation shows that $\tilde{f}_n := (\psi_n^*)^{-1} \circ (\cdot, v_n(\cdot)) \circ \psi_n^*$ arbitrarily approximates $f$ on $[0,1]^2$ if we increase $n$. We take a sufficiently large $n$.

Step 4. **Approximate the coordinate-wise independent transformation $v_n$:** We find an element of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ sufficiently approximating $(\cdot, v_n(\cdot))$ on $[0,1] \times [0,n]$. This is realized based on a lemma that we can construct an approximator for any element of $\mathcal{S}_{\mathrm{c}}^0$ of the form $(x,y) \mapsto (x, v(y))$ on any compact set in $\mathbb{R}^2$.

Step 5. **Approximate $\psi_n^*$ and combine the approximated constituents to approximate $\tilde{f}_n$:** We can also approximate $\psi_n^*$ and its inverse by ACFs based on the universality of $\mathcal{H}$. Finally, composing the approximated constituents gives an approximation of $f$ on $[0,1]^2$ with arbitrary precision (see Appendix D.6).

## 5.5 Related Work and Discussion

In this section, we relate the contribution of this dissertation to the literature on the representation power of invertible neural networks.

### 5.5.1 Normalizing Flows

The distributional universality of *normalizing flows* constructed using CF-INNs has been addressed in previous studies such as [131, 115]. Previously proposed architectures with distributional universality include the neural autoregressive flows [115] and the sum-of-squares polynomial flows [131]. Our findings elucidate the much stronger universalities of these architectures, namely the sup-universality for $\mathcal{D}^2$, which enhances the reliability of these models in the tasks where function approximation rather than distribution approximation is crucial, e.g., feature extraction [192, 267].

### 5.5.2 Other Invertible Neural Network Architectures

**One-dimensional case.** In the one-dimensional case ($d = 1$), strict monotonicity is a necessary and sufficient condition for a function to be invertible. In this case, there have been a few invertible neural network architectures with sup-universality for the set of all homeomorphisms on $\mathbb{R}$, e.g., *monotonic networks* [247] and *rational quadratic splines* [68]. These models complement CF-INNs in that they provide an invertible neural network only in the one-dimensional case, whereas the latter can be defined only in the multi-dimensional case.

**Relation to examples of functions that cannot be approximated by NODEs.** Neural ordinary differential equations (NODEs) [40, 67] can be considered as another design of invertible flow layers different from CFs. Zhang et al. [308] formulated its Theorem 1 to show that NODEs are not universal approximators by presenting a function that a NODE cannot approximate. The existence of this counterexample does not contradict our result because our approximation target $\mathcal{D}^2$ is different from the function class considered in Zhang et al. [308]: the class in Zhang et al. [308] can contain discontinuous maps whereas the elements of $\mathcal{D}^2$ are smooth and invertible. Also, in Proposition 5.1, we cap an affine transformation (realizable by $\mathrm{INN}_{\mathcal{G}}$) on top of the target function to reduce the approximation of $\mathcal{D}^2$ to that of $\mathrm{Diff}_{\mathrm{c}}^2$. Such an affine transformation may enhance the approximation capacity by allowing a certain set of transformations, e.g., coordinate-wise sign flipping.

### 5.5.3 The Strength of the Representation Power of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$

In this study, we showed the $L^p$-universal approximation property of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. While the $L^p$-universality is likely to suffice for developing probabilistic risk bounds for machine learning tasks

[167, 264] and for showing distributional universality, whether $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ is a sup-universal approximator for $\mathcal{D}^2$ remains an open question. Our conjecture is negative due to the following theoretical observation. The sup-universality requires a precise approximation uniformly everywhere while the $L^p$-universality can allow an approximation error on negligible regions. As described in Section 5.4.2, we used a smooth approximation of step functions to show the $L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. Intuitively, approximating the step functions and composing them can accumulate errors around the discontinuity points, so that it can retain the $L^p$-universality but it can affect the sup-universality. Since the step functions are devised to bypass the uniformity of the transformation by ACFs, we conjecture that the difficulty is intrinsic and a sup-universality is unlikely to hold for $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$.

## 5.6 Conclusion

In this study, we elucidated the representation power of CF-INNs by proving their $L^p$-universality or sup-universality for $\mathcal{D}^2$. Along the course, we invoked a structure theorem from differential geometry to establish an equivalence of the universalities for $\mathcal{D}^2$, $\mathcal{S}_{\mathrm{c}}^{\infty}$, and $\mathcal{T}^{\infty}$, which itself is of theoretical interest. Our result advances the theoretical understanding of CF-INNs by formally showing that most of the CF-INN architectures already yield $L^p$-universal approximators and that the different flow layer designs purely contribute to the efficiency of approximation, not much to the capacity of the model class. Comparing the approximation efficiency of different layer designs is an important area in future work. Also, the approximation efficiency for a better-behaved subset of $\mathcal{D}^2$ (e.g., bi-Lipschitz ones) remains as an open question for future research. In Chapter 6, we discuss further possibilities of future research directions.

# Chapter 6

# Conclusion and Future Prospects

In this chapter, we revisit the central statement of this dissertation and discuss the contributions of the previous chapters to this theme. Moreover, we also discuss important subjects and problems for future research that lie beyond this dissertation.

## 6.1 Conclusion

In this dissertation, we considered whether causal knowledge could be useful for predictive modeling. Intuitively, the benefit of causal models lies in the portability and invariance; causality is interesting for human beings because causal knowledge is believed to be valid even outside of the environment in which the knowledge was acquired. Therefore, learned causal knowledge can be potentially used as prior knowledge to enhance statistical machine learning, especially in small-data environments, because it captures some aspect (i.e., the data-generating process) of the data distribution that would otherwise be inferred from data. The concrete idea of this dissertation was to develop data augmentation methods to incorporate the statistical independence relations embedded in the structural causal models (SCMs; Section 2.3). This approach has the virtue of being model-independent: it can be combined with virtually any standard supervised machine learning method.

We presented a unified design principle for data augmentation to reflect the independence relations; the strategy was to mix the values of the sample points in an element-wise manner to synthesize hypothetical data. In Chapters 3 and 4, we have seen how this idea could be instantiated in two situations where the causal model is either partially known or estimable. The two cases correspond to different levels in the hierarchy of the structural causal framework: SCM is at the most granular level for which partial information could be estimated from data, and the graphical causal models (GCM; Section 2.3) is its coarse version. Chapter 5, while forming an interesting theoretical contribution on its own, supported the methodology employed in Chapter 4 by providing a theoretical guarantee.

Jointly, the three chapters provided evidence of the Central Statement; the knowledge of causal mechanisms estimated by GCMs/SCMs can enhance machine learning in small-data problems. In particular, we have developed the concrete methods to exploit the statistical independence relations implied by the causal models via data augmentation.

## 6.2 Future Prospects

Beyond the scope of this dissertation, there are various possible directions for future studies. In this section, we list some important research subjects that lie ahead.

### 6.2.1 Reducing Computational Complexity

The general methodological idea of this dissertation was based on data augmentation. However, as the approach from data augmentation inevitably increases the number of data points used for learning, the computational complexity of the proposed learning algorithms tends to become higher. Especially, if $n$ is the training sample size and $d$ is the dimensionality of the data, the computational complexities of causal-graph data augmentation (CDA; Algorithm 2) and causal mechanism transfer (CMT; Algorithm 3) can be of the order of $n^d$, which quickly increases with $d$. Therefore, to extend the applicability of the proposed method to higher dimensional data, the following problem should be addressed.

**Problem 6.1.** *Reduce the computational complexity of CDA and CMT to $\mathcal{O}(n^d)$ without losing much of their performance.*

### 6.2.2 Relaxing Model Assumptions

When the causal graph is not identifiable or if it is only partially known from the domain knowledge, we may only acquire an *equivalence class* of acyclic directed mixed graphs (ADMGs; Section 2.2) called *Partial Ancestral Graphs* (PAGs; Peters et al. [208, Table 9.1]). Thus, extending the approach of CDA to PAGs would make the idea more widely applicable.

**Problem 6.2.** *Generalize the approach of CDA to PAGs in place of ADMGs.*

The proposed framework of CMT considered the case of Markovian SCMs (Definition 2.4), i.e., those where all variables are assumed to be observable. When some variables are unobserved, we need to handle semi-Markovian SCMs (Definition 2.4). In order to extend the applicability of CMT, it is desirable to develop its generalization to such cases.

**Problem 6.3.** *Extend the applicability of CMT to semi-Markovian SCMs.*

The proposed framework of CMT cannot handle categorical variables because the identifiability of nonlinear independent component analysis (NLICA; e.g., Proposition 2.9) based on generalized contrastive learning (GCL; [124]) presumes that the data is real-valued. In building practical applications of CMT, categorical variables are as interesting as continuous variables. Thus, it would be important to consider extensions of CMT to categorical variables.

**Problem 6.4.** *Extend the approach of CMT to the case where there are also categorical variables in the data.*

### 6.2.3 Evaluating and Mitigating Model Misspecification

The use of prior knowledge in machine learning is a two-sided sword; it can facilitate learning by narrowing our attention down to specific subsets of the problem, but it may introduce a misspecification error. In the case of the causality-informed machine learning framework presented in this dissertation, the source of model misspecification lies in the following three layers of assumptions: the existence of an SCM, estimability of the causal models, and sufficient representation power of the implementation (Table 6.1). Among the three, the issue of sufficiency of the representation power has been addressed in this dissertation (Chapter 5).

**Existence of an SCM.** To further enhance the reliability of the proposed approaches, it is desirable to develop theories and algorithms for evaluating and mitigating the model misspecification error. Specifically, it would be interesting to see whether the proposed algorithms are justifiable even without the existence of the SCMs behind the data distributions. If the proposed methods,

Table 6.1: Three levels of model misspecification in causality-informed machine learning.

| Type of assumption | CDA | CMT |
|---|---|---|
| Existence of causal mechanism | Problem 6.5 | Problem 6.5 |
| Estimability of causal mechanism | Problem 6.2 | Problem 6.4, 6.7 |
| Representation Power of Implementation | Satisfied | Chapter 5 |

CDA (Chapter 3) and CMT (Chapter 4), are valid even without the existence of such SCMs behind the data, it may be sensible to apply the scheme to non-conventional data types such as images for which structural equations have not been considered.

**Problem 6.5.** *What is expected, in general, if we apply causality-informed machine learning methods when SCMs do not exist behind the data? Are CDA and CMT justifiable even in these cases?*

In this dissertation, we introduced the definition of the solutions of an SCM that requires the random variables to *almost surely* satisfy the structural equations (SEs). It may be fruitful to consider an alternative requirement that the SEs should be satisfied *with high probability*, which can allow the model to accommodate even more general situations.

**Problem 6.6.** *Is it possible to develop a relaxed version of the theory of SCMs where the solutions are required to satisfy the SEs only with high probability instead of almost surely? What are the implications of such theories to the methodologies developed in this dissertation?*

**Estimability of causal models.** In the case of CDA, the answer to Problem 6.2 could be a solution to the case that the causal graph is not fully estimable or not fully known. In the case of CMT, the invariance of the SEs played a crucial role in the estimability. In more practical applications, the invariance assumption could be violated, e.g., the assumption that all domains have the same intervention state may be violated when different hospitals employ different intervention policies. Thus, in the case that the data may have been generated from similar yet non-identical SEs, containing the error or adapting to the violation of the invariance assumption would be important to build practical applications of the proposed approach. Specifically, for example, Greene [93, p.284 Example 9.5] discussed the case of *group-wise heteroscedasticity* and employed different offset parameters for different countries. Considering whether a similar approach can be devised in CMT to accommodate the difference in the SEs, but with a more flexible model of the dissimilarities among the domains than just offsets, would be interesting.

**Problem 6.7.** *In the problem setup of CMT, relax the assumption that the SEs are identical across different domains. Design a method to adapt to the violation of this assumption when the change in the SE is (partially) known, e.g., what parts of the SEs are invariant and what are prone to change. For example, can we generalize the assumption that the different domains have the same SEs except for a difference in the offsets and take advantage of it in CMT? In general, what type of similarity of the SEs can be exploited in this approach, and under what conditions? Can we automatically learn the similarity and differences in the SEs using the data from different domains and exploit the similarity in the mechanism by extending the idea of CMT?*

### 6.2.4 Exploiting Other Aspects of SCMs

In this dissertation, we considered exploiting the causal models for machine learning via the statistical (conditional) independence relations that they imply. In general, causal models contain more statistical and causal information than only conditional independence (Peters et al. [208, Section 9.5]).

**Problem 6.8.** *In general, SCMs imply more statistical and causal information than only conditional independence. How can we exploit the knowledge in predictive modeling?*

In this case, it is also favorable to develop a generic strategy for designing data augmentation procedures that can reflect various statistical properties beyond statistical independence. Notably, statistical independence is expressed as a certain equality constraint of the density functions. Thus, the proposed approach of this dissertation could be more generalized if incorporating such general statistical assertions expressed as sets of equations among *statistical functionals*, i.e., functionals of the distributions.

**Problem 6.9.** *Develop a data augmentation strategy for more general statistical properties than (conditional) independence. Is it possible to design a meta-algorithm that produces a plausible data augmentation strategy given a set of equations among some statistical functionals as inputs?*

In the case that the SEs are parametrically specified, e.g., in the case of linear SEMs, the specific functions of the SEs may contain much richer information than the statistical independence.

**Problem 6.10.** *Exploit the specific functions of an estimated parametric SEM (e.g., linear SEMs) to further improve the sample efficiency of supervised learning. Is it possible to theoretically characterize the optimal predictor in such a case?*

### 6.2.5 Characterizing the Limitation of Causality-informed Learning

Throughout this dissertation, we investigated how causal assumptions on the data-generating processes could be exploited to train a predictor. With such additional assumptions, the problems are more specific and hence easier than the corresponding problems *without* the additional assumptions. While this is qualitatively obvious, to understand the relative difficulty more completely, it is important to quantitatively characterize the difficulty of the problems in terms of *minimax risk lower-bounds* [275]. Therefore, the following problem is important in future work.

**Problem 6.11.** *Provide a minimax lower bound on the generalization error for Problems 3.1 and 4.1.*

### 6.2.6 Application to Other Statistical Inference Tasks

The proposed data augmentation procedures are independent of the supervised learning setting, and they could be useful in other statistical tasks such as hypothesis testing or uncertainty estimation. However, if we apply the methods developed for independent random samples, we may suffer from the potential bias introduced by the data augmentation procedures. Therefore, it is important to investigate how the successive inference steps should be modified to avoid or mitigate such unintended effects when the goal is different from learning good predictors.

**Problem 6.12.** *Devise correction methods and theories for important statistical tasks such as hypothesis testing and uncertainty estimation to enable or justify the use of the data augmented by CDA and CMT in these tasks.*

Application of the proposed methods to unsupervised learning [236], such as representation learning [92], missing value imputation [277], and anomaly detection [4], can also be interesting. Unlike supervised learning investigated in this dissertation, unsupervised learning often uses different evaluation criteria. Therefore, an interesting question is how these tasks can benefit from CDA and CMT.

**Problem 6.13.** *Elucidate the utility of causal knowledge for unsupervised learning. From theoretical and practical perspectives, evaluate how CDA and CMT can be used to enhance the methods for unsupervised learning tasks such as representation learning, missing value imputation, and anomaly detection.*

### 6.2.7 More Detailed Levels of the Data-Generating Processes

The present dissertation proposed a framework for exploiting SCMs and GCMs in predictive modeling. In the hierarchical structure of the SCM framework (Peters et al. [208, Table 1.1]), SCMs and GCMs are the current deepest levels where some information of the model (namely RSEs and CGs) is estimable. On the other hand, more detailed models of the data-generating process, such as ordinary-differential-equation-based foundations of the SCMs (Mooij et al. [186]) have also been explored. Considering the possibility of estimating and exploiting such models is an important direction in the future development of causality-informed machine learning since the more detailed models of the data-generating processes may provide richer prior knowledge that can be used for training a predictor.

**Problem 6.14.** *Develop methods to estimate and exploit more detailed models of the data-generating process than SCMs and GCMs.*

### 6.2.8 Exploiting the Potential Outcome Framework

In many instances of econometrics [6, 1], political science [126], medicine [103], and epidemiology [103], the potential outcome framework [226, 126, 103] is often applied. It would also be interesting to develop methods that can directly exploit the causal knowledge expressed and estimated in the potential outcome framework (e.g., *average treatment effect*) in predictive modeling.

**Problem 6.15.** *Devise a method to exploit the causal knowledge captured by the potential outcome framework in predictive modeling.*

### 6.2.9 Extension to Continual Learning

*Continual learning* refers to the problem of learning from an infinite stream of data, with the goal of gradually extending acquired knowledge and using it for future learning [42, 61]. There are two main challenges in continual learning: (i) using previously acquired knowledge for performing well in newer tasks and (ii) avoiding the performance degradation on a previously learned task or domain after learning new tasks, a destructive phenomenon known as *catastrophic forgetting* [42]. One salient characteristic of the causal knowledge is its stability and portability; causal knowledge is believed to be valid in systems other than the one in which it was acquired (although the extent to which the stability is believed to hold may depend on the domain of interest; Woodward [296] and Woodward [295]). Thus, conceptually, learning causal knowledge may potentially provide a plausible approach to the first challenge: learning relevant knowledge from previously seen tasks and applying the knowledge to newer tasks.

**Problem 6.16.** *Develop continual learning methods that capture and leverage the causal knowledge of the data-generating processes.*

# Appendix A

# Appendices for Chapter 2

## A.1 Example of Figure 2.1

Here, we remark how the example of Figure 2.1 is obtained. Let $\mathbf{0}$ denote the zero-vector and $\mathbf{I}$ the unit matrix (of appropriate dimensions).

**First case: Figure 2.1(a).** This is the base case to which the distributions of the other two cases were matched by carefully choosing the parameters. Consider the following equation and its solution for $(x, y)$:

$$\begin{cases} x = e_1 \\ y = x + ae_2 \end{cases} \quad \Rightarrow \quad \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & a \end{pmatrix}}_{A} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

Then, if $(e_1, e_2)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have $(x, y)^\top \sim \mathcal{N}\left(\mathbf{0}, AA^\top\right)$ where

$$AA^\top = \begin{pmatrix} 1 & 1 \\ 1 & 1 + a^2 \end{pmatrix}.$$

**Second case: Figure 2.1(b).** Consider the following equation and its solution for $(x, y)$:

$$\begin{cases} x = by + ce_1 \\ y = de_2 \end{cases} \quad \Rightarrow \quad \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{pmatrix} c & bd \\ 0 & d \end{pmatrix}}_{B} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

Then, if $(e_1, e_2)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have $(x, y)^\top \sim \mathcal{N}\left(\mathbf{0}, BB^\top\right)$ where

$$BB^\top = \begin{pmatrix} c^2 + b^2 d^2 & bd^2 \\ bd^2 & d^2 \end{pmatrix}.$$

Thus, in order to match the distribution of $(X, Y)^\top$ to the first case,

$$b = (1 + a^2)^{-1}, \; c = a/(1 + a^2)^{1/2}, \; d = (1 + a^2)^{1/2}$$

is sufficient.

**Third case: Figure 2.1(c).** Consider the following equation and its solution for $(x, y, z)$:

$$\begin{cases} x = \alpha z + \gamma e_1 \\ y = \beta z + \delta e_2 \\ z = e_3 \end{cases} \quad \Rightarrow \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \underbrace{\begin{pmatrix} \gamma & 0 & \alpha \\ 0 & \delta & \beta \\ 0 & 0 & 1 \end{pmatrix}}_{C} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

Then, if $(e_1, e_2, e_3)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have $(x, y, z)^\top \sim \mathcal{N}\left(\mathbf{0}, CC^\top\right)$ where

$$CC^\top = \begin{pmatrix} \alpha^2 + \gamma^2 & \alpha\beta & \alpha \\ \alpha\beta & \beta^2 + \delta^2 & \beta \\ \alpha & \beta & 1 \end{pmatrix}.$$

Thus, in order to match the distribution of $(X, Y)^\top$ to the first case,

$$\alpha = \beta^{-1},\ 1 < \beta^2 < 1 + a^2,\ \gamma = (1 - 1/\beta^2)^{1/2},\ \delta = (1 + a^2 - \beta^2)^{1/2}$$

is sufficient.

With the choice of $a = .3$ and $\beta = 1.03$, Figure 2.1 was generated.

## A.2 Supplementary on Causal Models and Proofs

In this section, we provide the supplementary results, the facts, and the proofs relevant to the main text.

### A.2.1 Preparation from Probability Theory

**Definition A.1** (Markov kernels)**.** *Let $(D, \mathcal{D})$ and $(E, \mathcal{E})$ be measurable spaces. A* Markov kernel[1] *from $(D, \mathcal{D})$ to $(E, \mathcal{E})$ is a positive function $K : D \times \mathcal{E} \to \mathbb{R}$ such that*

- *$K(x, \cdot)$ is a probability measure for all $x \in D$, and*
- *$K(\cdot, B)$ is a measurable function for all $B \in \mathcal{E}$.*

*For simplicity, when $D = \emptyset$, we refer to probability measures on $(E, \mathcal{E})$ as Markov kernels.*

For a shorthand notation, we write $\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y)$ to denote the equation

$$\pi(A \times B) = \int_A \mu(\mathrm{d}x)K(x, B), \quad A \in \mathcal{D}, B \in \mathcal{E},$$

where $\mu$ is a measure on $(D, \mathcal{D})$, $K$ is a Markov kernel from $(D, \mathcal{D})to(E, \mathcal{E})$, and $\pi$ is a measure on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$. We use the following facts regarding Markov kernels.

**Fact A.1** (Integration Theorem [47, Theorem I.6.11])**.** *Let $\mu$ be a measure on $(D, \mathcal{D})$, and $K$ be a Markov kernel from $(D, \mathcal{D})$ to $(E, \mathcal{E})$. Then,*

$$\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y)$$

*defines a unique measure $\pi$ on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$.*

[1] Markov kernels are also known as *transition probability kernels* (Çinlar [47]).

**Fact A.2** (Disintegration Theorem [47, Theorem IV.2.18])**.** *Let $\pi$ be a probability measure on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$. Suppose that $(E, \mathcal{E})$ is standard. Then, there exist a probability measure $\mu$ on $(D, \mathcal{D})$ and a Markov kernel $K$ from $(D, \mathcal{D})$ to $(E, \mathcal{E})$ such that*

$$\pi(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(x, \mathrm{d}y).$$

The following is an easy lemma used in the main text.

**Lemma A.1** (Almost-surely deterministic relation)**.** *Let $(\Omega, \mathcal{U}, \mathbb{P})$ be a probability space, and $(D, \mathcal{D})$ and $(E, \mathcal{E})$ be measurable spaces. Let $X : \Omega \to D$ and $Y : \Omega \to E$ be random variables whose joint distribution is $\mathbb{P}_{X,Y}$, and let the marginal distribution of $X$ be $\mathbb{P}_X$. Assume that $(X, Y)$ almost surely satisfies $Y = f(X)$ for a measurable map $f : D \to E$. Then, for all $A \in \mathcal{D}$ and $B \in \mathcal{E}$,*

$$\mathbb{P}_{X,Y}(\mathrm{d}x, \mathrm{d}y) = \mathbb{P}_X(\mathrm{d}x)\delta_{f(x)}(\mathrm{d}y),$$
$$\mathbb{P}_Y = f_\sharp(\mathbb{P}_X) = \mathbb{P}_X \circ f^{-1}.$$

*Proof.* We have $\mathbb{P}_{X,Y}(A \times B) = \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A, Y \in B, Y = f(X)) = \mathbb{P}(X \in A, f(X) \in B, Y = f(X)) = \mathbb{P}(X \in A, f(X) \in B)$, and $\int_A \delta_{f(x)}(B)\mathbb{P}_X(\mathrm{d}x) = \int \mathbb{1}[x \in A]\,\mathbb{1}[f(x) \in B]\,\mathbb{P}_X(\mathrm{d}x) = \mathbb{P}_X(\{x : x \in A, f(x) \in B\}) = \mathbb{P}(X \in A, f(X) \in B)$. Thus, the first equality holds. On the other hand, for any $B \in \mathcal{E}$, we have $\mathbb{P}_Y(B) = \mathbb{P}(Y^{-1}(B)) = \mathbb{P}(Y^{-1}(B) \cap \{Y = f(X)\}) = \mathbb{P}((f(X))^{-1}(B) \cap \{Y = f(X)\}) = \mathbb{P}(X^{-1}(f^{-1}(B))) = \mathbb{P}_X(f^{-1}(B))$. □

For the rest of this section, we consider the following setup. Let $(D, \mathcal{D})$, $(E, \mathcal{E})$, $(F, \mathcal{F})$ be measurable spaces. Let $X, Y, Z$ be random variables taking values in the measurable spaces $D, E, F$, respectively, defined on the same probability space. Let $\mathbb{P}_X$, $\mathbb{P}_{X,Z}$, $\mathbb{P}_{X,Y}$, and $\mathbb{P}_{X,Y,Z}$ be the distributions of $X, (X, Z), (X, Y)$, and $(X, Y, Z)$, respectively.

**Fact A.3** (Conditional Independence and Markov Kernels [224, Definition 1.4.1, Theorem 3.5.3])**.** *Let $K$ be a Markov kernel from $(D, \mathcal{D})$ to $(E, \mathcal{E})$ satisfying $\mathbb{P}_{X,Y}(\mathrm{d}x, \mathrm{d}y) = K(x, \mathrm{d}y)\mathbb{P}_X(\mathrm{d}x)$. Then,*

$$Y \perp\!\!\!\perp Z \mid X \quad \Rightarrow \quad \mathbb{P}_{(X,Y,Z)}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = K(x, \mathrm{d}y)\mathbb{P}_{X,Z}(\mathrm{d}x, \mathrm{d}z).$$

**Fact A.4** (Conditional Independence and Markov Kernels [224, Definition 1.4.1, Theorem 3.5.5])**.** *Let $K$ be a Markov kernel from $(D, \mathcal{D})$ to $(E, \mathcal{F})$. Then,*

$$\mathbb{P}_{X,Y,Z}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = K(x, \mathrm{d}y)\mathbb{P}_{X,Z}(\mathrm{d}x, \mathrm{d}z) \quad \Rightarrow \quad Y \perp\!\!\!\perp Z \mid X.$$

### A.2.2 GCMs

With the aim of making the definition of *d-separation* more intuitively accessible, we adopt the terminology of *self-activeness* and *activation*, which is not part of the standard terminology ([204, Definition 1.2.3, Section 11.1.2]).

**Definition A.2** (d-separation [204, Definition 1.2.3, Section 11.1.2], [29])**.** *Let $\mathcal{G} = \langle \mathcal{I}, \mathfrak{D}, \mathfrak{B} \rangle$ be a directed mixed graph. Let $\{a\}, \{b\}, S \subset \mathcal{I}$ be distinct subsets.*

- *$a$ and $b$ are* d-separated given *$S$ if and only if no path between $a$ and $b$ is* active given *$S$.*[2]

[2] "d" connotes "directional" in "d-separated".

- *A path between $a$ and $b$ is* active given $S$ *if and only if either (i) the path is* self-active *and $S$ does not* deactivate *it, or (ii) the path is* self-inactive *and $S$* activates *it.*
- *A path between $a$ and $b$ is* self-active *if and only if it contains no* colliders.[3] *Otherwise, it is said to be* self-inactive.
- *A* self-active *path is* deactivated *by $S$ if and only if the path traverses $S$.*
- *A* self-inactive *path is* activated *by $S$ if and only if each collider is either contained in $S$ or has a descendant in $S$, and no non-collider is contained in $S$.*

*For disjoint subsets $A, B, S \subset \mathcal{I}$, we say $A$ and $B$ are* d-separated given $S$ *if and only if all pairs $(a, b) \in A \times B$ are d-separated given $S$, and we write*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid C.$$

**Definition A.3** (Directed Global Markov Property [29, Definition A.6])**.** *Let $\mathcal{G} = \langle \mathcal{I}, \mathfrak{D}, \mathfrak{B} \rangle$ be a directed mixed graph and $\mathbb{P}_{\boldsymbol{Z}}$ be a probability distribution on $\boldsymbol{Z} = \prod_{v \in \mathcal{I}} \mathcal{Z}^v$, where each $\mathcal{Z}^v$ is a standard measurable space. We say that $\mathbb{P}_{\boldsymbol{Z}}$ satisfies the* directed global Markov property *relative to $\mathcal{G}$ if for all subsets $A, B, C \subset \mathcal{I}$ we have*

$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid C \quad \Rightarrow \quad \boldsymbol{Z}^A \perp\!\!\!\perp \boldsymbol{Z}^B \mid \boldsymbol{Z}^C \text{ in } \mathbb{P}_{\boldsymbol{Z}}.$$

**Definition A.4** (Directed Local Markov Property [160, p.50])**.** *Let $\mathcal{G} = \langle \mathcal{I}, \mathfrak{D}, \mathfrak{B} \rangle$ be a directed mixed graph and $\mathbb{P}_{\boldsymbol{Z}}$ be a probability distribution on $\boldsymbol{Z} = \prod_{v \in \mathcal{I}} \mathcal{Z}^v$, where each $\mathcal{Z}^v$ is a standard measurable space. We say that $\mathbb{P}_{\boldsymbol{Z}}$ satisfies the* directed local Markov property *relative to $\mathcal{G}$ if for any $v \in \mathcal{I}$ we have*

$$\boldsymbol{Z}^v \perp\!\!\!\perp \boldsymbol{Z}^{\text{non-desc}(v)} \mid \boldsymbol{Z}^{\text{pa}(v)} \text{ in } \mathbb{P}_{\boldsymbol{Z}},$$

We use the following equivalence to translate the results in Bongers et al. [29] to our context.

**Proposition A.1** (Equivalence of Markov properties [160, Theorem 3.27], [224, Theorem 6.4.4])**.** *Let $\mathcal{G} = \langle V, E \rangle$ be a DAG, and $P$ be a probability distribution over $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$ where each $\mathcal{X}_v$ is a standard measurable space. Then, the following are equivalent:*

*(DF) $P$ admits a recursive factorization according to $\mathcal{G}$,*

*(DG) $P$ satisfies the directed global Markov property relative to $\mathcal{G}$,*

*(DL) $P$ satisfies the directed local Markov property relative to $\mathcal{G}$.*

*Proof.* By Lauritzen et al. [158, Proposition 4], it is known that (DG) is equivalent to (DL). Thus, it suffices to show that (DL) and (DF) are equivalent. We use mathematical induction on the number of vertices $|V|$ of $\mathcal{G}$. If $|V| = 1$, both (DL) and (DF) are trivially true, and hence they are equivalent. Assume that (DL) and (DF) are equivalent when $|V| = n$, and let us now consider the case $|V| = n + 1$. Let $v$ be a terminal vertex of $\mathcal{G}$. For simplicity of notation, let us denote $\boldsymbol{w} := V \setminus v$.

We first show (DL) $\Rightarrow$ (DF) for $n + 1$. (DL) implies $X_v \perp\!\!\!\perp X_{\boldsymbol{w}} \mid X_{\text{pa}(v)}$, and hence Fact A.3 implies

$$P(\mathrm{d}x_{\boldsymbol{w}}, \mathrm{d}x_v) = K'(x_{\text{pa}(v)}, \mathrm{d}x_v) P_{\boldsymbol{w}}(\mathrm{d}x_{\boldsymbol{w}}),$$

where $K'$ is a Markov kernel from $\mathcal{X}_{\text{pa}(v)}$ to $\mathcal{X}_v$ that satisfies $P_{v \cup \text{pa}(v)}(\mathrm{d}x_{\text{pa}(v)}, \mathrm{d}x_v) = K'(x_{\text{pa}(v)}, \mathrm{d}x_v) P_{\text{pa}(v)}(\mathrm{d}x_{\text{pa}(v)})$, which is guaranteed to exist by the disintegration theorem (Fact A.2)

[3] A node $v_i$ in a path $\langle \ldots, e_{i-1}, v_i, e_i, \ldots \rangle$ is called a *collider* if $e_{i-1}$ and $e_i$ have arrowheads at $v_i$, i.e., $(e_{i-1}, e_i) \in \{(\rightarrow, \leftarrow), (\rightarrow, \leftrightarrow), (\leftrightarrow, \leftarrow), (\leftrightarrow, \leftrightarrow)\}$.

and the fact that $\mathcal{X}_v$ is a standard measurable space, and $P_{\boldsymbol{w}}$, $P_{v\cup\mathrm{pa}(v)}$, $P_{\mathrm{pa}(v)}$ are the marginal distributions. Now, since (DL) holds for the pair $(P,\mathcal{G})$ and $v$ is a terminal node, $P_{\boldsymbol{w}}$ also obeys (DL) with respect to the subgraph $\mathcal{G}_{\boldsymbol{w}}$. Thus, by $|\boldsymbol{w}|=n$ and the inductive assumption, $P_{\boldsymbol{w}}$ satisfies (DF) with respect to $\mathcal{G}_{\boldsymbol{w}}$. Therefore, we have (DL) $\Rightarrow$ (DF) for $n+1$.

Next, we show (DF) $\Rightarrow$ (DL) for $n+1$. If (DF) holds for the pair $(P,\mathcal{G})$, then $P_{\boldsymbol{w}}$ also satisfies (DF) with respect to the subgraph $\mathcal{G}_{\boldsymbol{w}}$. Thus, by $|\boldsymbol{w}|=n$ and the inductive assumption, $P_{\boldsymbol{w}}$ satisfies (DL) with respect to $\mathcal{G}_{\boldsymbol{w}}$. Now, (DF) also implies

$$P(\mathrm{d}x_{\boldsymbol{w}},\mathrm{d}x_v)=K^v(x_{\mathrm{pa}(v)},\mathrm{d}x_v)\left(\prod_{w\in\boldsymbol{w}}K^w(x_{\mathrm{pa}(w)},\mathrm{d}x_w)\right) \tag{A.1}$$

for some Markov kernel $K^v$ from $\mathcal{X}_{\mathrm{pa}(v)}$ to $\mathcal{X}_v$. By marginalizing out $\mathrm{d}x_v$ from both sides (using the fact that $v$ is a terminal node), we obtain

$$P_{\boldsymbol{w}}(\mathrm{d}x_{\boldsymbol{w}})=\prod_{w\in\boldsymbol{w}}K^w(x_{\mathrm{pa}(w)},\mathrm{d}x_w).$$

Thus, Fact A.4 applied to Equation (A.1) implies $X_v \perp\!\!\!\perp X_{\boldsymbol{w}} \mid X_{\mathrm{pa}(v)}$. Therefore, we have (DF) $\Rightarrow$ (DL) for $n+1$. □

*Remark* A.1. Proposition A.1 is a straightforward generalization of Theorem 3.27 in Lauritzen [160] that showed the same assertion under the assumption that $P$ has a density with respect to some product measure, to accommodate the measure-theoretical definition of probabilistic graphical models as in Rønn-Nielsen and Hansen [224] and Wu et al. [298].

### A.2.3 SCMs

To prove Proposition 2.2, we use the fact that an equivalent structurally minimal SCM exists. The notions of equivalence and structural minimality are defined as follows.

**Definition A.5** (Structural minimality; Bongers et al. [29]). *An SCM $\mathcal{M}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ is* structurally minimal *if and only if, for all $v\in\mathcal{I}$, there exists a measurable map $\tilde{f}^{(v)}:\boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)}\times\boldsymbol{\mathcal{E}}^{\mathrm{pa}(v)}\to\mathcal{Z}^v$ such that $\boldsymbol{f}^v(\boldsymbol{z},\boldsymbol{e})=\tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)},\boldsymbol{e}^{\mathrm{pa}(v)})$ for all $\boldsymbol{z}\in\boldsymbol{\mathcal{Z}},\boldsymbol{e}\in\boldsymbol{\mathcal{E}}$, where $\mathrm{pa}(\cdot)$ is defined by* $\mathrm{graph}(\mathcal{M})$.

**Definition A.6** (Equivalent SCMs; Bongers et al. [29]). *Two SCMs $\mathcal{M}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ and $\tilde{\mathcal{M}}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\tilde{\boldsymbol{f}},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ are* equivalent *if and only if, for each $v\in\mathcal{I}$,*

$$\boldsymbol{z}^v=\boldsymbol{f}^v(\boldsymbol{z},\boldsymbol{e})\Leftrightarrow\boldsymbol{z}^v=\tilde{\boldsymbol{f}}^v(\boldsymbol{z},\boldsymbol{e}),\quad\forall\boldsymbol{z}\in\boldsymbol{\mathcal{Z}},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\text{-}a.s.(\boldsymbol{e})$$

*holds,*[4] *and we write $\mathcal{M}\simeq\tilde{\mathcal{M}}$.*

We use the following Fact A.5 to prove Proposition 2.2.

**Fact A.5** (Structurally minimal representation [29, Proposition 2.11]). *For an SCM $\mathcal{M}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$, there exists an equivalent SCM $\tilde{\mathcal{M}}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\tilde{\boldsymbol{f}},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ that is structurally minimal.*

**Lemma A.2** (Minimal SEs). *Let $\mathcal{M}=\langle\mathcal{I},\mathcal{J},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ be an SCM. Then, there exist measurable maps $\tilde{f}^{(v)}:\boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)}\times\boldsymbol{\mathcal{E}}^{\mathrm{pa}(v)}\to\mathcal{Z}^v$ $(v\in\mathcal{I})$ such that, for each $v\in\mathcal{I}$,*

$$\boldsymbol{z}^v=\boldsymbol{f}^v(\boldsymbol{z},\boldsymbol{e})\ \Leftrightarrow\ \boldsymbol{z}^v=\tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)},\boldsymbol{e}^{\mathrm{pa}(v)}),\quad\forall\boldsymbol{z}\in\boldsymbol{\mathcal{Z}},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\text{-}a.s.(\boldsymbol{e})$$

[4] Note that requiring the equivalence $\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{z},\boldsymbol{e})\Leftrightarrow\boldsymbol{z}=\tilde{\boldsymbol{f}}(\boldsymbol{z},\boldsymbol{e})$ for each $v\in\mathcal{I}$ is stronger than requiring the whole set of equations to be equivalent, i.e., $\boldsymbol{z}=\boldsymbol{f}(\boldsymbol{z},\boldsymbol{e})\Leftrightarrow\boldsymbol{z}=\tilde{\boldsymbol{f}}(\boldsymbol{z},\boldsymbol{e})$.

*holds where* $\mathrm{pa}(\cdot)$ *is defined by* $\mathrm{graph}(\mathcal{M})$. *In particular, any random variables* $(\boldsymbol{Z}, \boldsymbol{E})$ *satisfying* $\boldsymbol{Z}^v = \boldsymbol{f}^v(\boldsymbol{Z}, \boldsymbol{E})$ *(a.s.) satisfies* $\boldsymbol{Z}^v = \tilde{f}^{(v)}(\boldsymbol{Z}^{\mathrm{pa}(v)}, \boldsymbol{E}^{\mathrm{pa}(v)})$ *(a.s.).*

*Proof.* Take the structurally minimal SCM $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ that is equivalent to $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ (Fact A.5). Then, for each $v \in \mathcal{I}$, there exists a measurable map $\tilde{f}^{(v)} : \boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)} \times \boldsymbol{\mathcal{E}}^{\mathrm{pa}(v)} \to \mathcal{Z}^v$ such that $\tilde{\boldsymbol{f}}^v(\boldsymbol{z}, \boldsymbol{e}) = \tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})$ for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ (Definition A.5). With such $\{\tilde{f}^{(v)}\}_{v \in \mathcal{I}}$, for each $v \in \mathcal{I}$, for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$,

$$\boldsymbol{z}^v = \boldsymbol{f}^v(\boldsymbol{z}, \boldsymbol{e}) \;\Leftrightarrow\; \boldsymbol{z}^v = \tilde{\boldsymbol{f}}^v(\boldsymbol{z}, \boldsymbol{e}) \;\Leftrightarrow\; \boldsymbol{z}^v = \tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)}). \tag{A.2}$$

Let $\Omega_\neg$ be a $\mathbb{P}$-negligible set such that $\boldsymbol{Z}^v(\omega) = \boldsymbol{f}^v(\boldsymbol{Z}(\omega), \boldsymbol{E}(\omega))$ $(\omega \in \Omega \setminus \Omega_\neg)$, and $\boldsymbol{\mathcal{E}}_\neg$ be a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-negligible set such that Equation (A.2) holds for any $\boldsymbol{e} \in \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{E}}_\neg$ and any $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$. Then $\Omega_\neg \cup \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{E}}_\neg)$ is a $\mathbb{P}$-negligible set, and for any $\omega \in \Omega \setminus (\Omega_\neg \cup \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{E}}_\neg))$, we have $\boldsymbol{Z}^v(\omega) = \tilde{f}^{(v)}(\boldsymbol{Z}^{\mathrm{pa}(v)}(\omega), \boldsymbol{E}^{\mathrm{pa}(v)}(\omega))$ $(v \in \mathcal{I})$. That is, $(\boldsymbol{Z}, \boldsymbol{E})$ almost surely satisfies $\boldsymbol{Z}^v = \tilde{f}^{(v)}(\boldsymbol{Z}^{\mathrm{pa}(v)}, \boldsymbol{E}^{\mathrm{pa}(v)})$ $(v \in \mathcal{I})$. □

*Proof of Proposition 2.2.* As a result of Lemma A.2, there exist measurable maps $\{\tilde{f}^{(v)}\}_{v \in \mathcal{I}}$ such that for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$,

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}) \;\Leftrightarrow\; \boldsymbol{z}^v = \tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)}) \quad (v \in \mathcal{I}).$$

Since $\mathrm{graph}(\mathcal{M})$ is a DAG and the elements of $\mathcal{J}$ have no parents, we can find a topological ordering $\prec$ such that all elements of $\mathcal{J}$ precede the elements of $\mathcal{I}$, i.e., $u_1 \prec \cdots \prec u_{|\mathcal{J}|} \prec v_1 \prec \cdots v_{|\mathcal{I}|}$, where $\{u_l\}_{l=1}^{|\mathcal{J}|}$ and $\{v_l\}_{l=1}^{|\mathcal{I}|}$ are distinct elements of $\mathcal{J}$ and $\mathcal{I}$, respectively. Now, we solve the equations

$$\boldsymbol{z}^v = \tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)}) \quad (v \in \mathcal{I}) \tag{A.3}$$

for $\boldsymbol{z}$ by one-by-one eliminating $v_1, \ldots, v_{|\mathcal{I}|}$ from the right-hand side. Since $\mathrm{pa}(v_1)$ contains no elements of $\mathcal{I}$, we can eliminate $\boldsymbol{z}^{v_1}$ from the right-hand side of Equation (A.3) by imputing $\tilde{f}^{(v_1)}(\boldsymbol{z}^{\mathrm{pa}(v_1)}, \boldsymbol{e}^{\mathrm{pa}(u_1)})$ into each occurrence of $\boldsymbol{z}^{v_1}$ in $\tilde{f}^{(v_2)}, \ldots, \tilde{f}^{(v_{|\mathcal{I}|})}$. By similarly eliminating the endogenous variables from $v_2$ to $v_{|\mathcal{I}|}$ in an iterative manner, we obtain $\boldsymbol{F}$ satisfying

$$\text{Equation (A.3)} \quad \Rightarrow \quad \boldsymbol{z} = \boldsymbol{F}(\boldsymbol{e}).$$

Since $\boldsymbol{F}$ is constructed by a finite composition of measurable maps, it is measurable. Therefore, for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, we have $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}) \Rightarrow \boldsymbol{z} = \boldsymbol{F}(\boldsymbol{e})$, and hence $\boldsymbol{F}$ is an RSF of $\mathcal{M}$. □

### A.2.4 Compatibility of SCMs and GCMs

*Proof of Proposition 2.4.* The proof here invokes Bongers et al. [29, Corollary 8.3, Proposition 3.4] that showed the (DG) property (Definition A.3, Bongers et al. [29, Definition A.6]) of $\mathrm{obsDist}(\cdot)$ with respect to $\mathrm{obsGraph}(\cdot)$ for acyclic models. To do so, we construct an SCM $\mathcal{M}' := \langle \mathcal{I} \coprod \mathcal{J}, \mathcal{J}', \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{E}}', \boldsymbol{f}', \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ where $(\mathcal{J}', \boldsymbol{\mathcal{E}}')$ is a copy of $(\mathcal{J}, \boldsymbol{\mathcal{E}})$, and $\boldsymbol{f}' : (\boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}) \times \boldsymbol{\mathcal{E}}' \to \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}$ is defined by $\boldsymbol{f}'((\boldsymbol{z}, \boldsymbol{e}), \boldsymbol{e}') = (\boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}), \boldsymbol{e}')$. It is easy to confirm that $\mathcal{M}'$ is acyclic. The following diagram shows the proof outline.

$$\begin{array}{ccc} \mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle & \xrightarrow{\text{Duplicate}} & \mathcal{M}' = \langle \mathcal{I} \coprod \mathcal{J}, \mathcal{J}', \boldsymbol{\mathcal{Z}} \times \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{E}}', \boldsymbol{f}', \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle \\ \Big\downarrow{\scriptstyle \text{dist,graph}} & \swarrow{\scriptstyle \text{obsDist,obsGraph}} & \\ \mathcal{M}_1 = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}} \rangle & & \end{array}$$

We first show that $\mathrm{obsDist}(\mathcal{M}') = \mathrm{dist}(\mathcal{M})$ and $\mathrm{obsGraph}(\mathcal{M}') = \mathrm{graph}(\mathcal{M})$ hold. First, the SE of $\mathcal{M}'$ is

$$\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{e}),\ \boldsymbol{e} = \boldsymbol{e}'.$$

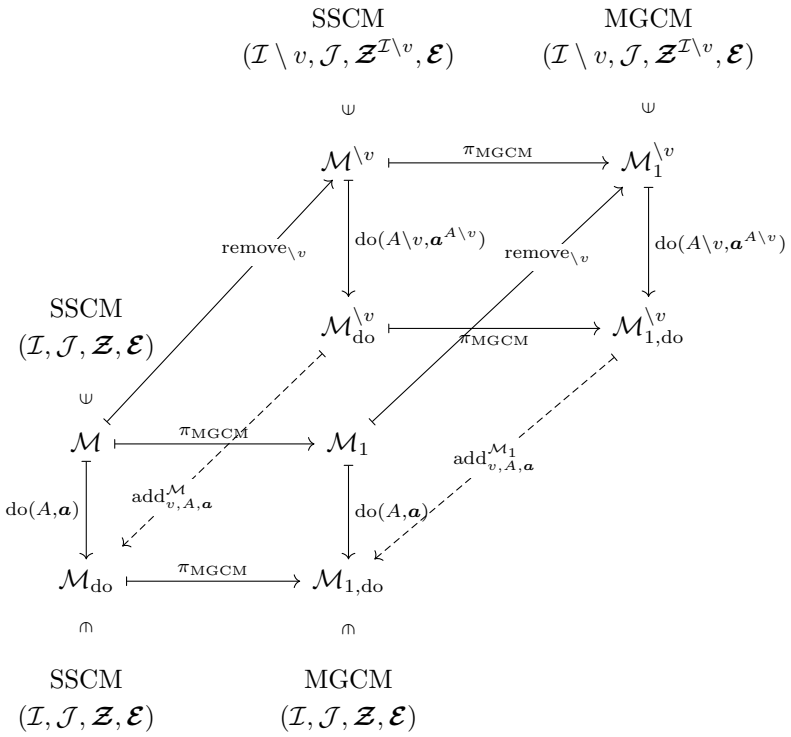
Thus, if $(\boldsymbol{Z}, \boldsymbol{E})$ is a solution of $\mathcal{M}$, then $((\boldsymbol{Z}, \boldsymbol{E}), \boldsymbol{E}')$ with $\boldsymbol{E}'(\cdot) = \boldsymbol{E}(\cdot)$ is a solution of $\mathcal{M}'$. Therefore, we have $\mathrm{obsDist}(\mathcal{M}') = \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} = \mathrm{dist}(\mathcal{M})$. Next, in fact, $\mathrm{graph}(\mathcal{M}')$ is obtained from $\mathrm{graph}(\mathcal{M})$ by adding $\mathcal{J}'$ as new nodes and adding directed edges from $v' \in \mathcal{J}'$ to its corresponding $v \in \mathcal{J}$. This can be confirmed by Definition 2.3; if $v \in \mathcal{I}$, then $\boldsymbol{f'}^{v} = \boldsymbol{f}^{v}$, and hence the same set of edges are drawn among $\mathcal{I}$ and $\mathcal{J}$, and on the other hand, if $v \in \mathcal{J}$, then $\boldsymbol{f'}^{v}$ is the projection map to $\boldsymbol{e'}^{v'}$ where $v' \in \mathcal{J}'$ is the element corresponding $v$, and hence the edge $(v' \to v)$ is drawn and $(u \to v)$ is not drawn for the other $u \neq v'$. Thus, all nodes in $\mathcal{J}'$ are root nodes in $\mathrm{graph}(\mathcal{M}')$, and the rules in Definition 2.3 construct $\mathrm{obsGraph}(\mathcal{M}')$ by removing $\mathcal{J}'$ from $\mathrm{graph}(\mathcal{M}')$ (i.e., taking the induced subgraph for the nodes in $\mathcal{I} \coprod \mathcal{J}$). Therefore, we have $\mathrm{obsGraph}(\mathcal{M}') = \mathrm{graph}(\mathcal{M})$.

Now, since $\mathcal{M}'$ is acyclic, Bongers et al. [29, Corollary 8.3, Proposition 3.4] asserts that $\mathrm{obsDist}(\mathcal{M}')$ satisfies the *directed global Markov property* (Definition A.3, Bongers et al. [29, Definition A.6]) with respect to $\mathrm{obsGraph}(\mathcal{M}')$. This, in turn, is equivalent to the recursive factorization of $\mathrm{obsDist}(\mathcal{M}')$ according to $\mathrm{obsGraph}(\mathcal{M}')$ (Proposition A.1). Therefore, we have that $\mathrm{dist}(\mathcal{M}) = \mathrm{obsDist}(\mathcal{M}') =: \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$ recursively factorizes according to $\mathrm{graph}(\mathcal{M}) = \mathrm{obsGraph}(\mathcal{M}') =: \mathcal{G}^{\mathrm{a}}$. Thus, $\mathcal{M}_1$ is a Markovian GCM.

By definition, $\mathbb{P}_{\boldsymbol{\mathcal{Z}}}$ is the marginal distribution of $\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}$. Since there is no directed edge pointing to $\mathcal{J}$ in $\mathcal{G}^{\mathrm{a}}$ by definition, the construction of $\mathcal{G}$ from $\mathcal{G}^{\mathrm{a}}$ in Definition 2.3 follows the steps of Algorithm 1. Therefore, we have $\mathcal{G} = \pi_{\mathrm{SGCM}}(\mathcal{G}^{\mathrm{a}})$, and as a result, by definition, we have $\mathcal{M}_2 = \pi_{\mathrm{SGCM}}(\mathcal{M}_1)$. □

*Proof of Proposition 2.5.* The proof is by induction on $|\mathcal{I}|$. If $|\mathcal{I}| = 1$, then $\mathrm{do}(A, \boldsymbol{a})(\mathcal{M}) = \mathrm{do}(A, \boldsymbol{a})(\mathcal{M}_1)$ immediately follows from the definitions. Now suppose that the assertion of the proposition holds for $|\mathcal{I}| = n$. Under this inductive hypothesis, we show that the assertion holds in the case of $|\mathcal{I}| = n + 1$.

The following diagram summarizes the proof outline.




(A.4)

Fix $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle \in \mathrm{SSCM}(\mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}})$. Let $v \in \mathcal{I}$ be a terminal node in $\mathcal{G}^{\mathrm{a}}$ (since $\mathcal{G}^{\mathrm{a}}$ is a DAG, there always exists a terminal node, and since $\mathcal{J}$ has no parents, there is one in $\mathcal{I}$). Also let $\mathcal{M}_1 = \pi_{\mathrm{MGCM}}(\mathcal{M})$.

First, because $|\mathcal{I} \setminus v| = n$, by the inductive hypothesis, the following diagram commutes.

$$\begin{array}{ccc} \mathrm{SSCM}(\mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}) & \xrightarrow{\pi_{\mathrm{MGCM}}} & \mathrm{MGCM}(\mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}) \\ {\scriptstyle \mathrm{do}(A\setminus v, \boldsymbol{a}^{A\setminus v})}\downarrow & & \downarrow{\scriptstyle \mathrm{do}(A\setminus v, \boldsymbol{a}^{A\setminus v})} \\ \mathrm{SSCM}(\mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}) & \xrightarrow{\pi_{\mathrm{MGCM}}} & \mathrm{MGCM}(\mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}) \end{array}$$

That is,

$$\pi_{\mathrm{MGCM}} \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) = \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \pi_{\mathrm{MGCM}}.$$

Next, we define two operators, $\mathrm{remove}_{\setminus v}$ and $\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}$, as follows. We first define $\mathrm{remove}_{\setminus v}$ as an operator that removes $v$ from the index set, the space of endogenous variables, and the SF, i.e.,

$$\mathrm{remove}_{\setminus v} : \underbrace{\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle}_{\mathcal{M}} \mapsto \underbrace{\langle \mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}, (\tilde{f}^{(v')})_{v' \neq v}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle}_{\mathcal{M}^{\setminus v}},$$

where $\tilde{f}^{(v)} : \boldsymbol{\mathcal{Z}}^{\mathrm{pa}(v)} \times \boldsymbol{\mathcal{E}}^{\mathrm{pa}(v)} \to \mathcal{Z}^v$ $(v \in \mathcal{I})$ are measurable maps obtained by applying Lemma A.2 to $\mathcal{M}$, and we considered $(\tilde{f}^{(v')})_{v' \neq v}$ as a measurable map from $\boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v} \times \boldsymbol{\mathcal{E}}$ to $\boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}$ in the natural way. On the other hand, we define $\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}$ as an operator that adds $v$ to the index set and $\mathcal{Z}^v$ to the space of endogenous variables, and also concatenates either $\tilde{f}^{(v)}$ (if $v \notin A$) or $\boldsymbol{a}^v$ (otherwise) to the SF, i.e., if $v \notin A$, we have

$$\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}} : \underbrace{\langle \mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}, (\tilde{f}^{(v')})_{v' \neq v}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle}_{\mathcal{M}^{\setminus v}} \mapsto \underbrace{\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, (\tilde{f}^{(v')})_{v' \in \mathcal{I}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle}_{\simeq \mathcal{M}},$$

and otherwise $\boldsymbol{a}^v$ is concatenated to the SF. Then, since $v$ is a terminal node, the following diagram commutes (with a slight abuse of terminology):

$$\begin{array}{ccc} \mathcal{M} & \xmapsto{\mathrm{remove}_{\setminus v}} & \mathcal{M}^{\setminus v} \\ {\scriptstyle \mathrm{do}(A,\boldsymbol{a})}\downarrow & & \downarrow{\scriptstyle \mathrm{do}(A\setminus v, \boldsymbol{a}^{A\setminus v})} \\ \mathcal{M}_{\mathrm{do}} & \xleftarrow[\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}]{} & \mathcal{M}^{\setminus v}_{\mathrm{do}} \end{array}$$

in the sense that

$$\mathrm{do}(A, \boldsymbol{a})(\mathcal{M}) \simeq \mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}} \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \mathrm{remove}_{\setminus v}(\mathcal{M}).$$

Similarly, we define $\mathrm{remove}_{\setminus v}$ and $\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}}$ for GCMs as follows. $\mathrm{remove}_{\setminus v}$ is defined as an operator that removes $v$ from the index set, the space of endogenous variables, and the graph, and marginalizes out $v$ in the distribution, i.e.,

$$\mathrm{remove}_{\setminus v} : \underbrace{\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle}_{\mathcal{M}_1} \mapsto \underbrace{\langle \mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}_{\mathcal{I}\setminus v}, \mathrm{marg}_{\setminus v}(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) \rangle}_{\mathcal{M}_1^{\setminus v}},$$

where $\mathrm{marg}_{\setminus v}$ is an operator to marginalize out $v$. On the other hand, we define $\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}}$ as an operator that adds $v$ to the index set, the space of endogenous variables, and adds to $v$ the graph either with the edges pointing to $v$ in $\mathcal{G}^{\mathrm{a}}$ (if $v \notin A$) or without the edges (otherwise), and integrates

(in the sense of Fact A.1) either the Markov kernel $K^v(\boldsymbol{z}^{\mathrm{pa}(v)}, \mathrm{d}\boldsymbol{z}^v)$ of $\mathcal{M}_1$ corresponding to $v$ (if $v \notin A$) or the Dirac measure $\delta_{\boldsymbol{a}^v}(\mathrm{d}\boldsymbol{z}^v)$ (otherwise), i.e., if $v \notin A$, we have

$$\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}} : \underbrace{\langle \mathcal{I} \setminus v, \mathcal{J}, \boldsymbol{\mathcal{Z}}^{\mathcal{I}\setminus v}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}_{\mathcal{I}\setminus v}, \mathrm{marg}_{\setminus v}(\mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}}) \rangle}_{\mathcal{M}_1^{\setminus v}} \mapsto \underbrace{\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{E}}, \mathcal{G}^{\mathrm{a}}, \mathbb{P}_{\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{E}}} \rangle}_{\mathcal{M}_1},$$

and otherwise $\delta_{\boldsymbol{a}^v}(\mathrm{d}\boldsymbol{z}^v)$ is integrated with the distribution. Then, since $v$ is a terminal node, the following diagram commutes (again, with a slight abuse of terminology):

$$\begin{array}{ccc} \mathcal{M}_1 & \xrightarrow{\mathrm{remove}_{\setminus v}} & \mathcal{M}_1^{\setminus v} \\ \Big\downarrow{\scriptstyle \mathrm{do}(A,\boldsymbol{a})} & & \Big\downarrow{\scriptstyle \mathrm{do}(A\setminus v, \boldsymbol{a}^{A\setminus v})} \\ \mathcal{M}_{1,\mathrm{do}} & \xleftarrow[\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}}]{} & \mathcal{M}_{1,\mathrm{do}}^{\setminus v} \end{array}$$

in the sense that

$$\mathrm{do}(A, \boldsymbol{a})(\mathcal{M}_1) = \mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}} \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \mathrm{remove}_{\setminus v}(\mathcal{M}_1).$$

The following diagram also commutes

$$\begin{array}{ccc} \mathcal{M} & \xrightarrow{\mathrm{remove}_{\setminus v}} & \mathcal{M}^{\setminus v} \\ \Big\downarrow{\scriptstyle \pi_{\mathrm{MGCM}}} & & \Big\downarrow{\scriptstyle \pi_{\mathrm{MGCM}}} \\ \mathcal{M}_1 & \xrightarrow{\mathrm{remove}_{\setminus v}} & \mathcal{M}_1^{\setminus v} \end{array}$$

in the sense that

$$\pi_{\mathrm{MGCM}} \circ \mathrm{remove}_{\setminus v}(\mathcal{M}) = \mathrm{remove}_{\setminus v} \circ \pi_{\mathrm{MGCM}}(\mathcal{M}),$$

because of the following consideration. First, we have $\mathrm{dist}(\mathcal{M}_1^{\setminus v}) := \mathrm{dist}(\mathcal{M}^{\setminus v}) = \mathrm{marg}_{\setminus v}(\mathrm{dist}(\mathcal{M}))$ by considering that an RSF of $\mathcal{M}^{\setminus v}$ can be obtained by ignoring $v$ in an RSF of $\mathcal{M}$ since $v$ is a terminal node. Second, we have that $\mathrm{graph}(\mathcal{M}_1^{\setminus v}) = \mathrm{graph}(\mathcal{M}^{\setminus v})$ is obtained by removing $v$ from $\mathrm{graph}(\mathcal{M}) = \mathrm{graph}(\mathcal{M}_1)$ since $v$ is a terminal node.

Also, the following diagram commutes

$$\begin{array}{ccc} \mathcal{M}_{\mathrm{do}} & \xleftarrow{\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}} & \mathcal{M}_{\mathrm{do}}^{\setminus v} \\ \Big\downarrow{\scriptstyle \pi_{\mathrm{MGCM}}} & & \Big\downarrow{\scriptstyle \pi_{\mathrm{MGCM}}} \\ \mathcal{M}_{1,\mathrm{do}} & \xleftarrow[\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}}]{} & \mathcal{M}_{1,\mathrm{do}}^{\setminus v} \end{array}$$

in the sense that

$$\pi_{\mathrm{MGCM}} \circ \mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}(\mathcal{M}_{\mathrm{do}}^{\setminus v}) = \mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}} \circ \pi_{\mathrm{MGCM}}(\mathcal{M}_{\mathrm{do}}^{\setminus v}),$$

because of the following. As for the graphs, since $\mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}$ only adds the SF corresponding to $v$, the difference of $\mathrm{graph}(\mathcal{M}_{\mathrm{do}})$ from $\mathrm{graph}(\mathcal{M}_{\mathrm{do}}^{\setminus v})$ is that $v$ is added as a node and the edges are either added from $\mathrm{pa}(v)$ to $v$ (if $v \notin A$) or not added (otherwise). This relation matches the definition of $\mathrm{add}^{\mathcal{M}_1}_{v,A,\boldsymbol{a}}$. As for the distribution, first notice that any solution of $\mathcal{M}_{\mathrm{do}} = \mathrm{add}^{\mathcal{M}}_{v,A,\boldsymbol{a}}(\mathcal{M}_{\mathrm{do}}^{\setminus v})$ satisfies, depending on whether $v \in A$,

$$\boldsymbol{Z}^v = \tilde{f}^{(v)}(\boldsymbol{Z}^{\mathrm{pa}(v)}, \boldsymbol{E}^{\mathrm{pa}(v)}) \quad \text{a.s.}, \qquad \text{or} \qquad \boldsymbol{Z}^v = \boldsymbol{a}^v \quad \text{a.s.}, \tag{A.5}$$

and also

$$\boldsymbol{Z}^{\mathcal{I}\setminus v} = \boldsymbol{F}_{\setminus v}(\boldsymbol{E}) \quad \text{a.s.}, \tag{A.6}$$

where $\boldsymbol{F}_{\setminus v}$ is an RSF of $\mathcal{M}_{\mathrm{do}}^{\setminus v}$, since $v$ is a terminal node. Therefore, letting

$$k(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)}) := \begin{cases} \tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)}) & \text{if } v \notin A, \\ \boldsymbol{a}^v & \text{otherwise,} \end{cases}$$

we obtain

$$\begin{aligned} \mathrm{dist}(\mathcal{M}_{1,\mathrm{do}}) = \mathrm{dist}(\mathcal{M}_{\mathrm{do}}) &= \delta_{k(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})}(\mathrm{d}\boldsymbol{z}^v)\mathrm{marg}_{\setminus v}(\mathrm{dist}(\mathcal{M}_{\mathrm{do}})) \\ &= \delta_{k(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})}(\mathrm{d}\boldsymbol{z}^v)\mathrm{dist}(\mathcal{M}_{\mathrm{do}}^{\setminus v}). \end{aligned}$$

where the third equality follows from Equation (A.5) and Lemma A.1, and the fourth equality follows from Equation (A.6). On the other hand, recall that $\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1}$ integrates the Markov kernel corresponding to $v$ in $\mathcal{M}_1$. The Markov kernel integrated by $\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1}$ is either $\delta_{\tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})}$ or $\delta_{\boldsymbol{a}^v}$, depending on whether $v \in A$, because of the following: since $\mathrm{dist}(\mathcal{M}_1) = \mathrm{dist}(\mathcal{M})$ and any solution $(\boldsymbol{Z}, \boldsymbol{E})$ of $\mathcal{M}$ almost surely satisfies $\boldsymbol{Z}^v = \tilde{f}^{(v)}(\boldsymbol{Z}^{\mathrm{pa}(v)}, \boldsymbol{E}^{\mathrm{pa}(v)})$ (Lemma A.2), we have, in light of Lemma A.1,

$$\begin{aligned} \mathrm{dist}(\mathcal{M}_1) = \mathrm{dist}(\mathcal{M}) &= \delta_{\tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})}(\mathrm{d}\boldsymbol{z}^v)\mathrm{marg}_{\setminus v}(\mathrm{dist}(\mathcal{M})) \\ &= \delta_{\tilde{f}^{(v)}(\boldsymbol{z}^{\mathrm{pa}(v)}, \boldsymbol{e}^{\mathrm{pa}(v)})}(\mathrm{d}\boldsymbol{z}^v)\mathrm{marg}_{\setminus v}(\mathrm{dist}(\mathcal{M}_1)). \end{aligned}$$

Therefore, $\mathrm{dist}(\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1}(\mathcal{M}_{1,\mathrm{do}}^{\setminus v})) = \mathrm{dist} \circ \mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}}(\mathcal{M}_{\mathrm{do}}^{\setminus v})$.

Finally, by chasing the diagram (A.4), we can see

$$\begin{aligned} &(\pi_{\mathrm{MGCM}} \circ \mathrm{do}(A, \boldsymbol{a}))(\mathcal{M}) \\ &= (\pi_{\mathrm{MGCM}} \circ (\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}} \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \mathrm{remove}_{\setminus v}))(\mathcal{M}) \\ &= ((\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1} \circ \pi_{\mathrm{MGCM}}) \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \mathrm{remove}_{\setminus v})(\mathcal{M}) \\ &= (\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1} \circ (\mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ \pi_{\mathrm{MGCM}}) \circ \mathrm{remove}_{\setminus v})(\mathcal{M}) \\ &= (\mathrm{add}_{v,A,\boldsymbol{a}}^{\mathcal{M}_1} \circ \mathrm{do}(A \setminus v, \boldsymbol{a}^{A\setminus v}) \circ (\mathrm{remove}_{\setminus v} \circ \pi_{\mathrm{MGCM}}))(\mathcal{M}) \\ &= (\mathrm{do}(A, \boldsymbol{a}) \circ \pi_{\mathrm{MGCM}})(\mathcal{M}). \end{aligned}$$

□

# Appendix B

# Appendices for Chapter 3

Table B.1 summarizes the abbreviations and the symbols used in the chapter. For notation simplicity, when $\overline{\mathcal{Z}}^j$ is a finite set, we identify it with $\mathbb{Z}/m\mathbb{Z}$ where $m$ is the cardinality of $\overline{\mathcal{Z}}^j$, to justify the subtractions inside the kernel functions.

## B.1 Real-world Data Experiment Details

Here, we describe the implementation details of the experiment using the real-world data. The experiment was implemented using the *hydra* package of Python [303]. All experiments were carried out on a 2.60 GHz Intel® Xeon® CPUs with 132 GB memory.

Our experiment code can be found at https://github.com/takeshi-teshima/incorporating-causal-graphical-prior-knowledge-into-predictive-modeling-via-simple-data-augmentation.

### B.1.1 Data Set Details

The following are the data acquisition procedures, the sample sizes, the variable definitions, and the preprocessing procedures used in our experiment. In all the data sets, after preprocessing as described below, we independently normalized each variable as a final preprocessing step.

**Sachs data [227].** This data set consists of continuous measurements from the flow cytometry of proteins and phospholipids in human immune system cells. The *consensus graph* is provided in Sachs et al. [227] based on the conventionally accepted cellular signaling networks (Figure 3.4(a)). Among the eight data sets corresponding to different intervention conditions [227], we use the one that is *observational*, i.e., without any interventions. The data set contains 853 observations of 11 variables, namely *Raf*, *Mek*, *Plcg*, *PIP2*, *PIP3*, *Erk*, *Akt*, *PKA*, *PKC*, *P38*, and *Jnk*. Among these, for demonstration purposes, we considered *PKA* as the target attribute. As preprocessing, we log-transformed *Raf*, *Mek*, and *PKA*.

**GSS data [240].** This data set is concerning the status attainment theory in sociology. This data set is originally part of the General Social Survey (GSS)[1], and we used a subset of the data that was previously used in the causal discovery literature [240]. The reference graph is based on domain knowledge of the status attainment model ([66]; Figure 3.4(b)). The acquired data set consists of 1380 observations of 6 variables, namely $x_1$: father's occupation level, $x_2$: son's income, $x_3$: father's education, $x_4$: son's occupation, $x_5$: son's education, and $x_6$: the number of siblings. We consider $x_4$ as the target variable.

**Boston Housing data [98].** This data set is concerning the house prices in Boston, and the objective is to predict the prices of the house from its attributes. We acquired the data from https://github.com/adityatiwari13/Boston_Dataset. The acquired data set consists of 506 observations of 13 variables, namely

[1] https://gss.norc.org/

Table B.1: List of abbreviations and symbols used in the chapter.

| Abbreviation / Symbol | Meaning |
|---|---|
| CG/GCM | Causal Graph / Graphical Causal Model |
| ADMG | Acyclic Directed Mixed Graph |
| DAG/PAG | Directed Acyclic Graph / Partial Ancestral Graph |
| MSE | Mean Squared Error |
| $\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{>0}, \mathbb{Z}, \mathbb{N}_0, \mathbb{N}$ | Set of all reals, nonnegative reals, positive reals, integers, nonnegative integers, and positive integers. |
| $\mathbb{1}[A]$ | Indicator function, i.e., 1 if $A$ holds true and 0 otherwise. |
| $X \perp\!\!\!\perp Y \mid Z$ | $X$ and $Y$ are conditionally independent given $Z$. |
| $\coprod$ | Disjoint union of sets. |
| $\operatorname{diag}((x_1, \ldots, x_d))$ | Diagonal matrix with diagonal elements $(x_1, \ldots, x_d)$ ($d \in \mathbb{N}$). |
| $\|\cdot\|, \|\cdot\|_\infty, \|\cdot\|_{\mathrm{op}}, \det$ | Euclidean norm of a vector, the supremum norm of a function, the operator norm and the determinant of a matrix. |
| $\lfloor \cdot \rfloor$ | $\lfloor a \rfloor := \max\{z \in \mathbb{Z} : z \leq a\}$ for $a \in \mathbb{R}$. |
| $\delta_{\boldsymbol{z}}$ | Dirac's delta function centered at $\boldsymbol{z}$ (e.g., [317, Section E.4.1]). |
| $\Delta_K$ | $(K-1)$-dimensional probability simplex [33, Example 2.5]. |
| $[N:M], [N]$ | $[N:M] := \{N, N+1, \ldots, M\}$ and $[N] := [1:N]$ where $N, M \in \mathbb{N}$ and $N \leq M$. |
| $\boldsymbol{x}^S$ | $\boldsymbol{x}^S := (x^{s_1}, \ldots, x^{s_{\lvert S \rvert}})$ where $\boldsymbol{x} = (x^1, \ldots, x^n)$ is an $n$-dimensional vector and $S = \{s_1, \ldots, s_{\lvert S \rvert}\} \subset [n]$ with $s_1 < \cdots < s_{\lvert S \rvert}$. |
| $[0] = \emptyset, \mathbb{R}^0 := \{0\}$<br>$\boldsymbol{x}^\emptyset = 0, [N]^0 := \{0\}$ | Conventions used in the chapter. |
| $d \in \mathbb{N}$ | Overall data dimensionality (with $X$ and $Y$ combined). |
| $\boldsymbol{\mathcal{Z}} = \prod_{j=1}^d \boldsymbol{\mathcal{Z}}_j$ | Overall data space (without distinguishing $X$ and $Y$). |
| $\mathcal{X} = \prod_{j \in [d] \setminus j^*} \overline{\mathcal{Z}}^j$<br>$\mathcal{Y} = \overline{\mathcal{Z}}^{j^*}$ | Input variable space and target variable space. |
| $p$ | Joint probability density of $\boldsymbol{Z} := (Z^1, \ldots, Z^d)$ taking values in $\boldsymbol{\mathcal{Z}}$. |
| $\mathrm{Rad}_{m,q}$ | Rademacher complexity of a function class. |
| $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ | Hypothesis set. |
| $\ell : \mathcal{H} \times \left(\prod_{j=1}^d \overline{\mathcal{Z}}^j\right) \to \mathbb{R}$ | Loss function. |
| $\mathcal{R}(h) = \mathbb{E}[\ell(h, \boldsymbol{Z})]$ | Risk functional for $h \in \mathcal{H}$. |
| $\mathcal{D} = \{\boldsymbol{Z}_i\}_{i=1}^n$ | Independently and identically distributed sample from $p$. |
| $\mathcal{G} = ([d], \hat{\mathfrak{D}}, \hat{\mathfrak{B}})$ | Underlying ADMG for which $p$ satisfies the topological ADMG |
| $\hat{\mathcal{G}} = ([d], \hat{\mathfrak{D}}, \hat{\mathfrak{B}})$ | factorization and its estimator. |
| $\mathrm{dis}(\cdot), \mathrm{pa}(\cdot), \mathrm{mp}(j)$ | District, parents, and Markov pillow of vertex $j \in [d]$. |
| $p_{j\mid \mathrm{mp}(j)}, p_{j,\mathrm{mp}(j)}, p_{\mathrm{mp}(j)}$ | Conditional density of $\boldsymbol{Z}^{(j)}$ given $\boldsymbol{Z}^{\mathrm{mp}(j)}$, the joint density of $(\boldsymbol{Z}^{(j)}, \boldsymbol{Z}^{\mathrm{mp}(j)})$, and the marginal density of $\boldsymbol{Z}^{\mathrm{mp}(j)}$. |
| $K^j : \overline{\mathcal{Z}}^{\mathrm{mp}(j)} \to \mathbb{R}$ | Kernel function (we define $K^j := 1$ if $\mathrm{mp}(j) = \emptyset$). |
| $\boldsymbol{Z}_{\boldsymbol{i}}$ | $\boldsymbol{Z}_{\boldsymbol{i}} = (Z^1_{i_1}, \ldots, Z^d_{i_d})$ for $\boldsymbol{i} = (i_1, \ldots, i_d) \in [n]^d$. |
| $\mathcal{D}_{\mathrm{aug}} := \{\boldsymbol{Z}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^d}$ | Augmented data set. |
| $\mathcal{W}_{\mathrm{aug}} := \{\hat{w}_{\boldsymbol{i}}\}_{\boldsymbol{i} \in [n]^d}$ | Instance weights on the augmented data set. |
| $\hat{\mathcal{R}}_{\mathrm{emp}}, \hat{\mathcal{R}}_{\mathrm{aug}}$ | Empirical risk and the proposed risk estimator. |
| $\Omega(h)$ | Regularization term for $h \in \mathcal{H}$. |
| $\lambda \in [0, 1]$ | Convex combination coefficient used in $(1-\lambda)\hat{\mathcal{R}}_{\mathrm{emp}}(h) + \lambda \hat{\mathcal{R}}_{\mathrm{aug}}(h) + \Omega(h)$. |
| $K^j_{j'}$ | Component of the product kernel $K^j$ for $j' \in \mathrm{mp}(j)$. |
| $\theta$ | Pruning threshold of the small weights in Algorithm 2. |

*CRIM*, *ZN*, *INDUS*, *CHAS*, *NOX*, *RM*, *AGE*, *DIS*, *RAD*, *TAX*, *PTRATIO*, *B*, *LSTAT*, and *MEDV*. The objective is to predict the value of prices of the house, i.e., *MEDV*, using the given features.

**Auto MPG data [212].** This data set concerns the city-cycle fuel consumption in miles per gallon (MPG). We acquired the data from https://archive.ics.uci.edu/ml/datasets/Auto+MPG. The acquired data set consists of 398 observations of 9 variables, namely *mpg*, *cylinders*, *displacement*, *horsepower*, *weight*, *acceleration*, *model year*, *origin*, and *car name*. Among these, we discard *origin* and *car name*, and we consider *mpg* as the predicted variable.

**White Wine data [52].** This data set is concerning the prediction of wine quality from its physicochemical attributes. We acquired the data from https://archive.ics.uci.edu/ml/datasets/wine+quality. The acquired data set consists of 4898 observations of 12 variables, namely *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates*, *alcohol*, and *quality*. Among the variables, we consider the *quality* variable as the target.

**Red Wine data [52].** This data set is concerning the prediction of wine quality from its physicochemical attributes. We acquired the data from https://archive.ics.uci.edu/ml/datasets/wine+quality. The acquired data set consists of 1599 observations of 12 variables, namely *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates*, *alcohol*, and *quality*. Among these, we consider the *quality* variable as the target.

### B.1.2 Predictor Model Details

For the implementation of the predictor model, we employed the *xgboost* library of Python [41]. See Chen and Guestrin [41] for the optimization method and the other details.

### B.1.3 Proposed Method Implementation Details

For continuous variables, we compute the kernel bandwidths as follows. We first specify the *bandwidth temperature* $\gamma > 0$ as a hyper-parameter. Then we calculate the rule-of-thumb bandwidth $h_j^{\text{thumb}}$ for each $j \in [d]$ using the training data $\{\boldsymbol{Z}_i^j\}_{i=1}^n$. Finally, we set $h_j = \gamma \cdot h_j^{\text{thumb}}$. In the experiment, we fix $\gamma = 10^{-3}$ throughout all runs.

For the rule-of-thumb kernel bandwidth, we employed *Silverman*'s rule-of-thumb [250, pp.45–47, Equations (3.28) and (3.30) therein] implemented in the *statsmodels* package of Python [207], namely, $h^{\text{thumb}} = \left(\frac{4}{3}\right)^{1/5} An^{-1/5}$ where $A = \min\{\hat{\sigma}, \text{IQR}/1.349\}$, $\hat{\sigma}$ is the square root of the unbiased estimator of the variance, and IQR is the interquantile range.

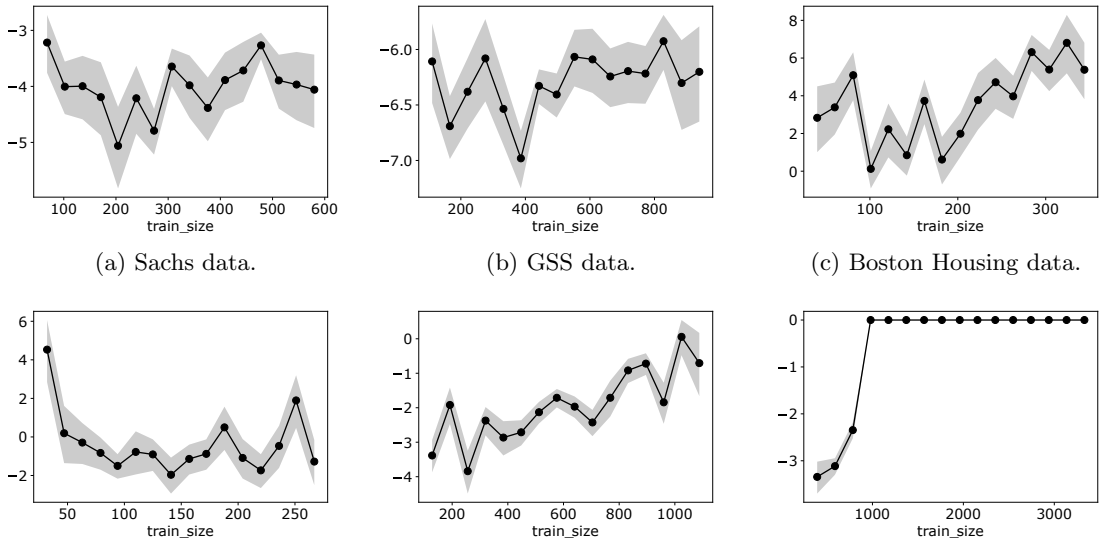
For the pruning threshold, we use $\theta = 10^{-3} \cdot n^{-1}$.

### B.1.4 Causal Discovery Method Configuration

We perform *DirectLiNGAM* [240] on the data sets to simulate a situation where we have access to domain knowledge. As the independence measure used in the DirectLiNGAM framework, we employ the pairwise likelihood ratio score [125] that is based on a nonparametric approximation to the mutual information.

### B.1.5 Supplementary Figures

Figure B.1 shows the average improvement achieved by the proposed method relative to the baseline without a device. The improvement in the small-data regime is consistently observed except in a few cases in the *Auto MPG* and the *Boston Housing* data. In the *Boston Housing* data set, the performance loss may be due to the failure of the CG estimation since the performance loss is magnified as the training set size is increased. In the *Auto MPG* data, the performance degradation for the smallest training set fraction may be due to the additional complexity and bias introduced by the kernel approximation.




(a) Sachs data. (b) GSS data. (c) Boston Housing data.

(d) Auto MPG data. (e) Red Wine data. (f) White Wine data.

Figure B.1: Average relative improvement in percentage. In all figures, the horizontal axis is the varied sizes of the original training data before augmentation. The vertical axis is the relative MSE improvement in percentage, i.e., $\frac{\mathrm{MSE}_{\mathrm{prop}}-\mathrm{MSE}_{\mathrm{base}}}{\mathrm{MSE}_{\mathrm{base}}} \times 100\ \%$ where $\mathrm{MSE}_{\mathrm{base}}$ and $\mathrm{MSE}_{\mathrm{prop}}$ are the MSE of the baseline and that of the proposed method, respectively (the lower the better). The markers and the lines indicate the average over the 20 independent runs, and the shades are drawn for the width of the standard errors both above and below the lines. In most of the cases, the proposed method shows a consistently improved performance compared to the baseline based on the empirical risk minimization with the same hypothesis class, particularly in the small-data regime.

## B.2 Synthetic-data Experiment Details

Here, we explain the implementation details of the synthetic-data experiment in Section 3.5.2. All experiments were carried out on a 2.60 GHz Intel® Xeon® CPUs with 132 GB memory.

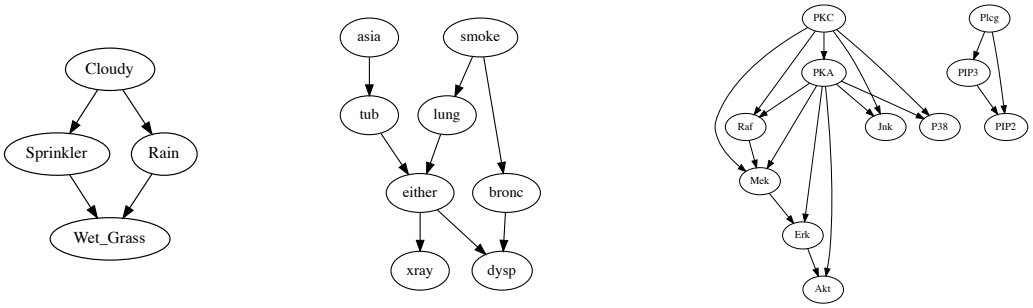
**Data sets.** We use three sets of CGs and conditional probability tables (CPDs) for generating the synthetic data sets, namely *sprinkler*, *asia*, and *sachs*. The ground-truth CGs are shown in Figure B.2. The data generation was implemented by using the *bnlearn* package of Python [266].

***sprinkler* data [266].** This data set consists of 4 variables, among which we considered the *Rain* variable as the target variable to be predicted. The CPD for generating the data was the one implemented in the *bnlearn* package [266].

***asia* data [159].** This data set consists of 8 variables, among which we considered the *smoke* variable as the target variable to be predicted. The CPD for generating the data was acquired from `https://erdogant.github.io/datasets/asia.zip`.

***sachs* data [227].** This data set consists of 11 variables, among which we considered the *PKA* variable as the target variable to be predicted. The CPD for generating the data was acquired from `https://erdogant.github.io/datasets/sachs.zip`.

**Predictor model class.** For the predictor model class, we employed the GBRTs [81, 41] using the same configuration as the main experiment (Section 3.5) with the following hyper-parameter candidates: the number of leaves was fixed as $M = 64$, the number of boosting rounds $K$ was searched in $\{10, 50, 250, 1250\}$, and the $\ell_2$-regularization coefficient $\rho$ in $\{10^{-1}, 10^{-2}, 10^{-3}\}$.




(a) sprinkler (b) asia (c) sachs

Figure B.2: Ground-truth CGs of the synthetic data sets used in our synthetic-data experiment: (a) *sprinkler* data, (b) *asia* data, and (c) *sachs* data.

**Implementation of the proposed method.** Because the synthetic data sets are fully categorical, i.e., $M_j := \left|\overline{\mathcal{Z}}^j\right| < \infty$, we implemented the proposed method in this experiment using the device described in Section 3.3.4. The efficient implementation is derived by the following calculation:

$$\begin{aligned}
\hat{p}_{j|\mathrm{mp}(j)}(Z^j|\boldsymbol{Z}^{\mathrm{mp}(j)}) &= \frac{\sum_{i=1}^n \mathbb{1}\left[Z^j = Z_i^j\right] \mathbb{1}\left[\boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_i^{\mathrm{mp}(j)}\right]}{\sum_{k=1}^n \mathbb{1}\left[\boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_k^{\mathrm{mp}(j)}\right]} \\
&= \frac{\sum_{r=1}^{M_j} \sum_{i:Z_i^j=r} \mathbb{1}\left[Z^j = Z_i^j\right] \mathbb{1}\left[\boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_i^{\mathrm{mp}(j)}\right]}{\sum_{k=1}^n \mathbb{1}\left[\boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_k^{\mathrm{mp}(j)}\right]} \\
&= \frac{\sum_{r=1}^{M_j} \mathbb{1}\left[Z^j = r\right] \left|\{i : Z_i^j = r, \boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_i^{\mathrm{mp}(j)}\}\right|}{\left|\{k : \boldsymbol{Z}^{\mathrm{mp}(j)} = \boldsymbol{Z}_k^{\mathrm{mp}(j)}\}\right|} \\
&= \sum_{r=1}^{M_j} \mathbb{1}\left[Z^j = r\right] \frac{\hat{m}_j(r, \boldsymbol{Z}^{\mathrm{mp}(j)})}{\sum_{r'=1}^{M_j} \hat{m}_j(r', \boldsymbol{Z}^{\mathrm{mp}(j)})},
\end{aligned}$$

where $\hat{m}_j(r, \boldsymbol{Z}^{\mathrm{mp}(j)}) := \left|\{i : Z_i^j = r, \boldsymbol{Z}_i^{\mathrm{mp}(j)} = \boldsymbol{Z}^{\mathrm{mp}(j)}\}\right|$.

## B.3 Details and Proof of the Theoretical Analysis

### B.3.1 Notation and Problem Setup

**Basic notation.** Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{N}$ that of positive integers, $\mathbb{R}_{>0}$ that of positive real numbers, $\mathbb{Z}$ that of integers, and $\mathbb{N}_0$ that of non-negative integers. For $(x_1, \ldots, x_k) \in \mathbb{R}^k$, $\mathrm{diag}((x_1, \ldots, x_k))$ denotes the diagonal matrix whose diagonal elements are $(x_1, \ldots, x_k)$. For a vector, $\|\cdot\|$ denotes its Euclidean norm. For a matrix, det denotes its determinant, and $\|\cdot\|_{\mathrm{op}}$ its operator norm. For a function, $\|\cdot\|_\infty$ denotes its supremum norm over a suitable set of inputs when the domain is clear from the context. For a finite set, $|\cdot|$ denotes its cardinality.

**Utility notation.** For $n \in \mathbb{N}$, define $[n] := \{1, 2, \ldots, n\}$. For $n, m \in \mathbb{N}$ with $n \leq m$, define $[n : m] := \{n, n+1, \ldots, m\}$. For an $n$-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $S \subset [n]$, we let $\boldsymbol{x}^S = (x_{s_1,\ldots,s_{|S|}})$ denote its sub-vector with indices in $S = \{s_1, \ldots, s_{|S|}\}$ with $s_1 < \cdots < s_{|S|}$. Similarly, for $j \in [n]$, we let $\boldsymbol{x}^j := \boldsymbol{x}^{\{j\}}$. For $S \subset [n]$, we also define $\boldsymbol{\mathcal{Z}}^S := \prod_{k \in S} \boldsymbol{\mathcal{Z}}^k$. To simplify the notation, we use the convention of $\mathbb{R}^0 := \{0\}$, $\boldsymbol{x}^\emptyset = 0$, and $[n]^{j-1} = \{0\}$.

**Distribution and sample.** Let $d \in \mathbb{N}$. In this theoretical analysis, we assume that $\boldsymbol{\mathcal{Z}}^j$ is a measurable subset of $\mathbb{R}$ ($j \in [d]$). We consider a probability distribution over $\mathcal{Z} := \prod_{j=1}^d \boldsymbol{\mathcal{Z}}^j$, and let $p$ denote its density function (assuming it exists). We are given $\mathcal{D} = \{\boldsymbol{Z}_i\}_{i=1}^n$, an independently and identically distributed sample from $p$. Let $\mathbb{E}$ denote the expectation with respect to $p$. Additionally, we are given an ADMG $\mathcal{G} = ([d], \hat{\mathfrak{D}}, \hat{\mathfrak{B}})$. Let $\mathrm{mp}(j) \subset [d]$ denote the Markov pillow of $j \in [d]$. Throughout this section, we assume $p$ satisfies the topological ADMG factorization relation according to $\mathcal{G}$ [24]:

$$p(\boldsymbol{z}) = \prod_{j=1}^d p_{j|\mathrm{mp}(j)}(\boldsymbol{z}^j | \boldsymbol{z}^{\mathrm{mp}(j)}) \quad \left( = \prod_{j=1}^d \frac{p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)})}{p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})} \right).$$

**Learning problem.** Let $\mathcal{H}$ denote a hypothesis class, and let $\ell : \mathcal{H} \times \mathbb{R}^d \to \mathbb{R}_{>0}$ be a loss function. To simplify the notation, we define $\ell_h := \ell(h, \cdot)$ and $\mathcal{L}_{\mathcal{H}} := \{\ell_h : f \in \mathcal{H}\}$. For each $h \in \mathcal{H}$, we define the risk functional $\mathcal{R}(h) := \mathbb{E}[\ell_h(\boldsymbol{Z})]$. The learning problem is to find a hypothesis $\hat{h} \in \mathcal{H}$ for which $\mathcal{R}$ is small, given the training data $\mathcal{D}$ and the graph $\mathcal{G}$.

**Proposed method.** For each $j \in [d]$, we fix a kernel function $K^j : \mathbb{R}^{|\mathrm{mp}(j)|} \to \mathbb{R}$. For notation simplicity, we define $K^j := 1$ for $j$ such that $\mathrm{mp}(j) = \emptyset$. We also fix $\boldsymbol{h} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^d) \in \mathbb{R}^d_{>0}$. Then, we define

$$\mathbf{H}_j := \mathrm{diag}(\boldsymbol{h}^{\mathrm{mp}(j)}), \qquad K^j_{\mathbf{H}}(u) := \frac{1}{|\det \mathbf{H}_j|} K^j(\mathbf{H}_j^{-1} u).$$

For $\boldsymbol{i} = (i_1, \ldots, i_d)$ and $\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathbb{R}^{|\mathrm{mp}(j)|}$, define

$$\hat{w}_i^j(\boldsymbol{z}^{\mathrm{mp}(j)}) := \frac{K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)})}{\sum_{i=1}^n K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)})} \mathbb{1}\left[\sum_{i=1}^n K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)}) \neq 0\right]$$

where $\boldsymbol{i} = (i_1, \ldots, i_d)$, $\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathbb{R}^{|\mathrm{mp}(j)|}$. Then, we recursively define

$$\hat{w}_{\boldsymbol{i}_{1:0}} = 1, \quad \hat{w}_{\boldsymbol{i}_{1:j}} = \hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} \cdot \hat{w}_{\boldsymbol{i}_{1:j-1}} \ (j \in [d], \boldsymbol{i}_{1:j-1} \in [n]^{j-1}),$$

where

$$\hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} := \hat{w}^j_{i_j}\left(\boldsymbol{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}\right), \quad Z_{\boldsymbol{i}_{1:j-1}} = \left(Z^1_{i_1}, \ldots, Z^{j-1}_{i_{j-1}}\right).$$

Here, we use the convention $Z^{\mathrm{mp}(1)}_{\boldsymbol{i}_{1:0}} := 0$ to be consistent with the notation. Using this notation, for $h \in \mathcal{H}$, define the augmented empirical risk estimator

$$\hat{\mathcal{R}}_{\mathrm{aug}}(h) := \sum_{\boldsymbol{i} \in [n]^d} \hat{w}_{\boldsymbol{i}} \ell_h(\boldsymbol{Z}_{\boldsymbol{i}}).$$

**Target of the theoretical analysis.** We aim to provide a stochastic upper bound on $\mathcal{R}(\hat{h}) - \mathcal{R}(h^*)$, where

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \{\hat{\mathcal{R}}_{\mathrm{aug}}(h)\}, \text{ and } h^* \in \arg\min_{h \in \mathcal{H}} \{\mathcal{R}(h)\},$$

assuming both exist.

**Notation for stating the results.** To state the main theorem, we use the following notation. For each $j \in [d]$ and $h \in \mathcal{H}$, define

$$\ell_{h,j} : \begin{pmatrix} \boldsymbol{z}^1 \\ \vdots \\ \boldsymbol{z}^j \end{pmatrix} \mapsto \int_{\boldsymbol{\mathcal{Z}}^{[j+1:d]}} \ell_h(\boldsymbol{z}) \left( \prod_{k=j+1}^d p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k | \boldsymbol{z}^{\mathrm{mp}(k)}) \right) \mathrm{d}\boldsymbol{z}^{j+1} \cdots \mathrm{d}\boldsymbol{z}^d.$$

Also define

$$\mathcal{L}_{\mathcal{H}}^{j} := \left\{ \ell_{h,j}(\boldsymbol{z}^1, \dots, \boldsymbol{z}^{j-1}, \cdot) : h \in \mathcal{H}, (\boldsymbol{z}^1, \dots, \boldsymbol{z}^{j-1}) \in \boldsymbol{\mathcal{Z}}^{[1:j-1]} \right\},$$
$$\mathcal{K}_{\mathbf{H}}^{j} := \left\{ K_{\mathbf{H}}^{j}(\boldsymbol{z}^{\mathrm{mp}(j)} - (\cdot)) : \boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)} \right\}.$$

For simplicity, throughout the theoretical analysis, we assume that all quantities appearing in the proof satisfy sufficient measurability conditions.

## B.3.2 Main Theorem

Here, we detail the assumptions, the statement, and a proof of Theorem 3.1.

**Preliminaries.** We use the following convenient *multi-index* notation (see, e.g., [258]).

**Definition B.1** (Multi-index notation)**.** *For $d \in \mathbb{N}$, we call a $d$-tuple $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$* multi-index*. For a multi-index $\alpha$, let $|\alpha| := \sum_{j=1}^d \alpha_j$ and $\alpha! := \prod_{j=1}^d \alpha_j!$, and $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Also, let $\partial^\alpha$ denote the partial differential operator defined by*

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

**Definition B.2** (Convolution)**.** *Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ be a measurable subset. For continuous bounded functions $f, g : \Omega \to \mathbb{R}$, we define a function $(f \underset{[\Omega]}{*} g) : \Omega \to \mathbb{R}$ by*

$$f \underset{[\Omega]}{*} g(\boldsymbol{x}) := \int_\Omega f(\boldsymbol{x} - \boldsymbol{y}) g(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}.$$

*When $\Omega = \mathbb{R}^d$, we drop $\Omega$ from the notation and denote $f * g$.*

We define the following class of functions.

**Definition B.3** (Hölder class; [258, 275])**.** *Let $d \in \mathbb{N}$, $\beta > 1$, $L > 0$, and let $\Omega \subset \mathbb{R}^d$ be an open subset. The $(\beta, L)$-Hölder class $\Sigma(\beta, L, \Omega)$ is defined as the set of $k = \lfloor \beta \rfloor$-times continuously differentiable functions $f : \Omega \to \mathbb{R}$ satisfying*

$$|\partial^\alpha f(x) - \partial^\alpha f(x')| \le L \left\| x - x' \right\|^{\beta - k} \quad \text{for} \quad x, x' \in \Omega \text{ and } |\alpha| = k,$$

*where $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index, and $\lfloor a \rfloor = \max\{z \in \mathbb{Z} : z \le a\}$ for $a \in \mathbb{R}$. When $\Omega = \mathbb{R}^d$, we also drop $\mathbb{R}^d$ from the notation and denote $\Sigma(\beta, L)$ when the dimension is clear from the context.*

*Remark* B.1. In the 1-dimensional case, a related analysis based on the notion of the Hölder class is presented in Section 1.2.3 of Tsybakov [275].

For function classes, we quantify their complexities using the Rademacher complexity.

**Definition B.4** (Rademacher complexity)**.** *Let $q$ denote a probability distribution on some measurable space $\mathcal{X}$. For a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, define*

$$\mathrm{Rad}_{m,q}(\mathcal{F}) := \mathbb{E}_q \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i f(X_i) \right| \right]$$

*where $m \in \mathbb{N}$, $\{\sigma_i\}_{i=1}^m$ are independent uniform $\{\pm 1\}$-valued random variables, and $\{X_i\}_{i=1}^m \overset{i.i.d.}{\sim} q$.*

**Assumptions.** For simplicity, throughout this theoretical analysis, we assume that all quantities appearing in the proof satisfy sufficient measurability conditions.

**Assumption B.1** (Boundedness assumptions)**.** *We assume that the following hold:*

- *The loss function is bounded, i.e., $B_\ell := \sup_{h \in \mathcal{H}} \sup_{\boldsymbol{Z} \in \mathbb{R}^d} |\ell(h, \boldsymbol{Z})| < \infty$.*

- $\mathbf{K} := \{K^j\}_{j=1}^d$ *are uniformly bounded from above, i.e.,* $B_{\mathbf{K}} := \max\{\left\|K^j\right\|_\infty : j \in [d]\} < \infty$.
- *For each* $j \in [d]$, $\boldsymbol{\mathcal{Z}}^j \subset \mathbb{R}$ *is a compact subset. Let* $B_j := \int_{\boldsymbol{\mathcal{Z}}^j} \mathrm{d}\boldsymbol{z}^j < \infty$.
- *For all* $j \in [d]$, $p_{\mathrm{mp}(j)}$ *is bounded away from zero over* $\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$. *Define* $\epsilon_{\mathrm{mp}(j)} := \inf_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})$.
- *For each* $j \in [d]$, $K^j$ *is continuous and strictly positive.*

*We define*

$$\begin{aligned}\phi_{K^j,\mathbf{H}_j} &:= \sup_{\substack{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}, \\ \boldsymbol{z}^{\mathrm{mp}(j)\prime} \in \mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}} \left|K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime})\right| \\ &= \sup_{\substack{\boldsymbol{z}^{\mathrm{mp}(j)} \in \mathbf{H}_j^{-1}\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}, \\ \boldsymbol{z}^{\mathrm{mp}(j)\prime} \in \mathbf{H}_j^{-1}(\mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)})}} \left|K^j(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime})\right| |\det \mathbf{H}_j|^{-1}\end{aligned}$$

*and assume* $\phi_{K^j,\mathbf{H}_j} < \infty$.

*Remark* B.2. Since $\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$ is compact and $K^j$ is continuous, if we define

$$\epsilon_{K^j}(\mathbf{H}_j) := |\det \mathbf{H}_j| \left(\inf_{\boldsymbol{x},\boldsymbol{x}' \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{x}')\right) = \inf_{\boldsymbol{x},\boldsymbol{x}' \in \mathbf{H}_j^{-1}\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j(\boldsymbol{x} - \boldsymbol{x}'),$$

this quantity is strictly positive under Assumption B.1.

From here, we fix $\beta > 1$ and $L > 0$.

**Assumption B.2** (Smoothness assumptions)**.** *We assume that the following hold for all* $j \in [d]$*:*

- $p_{\mathrm{mp}(j)}$ *has an extension* $\check{p}_{\mathrm{mp}(j)} \in \Sigma(\beta, L)$ *such that*
$$\check{I}_{\mathrm{mp}(j)} := \int_{\mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} |\check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})| \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)} < \infty.$$
- *For all* $\boldsymbol{z}^j \in \boldsymbol{\mathcal{Z}}^j$, $p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot)$ *has an extension* $\check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot) \in \Sigma(\beta, L)$ *such that* $\check{I}_{j,\mathrm{mp}(j)} := \int_{\boldsymbol{\mathcal{Z}}^j} \left(\int_{\mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} |\check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)})| \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)}\right) \mathrm{d}\boldsymbol{z}^j < \infty$.
- $K^j$ *is of order* $k = \lfloor \beta \rfloor$*, i.e.,*
$$\int_{\mathbb{R}^{|\mathrm{mp}(j)|}} K^j(u)\mathrm{d}u = 1, \qquad \int_{\mathbb{R}^{|\mathrm{mp}(j)|}} K^j(u)u^\alpha \mathrm{d}u = 0 \quad (1 \le |\alpha| \le k),$$

*where* $\alpha \in \mathbb{N}_0^{|\mathrm{mp}(j)|}$ *is a multi-index, and* $K^j$ *satisfies* $\int_{\mathbb{R}^{|\mathrm{mp}(j)|}} |K^j(u)| \cdot \|u\|^\beta \mathrm{d}u < \infty$.

*Remark* B.3 (Existence of the smooth extensions). The smooth extensions in Assumption B.2 exist, for example, if we consider a smooth density function $\check{p}_{\mathrm{mp}(j)}$ on $\mathbb{R}^{|\mathrm{mp}(j)|}$ and regard its restriction to $\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$ with appropriate scaling as $p_{\mathrm{mp}(j)}$.

**Statement.** We prove the following theorem. Theorem 3.1 is obtained by changing $\delta$ to $\frac{\delta}{2d}$ in the following theorem, substituting $\|\mathbf{H}_j\|_{\mathrm{op}} = \max_{j' \in \mathrm{mp}(j)} \boldsymbol{h}^{j'}$, and defining the appropriate constants.

**Theorem B.1** (Excess risk bound)**.** *Assume that Assumptions B.1 and B.2 hold. Let* $n \in \mathbb{N}$. *For* $j \in [d]$*, define*

$$C_{\mathbf{H}} := B_\ell \sum_{j=1}^d \frac{1}{\epsilon_{\mathrm{mp}(j)}} \left(B_j + \frac{B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}\right) \Phi(\beta, L, K^j) \|\mathbf{H}_j\|_{\mathrm{op}}^\beta,$$

$$C_p := B_\ell \sum_{j=1}^d \frac{\phi_{K^j,\mathbf{H}_j}}{\epsilon_{\mathrm{mp}(j)}} \left(\check{I}_{j,\mathrm{mp}(j)} + \frac{B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \check{I}_{\mathrm{mp}(j)}\right),$$

$$C_{\mathbf{K}} := \max_{j \in [d]} \left\{\frac{1}{\epsilon_{K^j}(\mathbf{H}_j)}, \frac{B_{\mathbf{K}}}{(\epsilon_{K^j}(\mathbf{H}_j))^2}\right\}, \quad R_{\mathcal{H},\mathbf{K}} := \sum_{j=1}^d |\det \mathbf{H}_j| \operatorname{Rad}_{n,p}\left(\mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}}\right),$$

$$R_{\mathbf{K}} := \sum_{j=1}^d |\det \mathbf{H}_j| \operatorname{Rad}_{n,p}(\mathcal{K}^j_{\mathbf{H}}).$$

*Then, for any $\delta \in (0,1)$, we have with probability at least $1-2d\delta$,*

$$\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \leq 2(C_{\mathbf{H}} + C_p) + 4C_{\mathbf{K}}(R_{\mathcal{H},\mathbf{K}} + B_\ell R_{\mathbf{K}}) + 2DB_\ell B_{\mathbf{K}} C_{\mathbf{K}} \sqrt{\frac{\log(2/\delta)}{2n}}.$$

**Proof.** Our proof derives ideas from the literature on *local empirical processes* and *kernel-type estimators*, namely Einmahl and Mason [69, 70] and Dony et al. [64]. Two elementary calculations are essential in the proof. The first one handles a difference between two products: let $N \in \mathbb{N}$, $(a_1, \ldots, a_N) \in \mathbb{R}^N$, and $(b_1, \ldots, b_N) \in \mathbb{R}^N$, then,

$$\left(\prod_{j=1}^{N} a_i\right) - \left(\prod_{j=1}^{N} b_i\right) = \sum_{j=1}^{N} a_1 \cdots a_{j-1}(a_j - b_j) b_{j+1} \cdots b_N. \tag{B.1}$$

The second one bounds a difference between two ratios from above: for $A, B, C, D \in \mathbb{R}$ with $B, D \neq 0$,

$$\left|\frac{A}{B} - \frac{C}{D}\right| = \left|\frac{A}{B} - \frac{C}{B} + \frac{C}{B} - \frac{C}{D}\right| \leq \left|\frac{1}{B}\right| \cdot |A - C| + \left|\frac{C}{BD}\right| \cdot |B - D|. \tag{B.2}$$

*Proof of Theorem B.1.* First, note

$$\begin{aligned}
\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) &= \mathcal{R}(\hat{h}) - \hat{\mathcal{R}}_{\text{aug}}(\hat{h}) + \hat{\mathcal{R}}_{\text{aug}}(\hat{h}) - \mathcal{R}(h^*) \\
&\leq \mathcal{R}(\hat{h}) - \hat{\mathcal{R}}_{\text{aug}}(\hat{h}) + \hat{\mathcal{R}}_{\text{aug}}(h^*) - \mathcal{R}(h^*) \leq \underbrace{2 \sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \hat{\mathcal{R}}_{\text{aug}}(h)\right|}_{(*)}.
\end{aligned}$$

For ease of notation, define $\hat{p}_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) = \sum_{i=1}^{n} \delta_{Z_i^j}(\boldsymbol{z}^j) \hat{w}_i^j(\boldsymbol{z}^{\text{mp}(j)})$ and temporarily denote $p_k := p_{k|\text{mp}(k)}$. With this notation,

$$\hat{\mathcal{R}}_{\text{aug}}(h) = \int_{\mathcal{Z}} \ell_h(\boldsymbol{z}) \prod_{j=1}^{d} \hat{p}_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z}.$$

Then, applying the argument of Equation (B.1), we have

$$\begin{aligned}
(*) &= \sup_{h \in \mathcal{H}} \left| \int_{\mathcal{Z}} \ell_h(\boldsymbol{z}) \prod_{j=1}^{d} p_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z} - \int_{\mathcal{Z}} \ell_h(\boldsymbol{z}) \prod_{j=1}^{d} \hat{p}_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z} \right| \\
&= \sup_{h \in \mathcal{H}} \left| \int_{\mathcal{Z}} \ell_h(\boldsymbol{z}) \sum_{j=1}^{d} \left( \prod_{k=j+1}^{d} p_k(\boldsymbol{z}^k | \boldsymbol{z}^{\text{mp}(k)}) \right) \left(p_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) - \hat{p}_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)})\right) \left( \prod_{k=1}^{j-1} \hat{p}_k(\boldsymbol{z}^k | \boldsymbol{z}^{\text{mp}(k)}) \right) \mathrm{d}\boldsymbol{z} \right| \\
&\leq \sum_{j=1}^{d} \underbrace{\sup_{h \in \mathcal{H}} \left| \int_{\mathcal{Z}} \ell_h(\boldsymbol{z}) \left( \prod_{k=j+1}^{d} p_k(\boldsymbol{z}^k | \boldsymbol{z}^{\text{mp}(k)}) \right) \left(p_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) - \hat{p}_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)})\right) \left( \prod_{k=1}^{j-1} \hat{p}_k(\boldsymbol{z}^k | \boldsymbol{z}^{\text{mp}(k)}) \right) \mathrm{d}\boldsymbol{z} \right|}_{(*j)}.
\end{aligned}$$

Now, for $h \in \mathcal{H}$ and $j \in [d]$, we define $\ell_{h,j}^{\boldsymbol{i}_{1:j-1}} : \boldsymbol{z}^j \mapsto \ell_{h,j}(Z_{\boldsymbol{i}_{1:j-1}}, \boldsymbol{z}^j)$. Then, for each $j \in [D]$, applying Lemma B.5, we obtain

$$\begin{aligned}
&(*j) \\
&= \sup_{h \in \mathcal{H}} \left| \sum_{i_1=1}^{n} \cdots \sum_{i_{j-1}=1}^{n} \left( \int_{\boldsymbol{\mathcal{Z}}^j} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(\boldsymbol{z}^j) p_j(\boldsymbol{z}^j | \boldsymbol{Z}_{\boldsymbol{i}_{1:j-1}}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \sum_{i_j=1}^{n} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j | \boldsymbol{i}_{1:j-1}} \right) \hat{w}_{i_{j-1} | \boldsymbol{i}_{1:j-2}} \cdots \hat{w}_{i_1}^1 \right| \\
&\leq 1 \cdot \left( \sup_{h \in \mathcal{H}} \max_{\boldsymbol{i}_{1:j-1} \in [n]^{j-1}} \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(\boldsymbol{z}^j) p_j(\boldsymbol{z}^j | \boldsymbol{Z}_{\boldsymbol{i}_{1:j-1}}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \sum_{i_j=1}^{n} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j}^j(Z_{\boldsymbol{i}_{1:j-1}}^{\text{mp}(j)}) \right| \right) \\
&\leq \max_{\boldsymbol{i}_{1:j-1} \in [n]^{j-1}} \sup_{h \in \mathcal{H}} \sup_{\boldsymbol{z}^{\text{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\text{mp}(j)}} \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(\boldsymbol{z}^j) p_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \sum_{i_j=1}^{n} \ell_{h,j}^{\boldsymbol{i}_{1:j-1}}(Z_{i_j}^j) \hat{w}_{i_j}^j(\boldsymbol{z}^{\text{mp}(j)}) \right| \\
&\leq \sup_{\ell'_{h,j} \in \mathcal{L}_{\mathcal{H}}^j} \sup_{\boldsymbol{z}^{\text{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\text{mp}(j)}} \underbrace{\left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) p_j(\boldsymbol{z}^j | \boldsymbol{z}^{\text{mp}(j)}) \mathrm{d}\boldsymbol{z}^j - \sum_{i_j=1}^{n} \ell'_{h,j}(Z_{i_j}^j) \hat{w}_{i_j}^j(\boldsymbol{z}^{\text{mp}(j)}) \right|}_{(**)},
\end{aligned}$$

where we used that $\left\{\boldsymbol{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}}\right\}_{\boldsymbol{i}_{1:j-1}\in[n]^{j-1}} \subset \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$ that follows from $\left\{\boldsymbol{Z}^{\mathrm{mp}(j)}_{i}\right\}_{i=1}^{n} \subset \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$. Define

$$
\begin{aligned}
r^j(f, \boldsymbol{z}^{\mathrm{mp}(j)}) &:= \int_{\boldsymbol{\mathcal{Z}}^j} f(\boldsymbol{z}^j) p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{z}^j, \\
\hat{r}^j(f, \boldsymbol{z}^{\mathrm{mp}(j)}) &:= \frac{1}{n}\sum_{i=1}^{n} f(Z_i^j) K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}^{\mathrm{mp}(j)}_i), \\
g^j(\boldsymbol{z}^{\mathrm{mp}(j)}) &:= p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}), \\
\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)}) &:= \frac{1}{n}\sum_{i=1}^{n} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}^{\mathrm{mp}(j)}_i).
\end{aligned}
$$

Then, for each $\ell'_{h,j} \in \mathcal{L}^j_{\mathcal{H}}$ and $\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$,

$$
\begin{aligned}
(**) &= \left| \frac{r^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} - \frac{\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \\
&\leq \underbrace{\left| \frac{r^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} - \frac{\mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right|}_{\rho_1} + \underbrace{\left| \frac{\mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} - \frac{\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right|}_{\rho_2}.
\end{aligned}
$$

By applying the argument of Equation (B.2), we can bound each ratio difference term as

$$
\begin{aligned}
\rho_1 &\leq \left| \tfrac{1}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \cdot |r^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| + \left| \tfrac{\mathbb{E}\hat{r}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \cdot |g^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \\
\rho_2 &\leq \left| \tfrac{1}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \cdot |\mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| + \left| \tfrac{\hat{r}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \cdot |\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|.
\end{aligned}
$$

Applying Lemma B.1 to the coefficients, Lemma B.2 to the deterministic difference terms bounding $\rho_1$, Lemma B.3 to the stochastic difference terms bounding $\rho_2$ along with the union bound, for any $\delta \in (0,1)$, we have with probability at least $1 - 2D\delta$,

$$
\begin{aligned}
&\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \\
&\leq 2\sum_{j=1}^{d} \Bigg( \frac{1}{\epsilon_{\mathrm{mp}(j)}} \left( B_\ell B_j \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|^{\beta}_{\mathrm{op}} + B_\ell \phi_{K^j,\mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \right) \\
&\quad + \frac{1}{\epsilon_{\mathrm{mp}(j)}} \cdot \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \left( \Phi(\beta, L, K^j) \left\|\mathbf{H}_j\right\|^{\beta}_{\mathrm{op}} + \phi_{K^j,\mathbf{H}_j} \check{I}_{\mathrm{mp}(j)} \right) \\
&\quad + \frac{|\det \mathbf{H}_j|}{\epsilon_{K^j}(\mathbf{H}_j)} \left( 2\mathrm{Rad}_{n,p}\left( \mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}} \right) + \frac{B_\ell B_{\mathbf{K}}}{|\det \mathbf{H}_j|} \sqrt{\frac{\log(2/\delta)}{2n}} \right) \\
&\quad + \frac{|\det \mathbf{H}_j|}{\epsilon_{K^j}(\mathbf{H}_j)} \cdot \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)} \left( 2\mathrm{Rad}_{n,p}(\mathcal{K}^j_{\mathbf{H}}) + \frac{B_{\mathbf{K}}}{|\det \mathbf{H}_j|} \sqrt{\frac{\log(2/\delta)}{2n}} \right) \Bigg).
\end{aligned}
$$

By reorganizing the terms, we obtain the assertion. □

**Lemmas.** Here, we prove the lemmas used in the proof of Theorem B.1.

**Lemma B.1** (Bounded coefficients)**.** *Assume Assumption B.1 holds. Let $j \in [d]$. Then,*

$$
\begin{aligned}
&\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \frac{1}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{1}{\epsilon_{\mathrm{mp}(j)}}, \qquad
&&\sup_{\ell'_{h,j} \in \mathcal{L}^j_{\mathcal{H}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \frac{\mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}, \\
&\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \frac{1}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{|\det \mathbf{H}_j|}{\epsilon_{K^j}(\mathbf{H}_j)}, \qquad
&&\sup_{\ell'_{h,j} \in \mathcal{L}^j_{\mathcal{H}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \frac{\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})} \right| \leq \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.
\end{aligned}
$$

*Proof.* By Assumption B.1, we have

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{1}{g^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| = \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)})} \le \frac{1}{\epsilon_{\mathrm{mp}(j)}}.$$

Also,

$$\begin{aligned}
&\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{1}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| \\
&\le \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|} \\
&= \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)}-\boldsymbol{z}^{\mathrm{mp}(j)\prime})g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime})\mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime}\right|} \\
&= \frac{1}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)}-\boldsymbol{z}^{\mathrm{mp}(j)\prime})g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime})\mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime}} \\
&\le \frac{1}{|\det\mathbf{H}_j|^{-1}\epsilon_{K^j}(\mathbf{H}_j)\int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} g^j(\boldsymbol{z}^{\mathrm{mp}(j)\prime})\mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime}} = \frac{|\det\mathbf{H}_j|}{\epsilon_{K^j}(\mathbf{H}_j)},
\end{aligned}$$

where we used the positivity of the integrand. Now,

$$\begin{aligned}
&\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{\mathbb{E}\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})}{\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| \\
&= \sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{|\det\mathbf{H}_j|\,\mathbb{E}\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})}{|\det\mathbf{H}_j|\,\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| \\
&\le \frac{\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left\|\ell'_{h,j}\right\|_\infty\cdot\left\|\left(|\det\mathbf{H}_j|\,K^j_{\mathbf{H}}\right)\right\|_\infty}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\det\mathbf{H}_j|\,|\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|} \le \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.
\end{aligned}$$

Similarly, we have $\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\det\mathbf{H}_j|\cdot\left|\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})\right| \ge \epsilon_{K^j}(\mathbf{H}_j)$. Therefore,

$$\begin{aligned}
&\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})}{\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| \\
&= \sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}\left|\frac{|\det\mathbf{H}_j|\,\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})}{|\det\mathbf{H}_j|\,\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})}\right| \\
&\le \frac{\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\det\mathbf{H}_j|\cdot\left|\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})\right|}{\inf_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\det\mathbf{H}_j|\cdot|\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|} \le \frac{B_\ell B_{\mathbf{K}}}{\epsilon_{K^j}(\mathbf{H}_j)}.
\end{aligned}$$

□

**Lemma B.2** (Deterministic terms)**.** *Assume that Assumptions B.1 and B.2 hold. Let $j\in[d]$. Then,*

$$\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|r^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})-\mathbb{E}\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})| \le B_\ell B_j\Phi(\beta,L,K^j)\left\|\mathbf{H}_j\right\|_{\mathrm{op}}^\beta + B_\ell\phi_{K^j,\mathbf{H}_j}\bar{I}_{j,\mathrm{mp}(j)},$$

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|g^j(\boldsymbol{z}^{\mathrm{mp}(j)})-\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \le \Phi(\beta,L,K^j)\left\|\mathbf{H}_j\right\|_{\mathrm{op}}^\beta + \phi_{K^j,\mathbf{H}_j}\bar{I}_{\mathrm{mp}(j)}.$$

*Proof.* By applying Lemma B.4 under Assumption B.2,

$$
\begin{aligned}
&\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} |g^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \\
&= \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) p_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&= \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&\leq \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)}) - \left( K^j_{\mathbf{H}} * \check{p}_{\mathrm{mp}(j)} \right)(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \\
&\quad + \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \int_{\mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{\mathrm{mp}(j)}(\boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right| \\
&\leq \Phi(\beta, L, K^j) \left\| \mathbf{H}_j \right\|_{\mathrm{op}}^{\beta} + \phi_{K^j, \mathbf{H}_j} \check{I}_{\mathrm{mp}(j)}.
\end{aligned}
$$

Similarly, for each $\ell'_{h,j} \in \mathcal{L}^j_{\mathcal{H}}$ and $\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$,

$$
\begin{aligned}
&|r^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| \\
&= \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{z}^j \right. \\
&\qquad \left. - \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) \left( \int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) p_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&= \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) \mathrm{d}\boldsymbol{z}^j \right. \\
&\qquad \left. - \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) \left( \int_{\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\leq \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) \left( \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\quad + \left| \int_{\boldsymbol{\mathcal{Z}}^j} \ell'_{h,j}(\boldsymbol{z}^j) \left( \int_{\mathbb{R}^{|\mathrm{mp}(j)|} \setminus \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)\prime}) \mathrm{d}\boldsymbol{z}^{\mathrm{mp}(j)\prime} \right) \mathrm{d}\boldsymbol{z}^j \right| \\
&\leq B_\ell \int_{\boldsymbol{\mathcal{Z}}^j} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \mathrm{d}\boldsymbol{z}^j + B_\ell \phi_{K^j, \mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \\
&\leq B_\ell B_j \sup_{\boldsymbol{z}^j \in \boldsymbol{\mathcal{Z}}^j} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| + B_\ell \phi_{K^j, \mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)} \\
&\leq B_\ell B_j \sup_{\boldsymbol{z}^j \in \boldsymbol{\mathcal{Z}}^j} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \\
&\quad + B_\ell \phi_{K^j, \mathbf{H}_j} \check{I}_{j,\mathrm{mp}(j)}.
\end{aligned}
$$

Applying Lemma B.4 under Assumption B.2, for each $\boldsymbol{z}^j \in \boldsymbol{\mathcal{Z}}^j$, we obtain

$$
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} \left| \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \boldsymbol{z}^{\mathrm{mp}(j)}) - (K^j_{\mathbf{H}} * \check{p}_{j,\mathrm{mp}(j)}(\boldsymbol{z}^j, \cdot))(\boldsymbol{z}^{\mathrm{mp}(j)}) \right| \leq \Phi(\beta, L, K^j) \left\| \mathbf{H}_j \right\|_{\mathrm{op}}^{\beta}.
$$

Therefore, we have the assertion. □

**Lemma B.3** (Probabilistic terms)**.** *Assume that Assumption B.1 holds. Let $j \in [d]$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$
\begin{aligned}
&\sup_{\ell'_{h,j} \in \mathcal{L}^j_{\mathcal{H}}} \sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{r}^j(\ell'_{h,j}, \boldsymbol{z}^{\mathrm{mp}(j)})| \\
&\qquad \leq 2\mathrm{Rad}_{n,p}\left( \mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}} \right) + \frac{B_\ell B_{\mathbf{K}}}{|\det \mathbf{H}_j|} \sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}
$$

*Similarly, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$
\sup_{\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}} |\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)}) - \hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})| \leq 2\mathrm{Rad}_{n,p}(\mathcal{K}^j_{\mathbf{H}}) + \frac{B_{\mathbf{K}}}{|\det \mathbf{H}_j|} \sqrt{\frac{\log(2/\delta)}{2n}}.
$$

*Proof.* Note

$$\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\mathbb{E}\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})-\hat{r}^j(\ell'_{h,j},\boldsymbol{z}^{\mathrm{mp}(j)})|$$
$$=\sup_{\ell'_{h,j}\in\mathcal{L}^j_{\mathcal{H}}}\sup_{k\in\mathcal{K}^j_{\mathbf{H}}}\left|\frac{1}{n}\sum_{i=1}^n\ell'_{h,j}(Z^j_i)k(\boldsymbol{Z}^{\mathrm{mp}(j)}_i)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\ell'_{h,j}(Z^j_i)k(\boldsymbol{Z}^{\mathrm{mp}(j)}_i)\right]\right|$$

and

$$\sup_{\boldsymbol{z}^{\mathrm{mp}(j)}\in\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}}|\mathbb{E}\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})-\hat{g}^j(\boldsymbol{z}^{\mathrm{mp}(j)})|=\sup_{k\in\mathcal{K}^j_{\mathbf{H}}}\left|\frac{1}{n}\sum_{i=1}^n k(\boldsymbol{Z}^{\mathrm{mp}(j)}_i)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n k(\boldsymbol{Z}^{\mathrm{mp}(j)}_i)\right]\right|.$$

Now, applying Fact B.3 to these expressions, we obtain the assertions of the lemma. □

**Facts.** Here, we state some facts used in the proof of Theorem B.1. The following is Taylor's formula with the integral form of the remainder, stated using the multi-index notation.

**Fact B.1** (Taylor's theorem; [317], Section 8.4.4)**.** *Let $\Omega\subset\mathbb{R}^n$ be an open subset. Let $n\in\mathbb{N}$, and let $f:\Omega\to\mathbb{R}$ be $k$-times continuously differentiable. Then, for any $x,u\in\Omega$ such that $x+tu\in\Omega$ for all $t\in[0,1]$, the following equality holds:*

$$f(x+u)-f(x)=\sum_{1\le|\alpha|<k}\frac{\partial^\alpha f(x)}{\alpha!}u^\alpha+\sum_{|\alpha|=k}\frac{|\alpha|}{\alpha!}u^\alpha\int_0^1(1-t)^{|\alpha|-1}\partial^\alpha f(x+tu)\mathrm{d}t.$$

The following elementary inequality is easily proved by using the strict convexity and the strict monotonicity of the logarithm function.

**Fact B.2** (Weighted AM-GM inequality)**.** *Let $n\in\mathbb{N}$, $x_1,\dots,x_n\ge0$, and $w_1,\dots,w_n\ge0$. Define $w:=w_1+\cdots+w_n$ and assume $w>0$. Then,*

$$\frac{w_1x_1+\cdots+w_nx_n}{w}\ge(x_1^{w_1}\cdots x_n^{w_n})^{\frac{1}{w}}.$$

The following standard Rademacher complexity bound is essentially due to McDiarmid's inequality, which is applied twice with the union bound [184, Theorem 3.3].

**Fact B.3** (Rademacher complexity bound; Theorem 3.3 in [184])**.** *Let $B>0$ and $m\in\mathbb{N}$. Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,B]$, and let $z$ be a $\mathcal{Z}$-valued random variable. Then, for any $\delta>0$, with probability at least $1-\delta$ over the draw of an independent and identically distributed sample $\{z_i\}_{i=1}^m\overset{i.i.d.}{\sim}z$, the following holds:*

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^m g(z_i)-\mathbb{E}[g(z)]\right|\le2\mathrm{Rad}_{m,p}(\mathcal{G})+B\sqrt{\frac{\log(2/\delta)}{2m}}.$$

**Basic Lemmas.** Here, we prove the basic lemmas used in the proof of Theorem B.1.

**Lemma B.4** (Convolution error bound for Hölder class)**.** *Let $d\in\mathbb{N}$, $\beta>1$, and $L>0$. Assume that the kernel function $K:\mathbb{R}^d\to\mathbb{R}$ is of order $k=\lfloor\beta\rfloor$ and satisfies*

$$\int_{\mathbb{R}^d}|K(u)|\cdot\|u\|^\beta\mathrm{d}u<\infty.$$

*Let $\mathbf{H}=\mathrm{diag}(h_1,\dots,h_d)$ with $h_1,\dots,h_d>0$, and define $K_{\mathbf{H}}(u):=\frac{1}{|\det\mathbf{H}|}K(\mathbf{H}^{-1}u)$. Then, for any $f\in\Sigma(\beta,L)$, the following holds:*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d}|f(\boldsymbol{x})-(K_{\mathbf{H}}*f)(\boldsymbol{x})|\le\Phi(\beta,L,K)\|\mathbf{H}\|_{\mathrm{op}}^\beta,$$

*where* $\Phi(\beta, L, K)$ *is defined as*

$$\Phi(\beta, L, K) := L\left(\int_0^1 (1-t)^{k-1}t^{\beta-k}\mathrm{d}t\right)\sum_{|\alpha|=k}\frac{\|\alpha\|^k}{\alpha!k^{k-1}}\int_{\mathbb{R}^d}|K(u)|\cdot\|u\|^\beta \mathrm{d}u$$

*and* $\alpha \in \mathbb{N}_0^d$ *runs over multi-indices.*

*Proof.* First, we fix $x \in \mathbb{R}^d$. We apply the change of variables formula and obtain

$$|f(x) - (K_{\mathbf{H}} * f)(x)| = \left|f(x) - \int_{\mathbb{R}^d} K(u)f(x-\mathbf{H}u)\mathrm{d}u\right|. \tag{*}$$

We apply Fact B.1 to obtain

$$\begin{aligned}
(*) &= \left|f(x) - \int_{\mathbb{R}^d} K(u)\left(f(x) + \sum_{1\le|\alpha|<k}\frac{\partial^\alpha f(x)}{\alpha!}(-\mathbf{H}u)^\alpha\right.\right.\\
&\qquad\left.\left. + \sum_{|\alpha|=k}\frac{|\alpha|}{\alpha!}(-\mathbf{H}u)^\alpha\int_0^1(1-t)^{|\alpha|-1}\partial^\alpha f(x+t(-\mathbf{H}u))\mathrm{d}t\right)\mathrm{d}u\right|\\
&= \left|\int_{\mathbb{R}^d} K(u)\left(\sum_{|\alpha|=k}\frac{|\alpha|}{\alpha!}(-\mathbf{H}u)^\alpha\int_0^1(1-t)^{|\alpha|-1}\partial^\alpha f(x-t\mathbf{H}u)\mathrm{d}t\right)\mathrm{d}u\right|\\
&= \left|\int_{\mathbb{R}^d} K(u)\left(\sum_{|\alpha|=k}\frac{|\alpha|}{\alpha!}(-\mathbf{H}u)^\alpha\int_0^1(1-t)^{|\alpha|-1}(\partial^\alpha f(x-t\mathbf{H}u)-\partial^\alpha f(x))\mathrm{d}t\right)\mathrm{d}u\right|\\
&\le \int_{\mathbb{R}^d}|K(u)|\left(\sum_{|\alpha|=k}\frac{|\alpha|}{\alpha!}|\mathbf{H}u|^\alpha\int_0^1(1-t)^{|\alpha|-1}|\partial^\alpha f(x-t\mathbf{H}u)-\partial^\alpha f(x)|\mathrm{d}t\right)\mathrm{d}u, \qquad (**)
\end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index and $|\mathbf{H}u|^\alpha := |h_1u_1|^{\alpha_1}\cdots|h_du_d|^{\alpha_d}$. Now, by the Hölder-condition of $\partial^\alpha f$, we have $|\partial^\alpha f(x-t\mathbf{H}u) - \partial^\alpha f(x)| \le L\|t\mathbf{H}u\|^{\beta-k}$. Also, by applying Fact B.2, we have

$$|\mathbf{H}u|^\alpha = |h_1u_1|^{\alpha_1}\cdots|h_du_d|^{\alpha_d} \le \left(\frac{1}{|\alpha|}\sum_{j=1}^d\alpha_j|h_ju_j|\right)^{|\alpha|} \le \left(\frac{1}{|\alpha|}\|\alpha\|\cdot\|hu\|\right)^{|\alpha|} = \frac{\|\alpha\|^k}{k^k}\|hu\|^k.$$

By applying these inequalities and imputing $|\alpha| = k$, we obtain

$$\begin{aligned}
(**) &\le \int_{\mathbb{R}^d}|K(u)|\left(\sum_{|\alpha|=k}\frac{\|\alpha\|^k}{\alpha!k^{k-1}}\|\mathbf{H}u\|^k\int_0^1(1-t)^{k-1}L\|t\mathbf{H}u\|^{\beta-k}\,\mathrm{d}t\right)\mathrm{d}u\\
&= L\left(\int_0^1(1-t)^{k-1}t^{\beta-k}\mathrm{d}t\right)\sum_{|\alpha|=k}\frac{\|\alpha\|^k}{\alpha!k^{k-1}}\int_{\mathbb{R}^d}|K(u)|\cdot\|\mathbf{H}u\|^\beta\,\mathrm{d}u.
\end{aligned}$$

Finally, applying $\|\mathbf{H}u\| \le \|\mathbf{H}\|_{\mathrm{op}}\|u\|$, we have the assertion. □

**Lemma B.5** (Bounded weights)**.** *For all* $j \in [d]$*,*

$$\sum_{i_1=1}^n\cdots\sum_{i_j=1}^n\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}\cdots\hat{w}^1_{i_1} \in \{0,1\}.$$

*Proof.* By direct computation, we have for any $\boldsymbol{z}^{\mathrm{mp}(j)} \in \boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}$,

$$\begin{aligned}
\sum_{i=1}^n\hat{w}_i^j(\boldsymbol{z}^{\mathrm{mp}(j)}) &= \begin{cases}\sum_{i=1}^n\frac{1}{n} & \text{if } \mathrm{mp}(j)=\emptyset,\\ \sum_{i=1}^n 0 & \text{if } K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)} - \boldsymbol{Z}_i^{\mathrm{mp}(j)}) = 0, \forall i,\\ \sum_{i=1}^n\frac{K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)}-\boldsymbol{Z}_i^{\mathrm{mp}(j)})}{\sum_{i=1}^n K^j_{\mathbf{H}}(\boldsymbol{z}^{\mathrm{mp}(j)}-\boldsymbol{Z}_i^{\mathrm{mp}(j)})} & \text{otherwise,}\end{cases}\\
&\in \{0,1\}.
\end{aligned}$$

For $j = 1$, since $\mathrm{mp}(1) = \emptyset$, we can directly show the assertion as

$$\sum_{i_1=1}^{n} \hat{w}_{i_1}^1 = \sum_{i_1=1}^{n} \frac{1}{n} = 1.$$

For $j \geq 2$,

$$\begin{aligned}\sum_{i_1=1}^{n} \cdots \sum_{i_j=1}^{n} \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \cdots \hat{w}_{i_1}^1 &= \sum_{i_1=1}^{n} \cdots \sum_{i_{j-1}=1}^{n} \hat{w}_{i_{j-1}|\boldsymbol{i}_{1:j-2}} \cdots \hat{w}_{i_1}^1 \left( \sum_{i_j=1}^{n} \hat{w}_{i_j|\boldsymbol{i}_{1:j-1}} \right) \\ &\in \left\{ 0, \left( \sum_{i_1=1}^{n} \cdots \sum_{i_{j-1}=1}^{n} \hat{w}_{i_{j-1}|\boldsymbol{i}_{1:j-2}} \cdots \hat{w}_{i_1}^1 \right) \right\}.\end{aligned}$$

By recursively applying the above argument for a finite number of times, we obtain the assertion for all $j \in [d]$. □

### B.3.3 Comparison of Complexity Measures

Here, we formally demonstrate the complexity reduction effect explained in Section 3.4. More concretely, as an example in which the effect can be demonstrated, we take the example represented by Assumption B.3 where the Lipschitz continuity of the functions are assumed and compare the upper bounds on the complexity terms appearing in the generalization error bound of the usual empirical risk minimization (ERM) and those in Theorem 3.1 (namely $R_{\mathcal{H},\mathbf{K}}$ and $R_{\mathbf{K}}$).

The complexity reduction effect in this example is demonstrated by the different dependencies of the upper bounds on the sample size, both derived based on the metric-entropy method; the one corresponding to ERM yields a bound of order $O(n^{-1/(2+D)})$ whereas the one for the proposed method yields $O(n^{-1/3})$. Although the comparison between the two upper bounds only provides circumstantial evidence, we believe that the reduced exponent demonstrates the complexity reduction effect as they are derived based on the same proof technique.

First, recall that the proposed method enjoys Theorem B.1 which states, for any $\delta \in (0, 1)$, we have with probability at least $1 - 2d\delta$,

$$\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \leq 2(C_{\mathbf{H}} + C_p) + \underbrace{4C_{\mathbf{K}}(R_{\mathcal{H},\mathbf{K}} + B_\ell R_{\mathbf{K}})}_{\text{Complexity terms}} + 2DB_\ell B_{\mathbf{K}} C_{\mathbf{K}} \sqrt{\frac{\log(2/\delta)}{2n}}.$$

On the other hand, the usual empirical risk minimization algorithm enjoys the following theoretical guarantee. Recall $\hat{\mathcal{R}}_{\mathrm{emp}}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h, \boldsymbol{Z}_i)$.

**Proposition B.1.** *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have that the solution to the usual empirical risk minimization*

$$\hat{f}_{\mathrm{emp}} \in \arg\min_{h \in \mathcal{H}} \{\hat{\mathcal{R}}_{\mathrm{emp}}(h)\}$$

*satisfies*

$$\mathcal{R}(\hat{f}_{\mathrm{emp}}) - \mathcal{R}(h^*) \leq \underbrace{4\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{H}})}_{\textit{Complexity term}} + 2B_\ell \sqrt{\frac{\log(2/\delta)}{2n}}.$$

*Proof.* The assertion is immediate from Fact B.3 and the following inequality:

$$\begin{aligned}\mathcal{R}(\hat{f}_{\mathrm{emp}}) - \mathcal{R}(h^*) &= \mathcal{R}(\hat{f}_{\mathrm{emp}}) - \hat{\mathcal{R}}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) + \hat{\mathcal{R}}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) - \mathcal{R}(h^*) \\ &\leq \mathcal{R}(\hat{f}_{\mathrm{emp}}) - \hat{\mathcal{R}}_{\mathrm{emp}}(\hat{f}_{\mathrm{emp}}) + \hat{\mathcal{R}}_{\mathrm{emp}}(h^*) - \mathcal{R}(h^*) \\ &\leq 2 \sup_{h \in \mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}_{\mathrm{emp}}(h)|.\end{aligned}$$

□

From here, we compare the dependency of the complexity terms $\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{H}})$ and $R_{\mathcal{H},\mathbf{K}} + B_\ell R_{\mathbf{K}}$ on $n$. In addition to Assumptions B.1 and B.2, assume the following:

**Assumption B.3** (Complexity assumptions)**.** *We assume the following:*

- *The functions in $\mathcal{L}_{\mathcal{H}}$ are $L_1$-Lipschitz continuous.*
- *The functions $K^j$ are $L_{K,j}$-Lipschitz continuous.*
- *The functions $p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|\cdot)$ are $L_{p,k}$-Lipschitz continuous for all $\boldsymbol{z}^k$.*

*For simplicity, we also assume* $\mathbf{H} = \mathrm{diag}((h, \ldots, h))$.

Under this assumption, we have the following:

**Proposition B.2** (Comparison of the complexity measures)**.** *Given Assumptions B.1, B.2, and B.3, we have the following:*

$$\mathrm{Rad}_{n,p}(\mathcal{L}_{\mathcal{H}}) \leq \mathcal{O}\left(n^{-\frac{1}{d+2}}\right), \qquad R_{\mathcal{H},\mathbf{K}} + B_\ell R_{\mathbf{K}} \leq \mathcal{O}\left(n^{-1/3}\right).$$

**Implications.** Proposition B.2 shows that the complexity terms appearing in Theorem B.1 has a better dependency on the sample size compared to those in Proposition B.1, demonstrating the complexity reduction effect in this example. Note here that we do not claim that the proposed method yields a rate-optimal predictor, but instead, we provide Theorem 3.1 and this supplementary analysis to obtain insights regarding how the proposed method may facilitate the learning.

*Proof of Proposition B.2.* By the Lipschitz continuity of the functions in $\mathcal{L}_{\mathcal{H}}$ and the boundedness of $\mathcal{Z}$, we can apply Fact B.6 to obtain

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{L}_{\mathcal{H}}) \leq C\left(\frac{L_1}{t}\right)^d$$

for a constant $C > 0$. By applying Fact B.4, and minimizing the right-hand side for $t$, we have the first assertion.

On the other hand, by Lemma B.6,

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}}) \leq \log \mathcal{N}_{(t_1,\|\cdot\|_\infty)}(\mathcal{L}^j_{\mathcal{H}}) + \log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}),$$

where $t_1, t_2$ are such that $B_{\mathbf{K}} t_1 + B_\ell t_2 = t$. Now, applying Lemma B.8,

$$\log \mathcal{N}_{(t_1,\|\cdot\|_\infty)}(\mathcal{L}^j_{\mathcal{H}}) \leq \log \sup_{z \in \boldsymbol{\mathcal{Z}}^{j-1}} \mathcal{N}_{(t_{1,1},\|\cdot\|_\infty)}(\mathcal{F}_z) + \log \mathcal{N}_{(t_{1,2},\|\cdot\|)}(\mathcal{B}^{j-1}(R_{\boldsymbol{\mathcal{Z}}}))$$

By combining Lemma B.7 and Lemma B.9, and applying Fact B.5, we have

$$\log \sup_{z \in \boldsymbol{\mathcal{Z}}^{j-1}} \mathcal{N}_{(t_{1,1},\|\cdot\|_\infty)}(\mathcal{F}_z) \leq C\frac{L_2}{t_{1,1}},$$
$$\log \mathcal{N}_{(t_{1,2},\|\cdot\|)}(\mathcal{B}^{j-1}(R_{\boldsymbol{\mathcal{Z}}})) \leq (j-1)\log\left(1 + \frac{2R_{\boldsymbol{\mathcal{Z}}}}{t_{1,2}}\right),$$

where $t_{1,1}, t_{1,2}$ are such that $t_1 = t_{1,1} + L_2 t_{1,2}$, and $L_2 = L_1 + B_\ell \sum_k L_{p,k}$.

On the other hand, by Lemma B.10, we have

$$\log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}) \leq |\mathrm{mp}(j)| \log\left(1 + \frac{2L_{K,H,j}R_{\boldsymbol{\mathcal{Z}}}}{t_2}\right).$$

where $L_{K,H,j} = h^{-|\mathrm{mp}(j)|-1} L_{K,j}$. Therefore, we have

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}}) \leq C\frac{L_2}{t_{1,1}} + (j-1)\log\left(1 + \frac{2R_{\boldsymbol{\mathcal{Z}}}}{t_{1,2}}\right) + |\mathrm{mp}(j)| \log\left(1 + \frac{2L_{K,H,j}R_{\boldsymbol{\mathcal{Z}}}}{t_2}\right).$$

By applying Fact B.4, letting

$$t_{1,1} = \frac{t}{3B_{\mathbf{K}}},\ t_{1,2} = \frac{t}{3B_{\mathbf{K}}L_2},\ t_2 = \frac{t}{3B_\ell},$$

and minimizing the upper bound for $t$, we have

$$|\det \mathbf{H}_j| \operatorname{Rad}_{n,p}\left(\mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}}\right) \leq \mathcal{O}\left(n^{-1/3}\right).$$

Therefore, we have

$$R_{\mathcal{H},\mathbf{K}} = \sum_{j=1}^{d} |\det \mathbf{H}_j| \operatorname{Rad}_{n,p}\left(\mathcal{L}^j_{\mathcal{H}} \otimes \mathcal{K}^j_{\mathbf{H}}\right) \leq \mathcal{O}\left(n^{-1/3}\right),$$
$$R_{\mathbf{K}} = \sum_{j=1}^{d} |\det \mathbf{H}_j| \operatorname{Rad}_{n,p}(\mathcal{K}^j_{\mathbf{H}}) \leq \mathcal{O}\left(n^{-1/2}\right),$$

and obtain the second assertion. □

### Lemmas and Facts

**Lemma B.6** (Metric entropy of products)**.** *Let $\mathcal{F}, \mathcal{G}$ be two classes of bounded measurable functions satisfying $\|f\|_\infty \leq M_{\mathcal{F}} (f \in \mathcal{F})$ and $\|g\|_\infty \leq M_{\mathcal{G}} (g \in \mathcal{G})$. Then, we have for any $t_1, t_2 > 0$,*

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{F} \otimes \mathcal{G}) \leq \log \mathcal{N}_{(t_1,\|\cdot\|_\infty)}(\mathcal{F}) + \log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{G})$$

*where $t = M_{\mathcal{G}} t_1 + M_{\mathcal{F}} t_2$.*

*Proof.* Let $\{f_i\}_i$ ($\{g_j\}_j$) be the $t_1$- (resp. $t_2$-)covering of $\mathcal{F}$ (resp. $\mathcal{G}$). Then, for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we have for some $i, j$ that

$$\begin{aligned}
\|f \otimes g - f_i \otimes g_j\|_\infty &\leq \|f \otimes g - f_i \otimes g\|_\infty + \|f_i \otimes g - f_i \otimes g_j\|_\infty \\
&\leq \|f - f_i\|_\infty M_{\mathcal{G}} + M_{\mathcal{F}} \|g - g_j\|_\infty \\
&\leq M_{\mathcal{G}} t_1 + M_{\mathcal{F}} t_2.
\end{aligned}$$

This implies the assertion. □

**Lemma B.7** (Lipschitz continuity of marginalized function class)**.** *Assume that $p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|\cdot)$ is $L_{p,k}$-Lipschitz continuous for all $\boldsymbol{z}^k$. Then, the elements of $\bar{\mathcal{L}}^j_{\mathcal{H}}$ are Lipschitz continuous with the constant $L_1 + B_\ell \sum_k L_{p,k}$.*

*Proof.* Since the functions in $\mathcal{L}_{\mathcal{H}}$ are $L_1$-Lipschitz continuous, the elements of $\bar{\mathcal{L}}^j_{\mathcal{H}}$ are also Lipschitz continuous:

$$\begin{aligned}
&|\ell_{h,j}(x) - \ell_{h,j}(y)| \\
&= \left| \int \ell_h((x,z)) \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(x,z)^{\mathrm{mp}(k)}) \mathrm{d}\boldsymbol{z} - \int \ell_h((y,z)) \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(y,z)^{\mathrm{mp}(k)}) \mathrm{d}\boldsymbol{z} \right| \\
&\leq \int |\ell_h((x,z)) - \ell_h((y,z))| \prod_k p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(y,z)) \mathrm{d}\boldsymbol{z} \\
&\quad + \sum_{k \geq j+1} \int |\ell_h((x,z))| p_{j+1|\mathrm{mp}(j+1)}(\boldsymbol{z}^{j+1}|(x,z)) \\
&\qquad \cdots (p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^k|(x,z)) - p_{k|\mathrm{mp}(k)}(\boldsymbol{z}^{\mid}(y,z))) \cdots p_{D|\mathrm{mp}(D)}(\boldsymbol{z}^D|(y,z)) \mathrm{d}\boldsymbol{z} \\
&\leq L_1 \|x - y\| \cdot 1 + B_\ell \sum_k 1 \cdot L_{p,k} \|x - y\| \cdot 1 \\
&\leq (L_1 + B_\ell \sum_k L_{p,k}) \|x - y\|.
\end{aligned}$$

□

**Lemma B.8** (Lipschitz continuity of curried function class)**.** *Let $j \in [2:d]$ and $R_{\boldsymbol{\mathcal{Z}}} = \sup_{z \in \boldsymbol{\mathcal{Z}}} \|z\|$. Also let $\mathcal{B}^{j-1}(R)$ denote the radius-$R$ ball in the $(j-1)$-dimensional Euclidean space, and define $\mathcal{F}_z := \{\ell_{h,j}(z,\cdot) :$*

$\ell_{h,j} \in \bar{\mathcal{L}}^j_{\mathcal{H}}\}$ *for* $z \in \boldsymbol{\mathcal{Z}}^{j-1}$. *Assume* $\bar{\mathcal{L}}^j_{\mathcal{H}}$ *consist of* $L_2$*-Lipschitz continuous functions. Then, we have*

$$\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{L}^j_{\mathcal{H}}) \le \log \sup_{z \in \boldsymbol{\mathcal{Z}}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z) + \log \mathcal{N}_{(v,\|\cdot\|)}(\mathcal{B}^{j-1}(R_{\boldsymbol{\mathcal{Z}}}))$$

*where* $t, u, v > 0$ *are such that* $t = u + L_2 v$.

*Proof.* Let $\{z_\mu\}_\mu \subset \boldsymbol{\mathcal{Z}}^{j-1}$ be a $v$-covering of $\boldsymbol{\mathcal{Z}}^{j-1}$. For each $z_\mu$, consider the set $\mathcal{F}_\mu = \{\ell_{h,j}(z_\mu, \cdot) : \ell_{h,j} \in \bar{\mathcal{L}}^j_{\mathcal{H}}\}$. Let $\{\ell^{\mu,k}_{h,j}\}_k \subset \mathcal{F}_\mu$ be a $u$-covering of $\mathcal{F}_\mu$. Then, for any $\ell_{h,j} \in \bar{\mathcal{L}}^j_{\mathcal{H}}$ and $z \in \boldsymbol{\mathcal{Z}}^{j-1}$, there exists $z_\mu$ such that $\|z_\mu - z\| \le v$. Moreover, since we have $\ell_{h,j}(z_\mu, \cdot) \in \mathcal{F}_\mu$, there exists $\ell^{\mu,k}_{h,j}$ such that $\|\ell_{h,j}(z_\mu, \cdot) - \ell^{\mu,k}_{h,j}(z_\mu, \cdot)\|_\infty \le u$. For such a pair $(z_\mu, \ell^{\mu,k}_{h,j})$, we have

$$\begin{aligned}
&\|\ell_{h,j}(z, \cdot) - \ell^{\mu,k}_{h,j}(z_\mu, \cdot)\|_\infty \\
&\le \|\ell_{h,j}(z, \cdot) - \ell_{h,j}(z_\mu, \cdot)\|_\infty + \|\ell_{h,j}(z_\mu, \cdot) - \ell^{\mu,k}_{h,j}(z_\mu, \cdot)\|_\infty \le L_2 v + u
\end{aligned}$$

Therefore, the set $\bigcup_\mu \{z_\mu\}_\mu \times \{\ell^{\mu,k}_{h,j}\}_k$ induces a $(L_2 v + u)$-covering of $\mathcal{L}^j_{\mathcal{H}}$. Noting that the cardinality of $\bigcup_\mu \{z_\mu\}_\mu$ is bounded by $\mathcal{N}_{(v,\|\cdot\|)}(\mathcal{B}^{j-1}(R_{\boldsymbol{\mathcal{Z}}}))$ and that of $\{\ell^{\mu,k}_{h,j}\}_k$ by $\sup_{z \in \boldsymbol{\mathcal{Z}}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z)$, we have the assertion. □

**Lemma B.9** (Metric entropy of functions curried by a specific input)**.** *Assume that the elements of* $\bar{\mathcal{L}}^j_{\mathcal{H}}$ *are* $L_2$*-Lipschitz continuous. Then, there exists a constant* $C > 0$ *such that for sufficiently small* $u > 0$,

$$\sup_{z \in \boldsymbol{\mathcal{Z}}^{j-1}} \mathcal{N}_{(u,\|\cdot\|_\infty)}(\mathcal{F}_z) \le C \frac{L_2}{u}.$$

*Proof.* Since the elements of $\bar{\mathcal{L}}^j_{\mathcal{H}}$ are $L_2$-Lipschitz continuous, so are the elements of $\mathcal{F}_z$ with Lipschitz constant $L_2$. Indeed, for any $x, y \in \boldsymbol{\mathcal{Z}}^j$ and $z \in \boldsymbol{\mathcal{Z}}^{j-1}$, we have

$$|\ell_{h,j}(z, x) - \ell_{h,j}(z, y)| \le L_2 \left\| \begin{pmatrix} z \\ x \end{pmatrix} - \begin{pmatrix} z \\ y \end{pmatrix} \right\| = L_2 \|x - y\|.$$

Therefore, by applying Lemma B.6, we have the assertion. □

**Lemma B.10** (Shifted kernel complexity)**.** *Assume that* $K^j : \mathbb{R}^{|\mathrm{mp}(j)|} \to \mathbb{R}$ *is* $L_{K,j}$*-Lipschitz continuous. Let* $L_{K,H,j} = \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \left\|\mathbf{H}_j^{-1}\right\|_{\mathrm{op}}$. *Then, we have the following:*

$$\log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}) \le |\mathrm{mp}(j)| \log \left(1 + \frac{2 L_{K,H,j} R_{\boldsymbol{\mathcal{Z}}}}{t_2}\right).$$

*Proof.* Recalling $K^j_{\mathbf{H}}(u) = \frac{1}{|\det \mathbf{H}_j|} K^j(\mathbf{H}_j^{-1} u)$, for any $K^j_{\mathbf{H}}(z_1 - \cdot), K^j_{\mathbf{H}}(z_2 - \cdot) \in \mathcal{K}^j_{\mathbf{H}}$, we have

$$\begin{aligned}
\|K^j_{\mathbf{H}}(z_1 - \cdot) - K^j_{\mathbf{H}}(z_2 - \cdot)\|_\infty &\le \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \|\mathbf{H}_j^{-1}(z_1 - z_2)\| \\
&\le \frac{1}{|\det \mathbf{H}_j|} L_{K,j} \left\|\mathbf{H}_j^{-1}\right\|_{\mathrm{op}} \|z_1 - z_2\|
\end{aligned}$$

Therefore, we have

$$\log \mathcal{N}_{(t_2,\|\cdot\|_\infty)}(\mathcal{K}^j_{\mathbf{H}}) \le \log \mathcal{N}_{\left(t_2/L_{K,H,j},\|\cdot\|\right)}(\boldsymbol{\mathcal{Z}}^{\mathrm{mp}(j)}).$$

Applying Fact B.5, we obtain the assertion. □

**Fact B.4** (One-step discretization bound)**.** *Let* $\mathcal{F}$ *be a class of measurable functions. There exist constants* $c$ *and* $B$ *such that for any* $t \in (0, B]$, *the following relation between the Rademacher complexity and the metric entropy holds:*

$$\mathrm{Rad}_{m,q}(\mathcal{F}) \le t + c \sqrt{\frac{\log \mathcal{N}_{(t,\|\cdot\|_\infty)}(\mathcal{F})}{m}}$$

**Fact B.5** (Euclidean ball metric entropy bound; [286], Example 5.8, p.126)**.** *Let $R > 0$ and $d \in \mathbb{N}$. Let $\mathcal{B}(R)$ denote the radius-$R$ ball in the $d$-dimensional Euclidean space. Then, we have the following metric entropy bound:*

$$\log \mathcal{N}_{(\delta, \|\cdot\|)}(\mathcal{B}(R)) \leq d \log \left(1 + \frac{2R}{\delta}\right).$$

**Fact B.6** (Lipschitz functions metric entropy bound; [286], Example 5.10, p.129)**.** *Let $L, R > 0$ and $d \in \mathbb{N}$. Let $\mathrm{Lip}(R, L)$ denote the set of $L$-Lipschitz functions on $[0, R]^d$. Then, we have the following metric entropy bound for sufficiently small $\delta > 0$:*

$$\log \mathcal{N}_{(\delta, \|\cdot\|_\infty)}(\mathrm{Lip}(R, L)) \leq C \left(\frac{LR}{\delta}\right)^d,$$

*where $C > 0$ is a constant.*

## B.4 Computational Complexity of Algorithm 2

Here, we remark why the worst-case computational complexity of Algorithm 2 is $\mathcal{O}\left(n^d\right)$. The main computation cost of Algorithm 2 comes from the computation of the weights $\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}$. There are $n^{j-1}$ nodes at depth $j$ (Figure 3.3), each with $n$ weighted edges connected to depth $j+1$. The set of weights corresponding to each node, $\{\hat{w}_{i_j|\boldsymbol{i}_{1:j-1}}\}_{i_j \in [n]}$, is computed by constructing a matrix of shape $n \times n^{j-1}$ each of whose element is the kernel value for two vectors of dimensionality $|\mathrm{mp}(j)| (\leq j-1)$. In the case of Gaussian kernels, each kernel value requires $\mathcal{O}(j-1)$ operations to compute. Subsequently, the kernel matrix is normalized by the column sum, which requires $\mathcal{O}(n)$ summations and $n^j$ divisions. The same computation takes place for each of the $\boldsymbol{i}_{1:j-1} \in [n]^{j-1}$ nodes at depth $j$, therefore, the edge weights between depth $j$ and depth $j+1$ can be computed by $\mathcal{O}\left(n^j\right)$ operations. The edge weights are multiplied to obtain the node weights, which requires $\mathcal{O}\left(n^d\right)$ multiplications since the number of multiplications that take place is equal to the number of edges in Figure 3.3. Overall, Algorithm 2 requires $\mathcal{O}\left(n^d\right)$ operations for the edge weight computation and $\mathcal{O}\left(n^d\right)$ for the node weight computation, amounting to $\mathcal{O}\left(n^d\right)$ operations in total, in the worst case that no edge is pruned by the threshold $\theta$.

# Appendix C

# Appendices for Chapter 4

Table C.1 summarizes the abbreviations and the symbols used in the paper.

Table C.1: List of abbreviations and symbols used in the chapter.

| Abbreviation / Symbol | Meaning |
|---|---|
| DA | Domain adaptation |
| TA | Transfer assumption |
| SEM | Structural equation model |
| GCM | Graphical causal model |
| SCM | Structural causal model |
| IC | Independent component |
| ICA | Independent component analysis |
| GCL | Generalized contrastive learning |
| i.i.d. | Independent and identically distributed |
| $[N]$ | $\{1, 2, \ldots, N\}$ where $N \in \mathbb{N}$ |
| $\|\cdot\|_{W^{k,p}}$ | The $(k, p)$-Sobolev norm |
| $X$ | The predictor random vector ($\mathbb{R}^{d-1}$-valued) |
| $Y$ | The predicted random variable ($\mathbb{R}$-valued) |
| $\boldsymbol{Z} = (X, Y)$ | The joint random variable ($\mathbb{R}^d$-valued) |
| $S$ | The independent component vector ($\mathbb{R}^d$-valued) |
| $\mathcal{X} \subset \mathbb{R}^{d-1}$ | The space of $X$ |
| $\mathcal{Y} \subset \mathbb{R}$ | The space of $Y$ |
| $\mathcal{Z} \subset \mathbb{R}^d$ | The space of $\boldsymbol{Z} = (X, Y)$ |
| $\mathcal{H} \subset \{h : \mathbb{R}^{d-1} \to \mathbb{R}\}$ | Predictor hypothesis class |
| $\ell : \mathcal{H} \times \mathcal{Z} \to [0, B_\ell]$ | Loss function |
| $\mathcal{R}(h)$ | Target domain risk $\mathbb{E}_{p_{\mathrm{Tar}}} \ell(h, \boldsymbol{Z})$ |
| $h^* \in \mathcal{H}$ | Minimizer of target domain risk |
| $\mathcal{Q}$ | The set of independent distributions |
| $\boldsymbol{F}$ | Ground truth mixing function |
| $p_{\mathrm{Tar}}$ | The target joint distribution |
| $p_k$ | The joint distribution of source domain $k$ |
| $q_{\mathrm{Tar}} \in \mathcal{Q}$ | The target independent component (IC) distribution |
| $q_k \in \mathcal{Q}$ | The IC distribution of source domain $k$ |
| $d$ | The dimension of $\mathcal{Z}$ |
| $K$ | The number of source domains |
| $n_{\mathrm{Tar}}$ | The size of the target labeled sample |
| $n_k$ | The size of the labeled sample from source domain $k$ |
| $\mathcal{D}_{\mathrm{Tar}} = \{Z_i\}_{i=1}^{n_{\mathrm{Tar}}}$ | Target labeled data set |
| $\mathcal{D}_k = \{\boldsymbol{Z}_{k,i}^{\mathrm{Src}}\}_{i=1}^{n_k}$ | Source labeled data set of source domain $k$ |
| $\hat{\mathcal{R}}(h)$ | The ordinary empirical risk estimator |
| $\check{\mathcal{R}}(h)$ | The proposed risk estimator (Equation (4.2)) |
| $\hat{\boldsymbol{F}}$ | The estimator of $\boldsymbol{F}$ |
| $\{\psi\}_{j=1}^d$ | The penultimate layer functions composed with $\boldsymbol{F}$ during GCL |

## C.1 Nonlinear ICA

Here, we use the same notation as the main text. The recently developed nonlinear ICA provides an algorithm to estimate the mixing function $\boldsymbol{F}$. For the case of nonlinear $\boldsymbol{F}$, the impossibility of identification (i.e., consistent estimation) of $\boldsymbol{F}$ in the one-sample i.i.d. case had been established more than two decades ago [120]. However, recently, various conditions have been proposed under which $\boldsymbol{F}$ can be identified with the help of auxiliary information [123, 122, 124, 141].

The identification condition that is directly relevant to this chapter is that of the generalized contrastive learning (GCL) proposed in Hyvärinen et al. [124]. Hyvärinen et al. [124] assumes that an auxiliary variable $u_i$ from some measurable set $\mathcal{U}$ is obtained for each data point as $\{(z_i, u_i)\}_{i=1}^n$ and that the ICs $S = (S^{(1)}, \ldots, S^{(d)})$ are conditionally independent given $u$:

$$q(s|u) = \prod_{j=1}^{d} q^{(j)}(s^{(j)}|u).$$

Under such conditions, GCL estimates $\boldsymbol{F}$ by training a classification function

$$r_{\hat{\boldsymbol{F}},\boldsymbol{\psi}}(z, u) = \sum_{j=1}^{d} \psi_j(\hat{\boldsymbol{F}}^{-1}(z)_j, u) \tag{C.1}$$

parametrized by $\hat{\boldsymbol{F}}$ and $\{\psi_j\}_{j=1}^d$ with the logistic loss for classifying

$$(z, u) \text{ vs. } (z, \tilde{u}),$$

where $\tilde{u} \in \mathcal{U} \setminus \{\mathbf{u}\}$. The key condition for the identification of $\boldsymbol{F}$ is the following.

**Assumption C.1** (Assumption of variability; [124, Theorem 1])**.** *For any $z$, there exist $2d+1$ distinct points in $\mathcal{U}$, denoted by $\{u_j\}_{j=0}^{2d}$, such that the set of $(2d)$-dimensional vectors $\{w(z|u_j) - w(z|u_0)\}_{j=1}^{2d}$ are linearly independent, where*

$$w(z|u) := \left( \frac{\partial \log q^{(1)}(z_1|u)}{\partial z_1}, \ldots, \frac{\partial \log q^{(d)}(z_d|u)}{\partial z_d}, \frac{\partial^2 \log q^{(1)}(z_1|u)}{\partial z_1^2}, \ldots, \frac{\partial^2 \log q^{(d)}(z_d|u)}{\partial z_d^2} \right).$$

Under Assumption C.1 and some regularity conditions, Theorem 1 of Hyvärinen et al. [124] states that the transformation $\hat{\boldsymbol{F}}$ in Equation (C.1) trained by GCL is a consistent estimator of $\boldsymbol{F}$ upto additional dimension-wise invertible transformations. Note that the assumption is intrinsically difficult to confirm based on data due to the unsupervised nature of the problem setting. In this chapter, we use the source domain index as the auxiliary variable and employ GCL for domain adaptation. The present version of Assumption C.1 requires that we have at least $2d+1$ distinct source domains. Although this condition can be restrictive in high-dimensional data, we conjecture that there is a possibility for this assumption to be made less stringent in the future because the identification condition is only known to be a sufficient condition, not a necessary condition. However, pursuing a refinement of the identification condition is out of the scope of this chapter. Among the various methods for nonlinear ICA, we chose to use GCL [124] because it can operate under a nonparametric assumption on the IC distributions whereas other nonlinear ICA methods [123, 122, 141] may require parametric assumptions.

## C.2 Details of Real-world Data Experiment

Here, we describe more implementation details of the experiment. Our experiment code can be found at https://github.com/takeshi-teshima/few-shot-domain-adaptation-by-causal-mechanism-transfer.

### C.2.1 Dataset Details

**Gasoline consumption data.** The data was downloaded from http://bcs.wiley.com/he-bcs/Books?action=resource&bcsId=4338&itemId=1118672321&resourceId=13452.

### C.2.2 Model Details: Invertible Neural Networks

Here, we describe the details of the Glow architecture [145] used in our experiments. Glow consists of three types of layers that are invertible *by design*, namely affine coupling layers, $1 \times 1$ convolution layers, and activation normalization (actnorm) layers. In our implementation, we use actnorm as the first layer, and each of the subsequent layers consists of a $1 \times 1$ convolution layer followed by an affine coupling layer.

**Affine coupling layers.** The coefficients $s$ and $t$ for affine coupling layers in the notation of Kingma and Dhariwal [145] are parametrized by two one-hidden-layer neural networks whose number of hidden units is the same and the first layer parameter is shared. The activation functions of the first layer, the second layer of $s$, and the second layer of $t$ are the rectified linear unit (ReLU) activation [161], the hyperbolic tangent function, and the linear activation function, respectively. A standard practice of affine coupling layers is to compose the coefficient $s$ with an exponential function $x \mapsto \exp(x)$ so as to simplify the computation of the log-determinant of the Jacobian [145]. In our implementation, since we do not require the computation of the log-determinant, we omit this device and instead compose $x \mapsto (x + 1)$. The addition of 1 shifts the parameter space so that $(s, t) = (0, 0)$ corresponds to the the identity map, where 0 denotes the constant zero function. The split of the affine coupling layers is fixed at $(\lfloor \frac{d}{2} \rfloor, d - \lfloor \frac{d}{2} \rfloor)$.

**$1 \times 1$ convolution layers.** We initialize the parameters of the neural networks by $\mathcal{N}(0, \frac{1}{m})$ where $m$ is the number of parameters of each layer and $\mathcal{N}$ is the normal distribution.

### C.2.3 Model Details: Penultimate Layer Networks

We initialize the parameter for each layer of $\psi_j$ by $\mathrm{Unif}(-\sqrt{\frac{1}{m}}, \sqrt{\frac{1}{m}})$, where $m$ is the number of input features and Unif is the uniform distribution.

### C.2.4 Training Details

During the training of GCL, we fix the batch size at 32.

### C.2.5 Compared Methods Details

Here, we detail the methods compared through the experiment. Note that the present chapter focuses on regression problems as our approach is based on ICA, and hence the methods for classification domain adaptation are not comparable.

**TrAdaBoost.** As suggested in Pardoe and Stone [200], we use the linear loss function and set the maximum number of internal boosting iterations at 30.

**GDM.** We fix the number of sampling required for approximating the maximization in the generalization discrepancy at 200. This method presumes using hypothesis classes in a reproducing kernel Hilbert space (RKHS).

**Copula.** For this model, the probabilistic model of non-parametric R-vine copula of depth 1 is used following Lopez-paz et al. [172]. Kernel density estimators with RBF kernel are used for estimating the marginal distributions and the copulas. The bandwidths of the RBF kernels are determined using the rule-of-thumb implemented as "normal-reference" in the *np* package of *R* language [100]. The predictions are made by numerically aggregating the estimated conditional distribution over the interval $[\min_i Y_i - 2\sigma, \max_i Y_i + 2\sigma]$ where $\sigma$ denotes the square root of the unbiased variance of $\{Y_i\}_{i=1}^{n_{\mathrm{src}}}$. The aggregation is performed by discretizing the interval into a grid of 300 points. The level of the two-sample test is fixed at 0.05 for all combination of the two-sample tests following the experiment code of Lopez-paz et al. [172]. This method is a single-source domain adaptation method and we pool all source domain data for adaptation.

## C.3 Details of Synthetic-data Experiment

Here, we describe the details of the synthetic-data experiment reported in Section 4.5.4.

### C.3.1 Data-generating Process

To randomly generate a mixing map, we first generate an invertible neural network $\tilde{\boldsymbol{F}}$, a matrix $B$, and a scaling parameter $a > 0$, and we define $\boldsymbol{F} = \frac{1}{a}\tilde{\boldsymbol{F}} \circ B$.

The architecture of $\tilde{\boldsymbol{F}}$ is a composition of 2 blocks of invertible layers of the form

$$x \mapsto R(\mathrm{Coupling}(x) + t_{\mathrm{global}})$$

where $R \in \mathbb{R}^{d\times d}$ is a regular matrix and $t_{\mathrm{global}} \in \mathbb{R}^d$. For simplicity, let us denote $L := \lfloor \frac{d}{2} \rfloor$. Here, Coupling is a *general incompressible-flow network* (GIN; [252]) layer defined as

$$\mathrm{Coupling}(x) = (x_1 \odot \exp(0.2(\tanh(s(x_2) - \bar{s}(x_2)))) + 0.1t(x_2), x_2),$$

where $x_1$ denotes the first $d - L$ elements of $x$ and $x_2$ the last $L$ elements, $\odot$ denotes the element-wise product, the map $(s, t) : \mathbb{R}^L \to \mathbb{R}^{2(d-L)}$ is modeled by a multi-layer perceptron (MLP; [99]), $\bar{s}(x_2) := \frac{1}{d-L}\sum_{j=1}^{d-L} s^{(j)}(1, \ldots, 1)^\top$, and $\exp, \tanh$ are applied in an element-wise manner.

Each block in $\tilde{\boldsymbol{F}}$ is initialized as follows: $t_{\mathrm{global}}$ is initialized as a zero vector, the map $(s, t)$ is modeled by a one-hidden-layer feed-forward neural network [99] with 128 hidden units whose weights are randomly initialized using the Xavier initialization [87], and $R \in \mathrm{SO}(d)$ is sampled from the Haar measure [256].

Given $r \in [d]$, we obtain $B$ as $B = r^{-1}VU^\top$ by generating two matrices $U, V \in \mathbb{R}^{d\times r}$. We generate $U$ and $V$ by sampling each element independently from the uniform distribution over $[-\sqrt{3}, \sqrt{3}]$. After the generation, we confirmed that the rank of $B$ is indeed $r$, i.e., that it is not smaller than $r$.

The ICs $S_{k,i}^{\mathrm{Src}}$ $(k \in [K], i \in [n_k])$ and $S_i$ $(i \in [n_{\mathrm{Tar}} + n_{\mathrm{test}}])$ are generated by the following procedure:

1. A scaling coefficient matrix $L = (L_1 \cdots L_K L_{\mathrm{Tar}}) \in \mathbb{R}^{d\times(K+1)}$ is generated by sampling each element independently from $\mathrm{Unif}([0, 1])$.
2. For each $k \in [K]$, a matrix $A_k = (A_{k,1} \cdots A_{k,n_k}) \in \mathbb{R}^{d\times n_k}$ is generated by sampling each element independently from $\mathrm{Lap}(0, 1/\sqrt{2})$, where $\mathrm{Lap}(\mu, \lambda)$ is the Laplace distribution with

the location parameter $\mu \in \mathbb{R}$ and the scale parameter $\lambda > 0$. For the target domain, $A_{\mathrm{Tar}} = (A_{\mathrm{Tar},1} \cdots A_{\mathrm{Tar},n_{\mathrm{Tar}}+n_{\mathrm{test}}}) \in \mathbb{R}^{d \times (n_{\mathrm{Tar}}+n_{\mathrm{test}})}$ is generated in the same manner.

3. The ICs are generated by $S^{\mathrm{Src}}_{k,i} = A_{k,i} \odot L_k$ and $S_i = A_{\mathrm{Tar},i} \odot L_{\mathrm{Tar}}$.

After generating the ICs, in order to approximately standardize the scales of the generated data sets, we define $a$ to be the empirical standard deviation of $\{(\tilde{\boldsymbol{F}}(B(S^{\mathrm{Src}}_{k,i})))_j\}_{j \in [d], k \in [K], i \in [n_k]}$, and we define $\boldsymbol{F} = \frac{1}{a}\tilde{\boldsymbol{F}} \circ B$.

Finally, the data sets $\mathcal{D}_k = \{\boldsymbol{Z}^{\mathrm{Src}}_{k,i}\}_{i=1}^{n_k}$ $(k \in [K])$, $\mathcal{D}_{\mathrm{Tar}} = \{\boldsymbol{Z}_i\}_{i=1}^{n_{\mathrm{Tar}}}$, and $\mathcal{D}_{\mathrm{test}} = \{\boldsymbol{Z}_{n_{\mathrm{Tar}}+i}\}_{i=1}^{n_{\mathrm{test}}}$ are generated by $\boldsymbol{Z}^{\mathrm{Src}}_{k,i} = \boldsymbol{F}(S^{\mathrm{Src}}_{k,i})$ and $\boldsymbol{Z}_i = \boldsymbol{F}(S_i)$.

### C.3.2 Proposed Method Configuration

For $\hat{\boldsymbol{F}}$, we use the same architecture and initialization procedures as that of $\tilde{\boldsymbol{F}}$. Note that $B$ and the scaling $a$ are not included in the model $\hat{\boldsymbol{F}}$. For training, we used the Adam optimizer [144] with fixed parameters $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$, fixed initial learning rate $10^{-4}$, and the maximum number of epochs 128. To sample the IC candidates, we randomly sampled the indices for each dimension from an independent uniform distribution over $[n_{\mathrm{Tar}}]$. We trained a one-class support vector machine (OCSVM; [235]) on the union of the source-domain data for the same reason as Section 4.5.2. The configuration of the OCSVM was the same as Section 4.5.2. The trained OCSVM was applied to the pseudo-data generated by applying $\hat{\boldsymbol{F}}$ to the IC candidates. Since the generated data after the filtering may not add up to $n_{\mathrm{aug}}$ points, we randomly sampled up to $3n_{\mathrm{aug}}$ points and selected up to the first $n_{\mathrm{aug}}$ points that remained after the filtering by the OCSVM. For training the predictor, we always concatenated the original target-domain data set $\mathcal{D}_{\mathrm{Tar}}$ with the pseudo-data generated by the proposed method.

### C.3.3 Evaluation

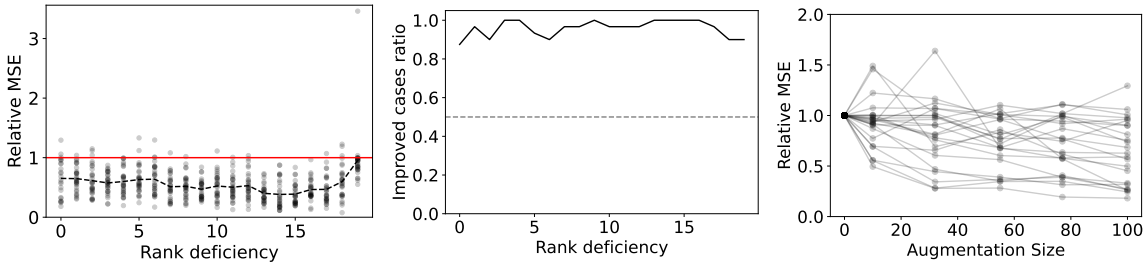
We set $n_k = n_{\mathrm{test}} = 512$ $(k \in [K])$. For the predictor model class, we employed the GBRTs [81, 41] using the same configuration as the experiment reported in Section 3.5 with the following hyper-parameter candidates: the number of leaves was fixed as 64, the number of boosting rounds was searched in $\{500, 1000, 2000\}$, and the $\ell_2$-regularization coefficient was searched in $\{10^{-1}, 10^{-2}, 10^{-3}\}$. To select the hyperparameters of the predictor hypothesis class, we performed the grid-search based on 3-fold cross-validation on the union of the original training data and the augmented data. For the implementation of the predictor model, we employed the *xgboost* library of Python [41]. See Chen and Guestrin [41] for the optimization method and the other details.

### C.3.4 Supplementary Figures

In Figure C.1, we report the results of the same experiment as Section 4.5.4 with $n_{\mathrm{Tar}} = 10$ and $n_{\mathrm{Tar}} = 20$. As observed in Section 4.5.4, we can observe the robustness of the proposed method to the ill-conditioning of the IC estimation problem as well as the tendency that a $\mathcal{O}(n_{\mathrm{Tar}})$ augmentation yields similar performance improvements to an augmentation of $\mathcal{O}(n^2_{\mathrm{Tar}})$ points.

## C.4 Details and Proofs of Theorem 4.2

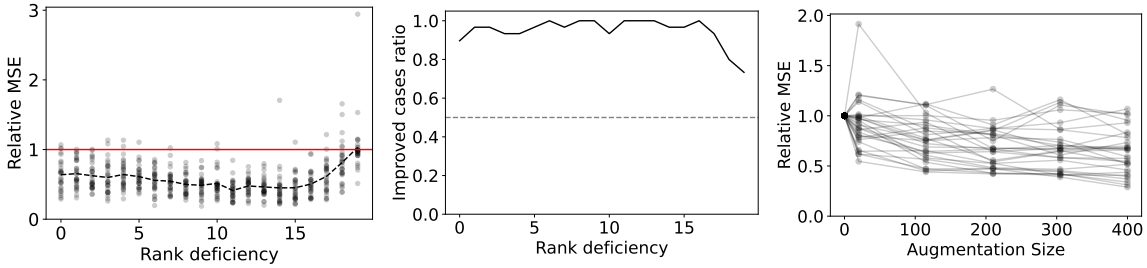
Here, we detail the assumptions, the statement, and the proof of Theorem 4.2.




(a) relMSE in relation to rank deficiency $d - r$ ($n_{\text{Tar}} = 10$).

(b) Ratio of improved cases (relMSE $< 1$) in relation to rank deficiency $d - r$ ($n_{\text{Tar}} = 10$).

(c) relMSE in relation to augmentation size $n_{\text{aug}}$ ($n_{\text{Tar}} = 10$).




(d) relMSE in relation to rank deficiency $d - r$ ($n_{\text{Tar}} = 20$).

(e) Ratio of improved cases (relMSE $< 1$) in relation to rank deficiency $d - r$ ($n_{\text{Tar}} = 20$).

(f) relMSE in relation to augmentation size $n_{\text{aug}}$ ($n_{\text{Tar}} = 20$).

Figure C.1: Additional results of the synthetic data experiments. For (a), (b), (d), and (e), we fixed $n_{\text{aug}} = n_{\text{Tar}}^2$. For (c) and (f), we fixed $d - r = 0$.

### C.4.1 Notation

To make the proof self-contained, we first recall some general and problem-specific notation. In the notation here, we omit the domain identifiers from the distributions and the sample size, such as *Tar* or *Src*, because only the target domain data or their distributions appear in the proofs. The theorem holds regardless of how $\hat{\boldsymbol{F}}$ is estimated as long as $\hat{\boldsymbol{F}}$ is independent of the target domain data. In the proof, we extend the maximal discrepancy bound of U-statistics previously proved for the case of degree-2 in Rejchel [216], to allow higher degrees.

**General mathematical notation.** We denote the set of natural numbers (resp. real numbers) by $\mathbb{N}$ (resp. $\mathbb{R}$). For any $N \in \mathbb{N}$, we define $[N] := \{1, 2, \ldots, N\}$. We use $\binom{a}{b}$ to denote the number of $b$-combinations of $a$ elements. For a finite set $A$, the notation $\overline{\sum_{a \in A}}$ denotes the operator to take an average over $A$, i.e., $\overline{\sum_{a \in A}} h = \frac{1}{|A|} \sum_{a \in A} h(a)$. For a $d$-dimensional function $h$, we denote its $j$-th dimension ($j \in [d]$) by suffixing $h_j$. For a vector $s$, we denote its $j$-th element by $s^{(j)}$. We denote the Jacobian determinant of a differentiable function $\psi$ at $a$ by $\mathrm{J}\psi(a) := \det \frac{\mathrm{d}\psi(a)}{\mathrm{d}a}$. We denote the identity matrix by $I$ regardless of the size of the matrices when there is no ambiguity. For finite dimensional vectors, we denote the 2-norm by $\|\cdot\|_{\ell^2}$ and the 1-norm by $\|\cdot\|_{\ell^1}$. For square matrices, we denote the operator-2 norm by $\|\cdot\|_{\text{op}}$ and the operator-1 norm by $\|\cdot\|_{\text{op}(1)}$. We use $W^{k,p}$ to denote the Sobolev space (on $\mathbb{R}^d$) of order $k$ and define its associated norm by $\|h\|_{W^{k,p}} := \left(\sum_{|\alpha| \le k} \left\|h^{(\alpha)}\right\|_{L^p}^p\right)^{1/p}$ where $\alpha$ is a multi-index and $h^{(\alpha)}$ denotes the partial derivative $\frac{\partial^{|\alpha|} h}{\partial s_1^{\alpha_1} \cdots \partial s_d^{\alpha_d}}$ [3, Paragraph 3.1]. We let $\mathfrak{S}_d$ be the degree-$d$ symmetric group, $\mathfrak{G}_j^d := \{\tau : [d] \to [j] \mid \tau \text{ is surjective}\}$ be the set of $j$ grouping of indices in $[d]$, and $\mathfrak{I}_j^n := \{\rho : [j] \to [n] \mid \rho \text{ is injective}\}$ be the set of all size-$j$ combinations (without replacement) of indices in $[n]$.

**Distributions and expectations.** We denote by $\mathcal{Q}$ the set of all factorized distributions on $\mathbb{R}^d$ with absolutely continuous marginals. For a measure $P$, we denote its $j$-product measure by $P^j := P \otimes \cdots \otimes P$ (repeated $j$ times). We assume that all measures appearing in this proof are absolutely continuous with respect to the Lebesgue measure. The push-forward of a distribution $p$ by a function $h$ is denoted by $h_\sharp(p)$. The expectation of a function $h$ with respect to measure $P$ is denoted by $Ph$ (if it exists) by abuse of notation. We also abuse the notation to use $\psi(s, P, \ldots, P)$ as the shorthand for $P^{d-1}\psi(s, S'_2, \ldots, S'_d)$ where $\{S'_j\}_{j=2}^d \overset{\text{i.i.d.}}{\sim} P$.

### C.4.2 Problem Setup

We denote the target domain distribution by $p$. We fix a hypothesis class $\mathcal{H}(\subset \{h : \mathbb{R}^{d-1} \to \mathbb{R}\})$, and our goal is to find a $h \in \mathcal{H}$ such that the risk functional

$$\mathcal{R}(h) := \int p(\boldsymbol{z})\ell(h, \boldsymbol{z})\mathrm{d}\mathbf{x}$$

is small, where $\ell : \mathcal{H} \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a loss function. We denote by $h^*$ a minimizer of $\mathcal{R}$ (assuming it exists). To this end, we are given the training data $\mathcal{D} := \{\boldsymbol{Z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$. Throughout, we assume $n \geq d$. To complement the smallness of $n$, we assume the existence of a generative mechanism. Concretely, we assume that there exists a diffeomorphism $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^d$ such that $q := (\boldsymbol{F}^{-1})_\sharp(p)$ satisfies $q \in \mathcal{Q}$. With this transform, the original risk functional is also expressed as

$$\mathcal{R}(h) = \int q(s)\ell(h, \boldsymbol{F}(s))\mathrm{d}s.$$

As an estimator of $\boldsymbol{F}$, we are given another diffeomorphism $\hat{\boldsymbol{F}} : \mathbb{R}^d \to \mathbb{R}^d$ such that $\hat{\boldsymbol{F}} \simeq \boldsymbol{F}$. With this $\hat{\boldsymbol{F}}$, the proposed method converts the dataset $\mathcal{D}$ by $S_i := \hat{\boldsymbol{F}}(\boldsymbol{Z}_i)$. We can regard $\check{\mathcal{D}} := \{S_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \check{q}$, where $\check{q} := (\hat{\boldsymbol{F}}^{-1} \circ \boldsymbol{F})_\sharp(q)$. We use $Q$ (resp. $\check{Q}$) to denote the probability measure corresponding to the density $q$ (resp. $\check{q}$). This conversion results in the relation:

$$\check{q}(s) = q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s)) \left|(\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right|.$$

As a candidate hypothesis $h \in \mathcal{H}$, the proposed method selects a minimizer $\check{h} \in \mathcal{H}$ of the proposed risk estimator $\check{\mathcal{R}}$ defined as

$$\check{\mathcal{R}}(h) := \frac{1}{n^d} \sum_{(i_1,\ldots,i_d)\in[n]^d} \ell(h, \hat{\boldsymbol{F}}(\hat{s}_{i_1}^{(1)}, \ldots, \hat{s}_{i_d}^{(d)})). \tag{C.2}$$

In the proof, we evaluate its concentration around the expectation $\bar{\check{\mathcal{R}}}(h) := \mathbb{E}_{\check{\mathcal{D}}}\check{\mathcal{R}}(h)$. We use $\mathbb{E}_{\check{\mathcal{D}}}$ to denote the expectation with respect to $\check{\mathcal{D}}$. Let $\bar{\check{h}}$ denote a hypothesis which minimizes $\bar{\check{\mathcal{R}}}(h)$ (assuming it exists).

In what follows, for notational simplicity, we define the $d$-variate symmetric function $\tilde{\ell}$ as

$$\tilde{\ell}(s_1, \ldots, s_d) = \overline{\sum_{\pi\in\mathfrak{S}_d}} \ell(h, \hat{\boldsymbol{F}}(s_{\pi(1)}^{(1)}, \ldots, s_{\pi(d)}^{(d)})),$$

where $\overline{\sum_{\pi\in\mathfrak{S}_d}}$ indicates an averaging operation over all permutations (without replacement) of $[d]$. We use $\hat{\mathbb{E}}_n$ to denote the sample average operator with respect to $\mathcal{D}$ or $\check{\mathcal{D}}$, depending on the context.

### C.4.3 Assumptions

**Assumption C.2** (The underlying density function is bounded and Lipschitz continuous)**.** *Assume*

$$B_q := \sup_{s\in\mathbb{R}^d} q(s) < \infty, \quad L_q := \sup_{s_1\neq s_2} \frac{|q(s_1)-q(s_2)|}{\|s_1-s_2\|} < \infty.$$

**Assumption C.3** ($\boldsymbol{F}^{-1}$ is Lipschitz continuous and Hölder continuous)**.** *We assume* $\boldsymbol{F}^{-1} \in C^{1,1}$ *where* $C^{1,1}$ *is the* $(1,1)$*-Hölder space [3, Paragraph 1.29] and*

$$L_{\boldsymbol{F}^{-1}} := \sup_{z_1\neq z_2} \frac{\|\boldsymbol{F}^{-1}(z_1)-\boldsymbol{F}^{-1}(z_2)\|}{\|z_1-z_2\|} < \infty.$$

**Assumption C.4** (Bounded derivatives of $\boldsymbol{F}$ and $\boldsymbol{F}^{-1}$)**.** *Assume that*

$$B^{\infty}_{\partial\boldsymbol{F}} := \sup_{s\in\mathbb{R}^d} \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) \right\|_{\infty} < \infty, \quad B^{\infty}_{\partial\boldsymbol{F}^{-1}} := \sup_{z\in\mathbb{R}^d} \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(z) \right\|_{\infty} < \infty.$$

*where* $\|\cdot\|_{\infty}$ *denotes the maximum absolute value of the elements of a matrix.*

**Assumption C.5** (Loss function is bounded and uniformly Lipschitz continuous in $\boldsymbol{Z}$)**.** *The considered loss function takes values in a bounded interval:*

$$\ell : \mathcal{H}\times\mathcal{Z} \to [0, B_\ell],$$

*where* $0 < B_\ell < \infty$*. Also assume*

$$L_{\ell_{\mathcal{H}}} := \sup_{h\in\mathcal{H}} \sup_{z_1\neq z_2} \frac{|\ell(h,z_1)-\ell(h,z_2)|}{\|z_1-z_2\|} < \infty.$$

**Assumption C.6** (Estimated feature extractor)**.** *Assume* $\hat{\boldsymbol{F}}$ *is independent of* $\mathcal{D}$ *and that* $\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \in W^{1,m}$ *for all* $(j,m) \in [d]\times[d]$*.*

Although $\hat{\boldsymbol{F}}$ and $f$ are assumed to be diffeomorphisms in the classical sense (implying that they are strongly differentiable), we introduce the Sobolev space because we want to measure their difference and their difference of derivatives in terms of integration.

**Assumption C.7** (Entropic condition: Euclidean class [237])**.** *The function class* $\Phi := \{\tilde{\ell} : h \in \mathcal{H}\}$ *is Euclidean for the envelope* $F$ *and constants* $A$ *and* $V$ *[237], i.e., if* $\mu$ *is a measure for which* $\mu F^2 < \infty$*, then*

$$\mathrm{Pack}(t, \mathrm{dist}_\mu, \Phi) \le At^{-V}, \quad 0 < t \le 1,$$

*where* $\mathrm{Pack}(t, \mathrm{dist}_\mu, \Phi)$ *denotes the packing number of* $\Phi$ *with respect to the radius* $t$ *and the pseudometric* $\mathrm{dist}_\mu$ *defined by*

$$\mathrm{dist}_\mu(\phi_1,\phi_2) := \left[\mu|\phi_1-\phi_2|^2/\mu F^2\right]^{1/2}$$

*for* $\phi_1, \phi_2 \in \Phi$*. Without loss of generality, we take the envelope* $F$ *such that* $F(\cdot) \le B_\ell$*.*

**Assumption C.8.** *The hypothesis class $\mathcal{H}$ is expressive enough so that the model approximation error does not expand due to $\hat{\boldsymbol{F}}$, i.e.,*

$$\inf_{h\in\mathcal{H}} \bar{\bar{\mathcal{R}}}(h) \leq \inf_{h\in\mathcal{H}} \mathcal{R}(h)$$

The following complexity measure of $\mathcal{H}$, which is a version of Rademacher complexity for our problem setting, is used to state the theorem.

**Definition C.1** (Effective Rademacher complexity)**.** *Define*

$$\mathfrak{R}(\mathcal{H}) := \frac{1}{n}\mathbb{E}_{\check{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S_2',\ldots,S_d'}[\tilde{\ell}(S_i,S_2',\ldots,S_d')]\right|\right]$$

*where $\{\sigma_i\}_{i=1}^n$ are independent uniform sign variables and $S_2',\ldots,S_d' \overset{i.i.d.}{\sim} \check{Q}$ are independent of all other random variables.*

We provide the definition of the ordinary Rademacher complexity in Section C.4.8 and make a comparison of the two complexity measures in terms of how they depend on the input dimensionality.

### C.4.4 Theorem Statement

Our goal is to prove the following theorem. This is a detailed version of the theorem appearing in Chapter 4.

**Theorem C.1** (Excess risk bound)**.** *Assume Assumptions C.2, C.3, C.4, C.5, C.6, C.7, and C.8. Then for arbitrary $\delta,\delta' \in (0,1)$, we have with probability at least $1-(\delta+\delta')$,*

$$\begin{aligned}&\mathcal{R}(\check{h}) - \mathcal{R}(h^*)\\ &\leq \underbrace{C\sum_{j=1}^{d}\left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,1}}}_{\textit{Approximation error}} + \underbrace{4d\mathfrak{R}(\mathcal{H}) + 2dB_\ell\sqrt{\frac{\log 2/\delta}{2n}}}_{\textit{Estimation error}} + \underbrace{\kappa_1(\delta',n) + dB_\ell B_q\kappa_2(\boldsymbol{F}-\hat{\boldsymbol{F}})}_{\textit{Higher order terms}}.\end{aligned}$$

*where*

$$\begin{aligned}C &:= B_qL_{\ell_{\mathcal{H}}} + dB_\ell(L_qL_{\boldsymbol{F}^{-1}} + B_qdC_1'),\\ C_1' &:= (d+1)^{3/2}\left(B^{\infty}_{\partial\boldsymbol{F}}\left(\sum_{k=1}^{d}\left\|\boldsymbol{F}_k^{-1}\right\|_{C^{1,1}}\right) + B^{\infty}_{\partial\boldsymbol{F}^{-1}}\right),\\ \kappa_1(\delta',n) &= \mathcal{O}(n^{-1})/\delta' + \mathcal{O}(n^{-1}),\\ \kappa_2(\boldsymbol{F}-\hat{\boldsymbol{F}}) &= \sum_{m=2}^{d}\binom{d}{m}C_m'\sum_{j=1}^{d}\left\|\boldsymbol{F}_j-\hat{\boldsymbol{F}}_j\right\|^m_{W^{1,m}}.\end{aligned}$$

*and $C_m'(m=1,\ldots,d)$ are constants determined in Lemma C.11.*

*Proof of Theorem C.1.* By adding and subtracting terms, we have

$$\begin{aligned}\mathcal{R}(\check{h}) - \mathcal{R}(h^*) = \underbrace{(\mathcal{R}-\bar{\bar{\mathcal{R}}})(\check{h})}_{\text{(A) Approximation error}} + \underbrace{\bar{\bar{\mathcal{R}}}(\check{h}) - \bar{\bar{\mathcal{R}}}(\bar{\check{h}})}_{\text{(B) Pseudo estimation error}}\\ + \underbrace{\bar{\bar{\mathcal{R}}}(\bar{\check{h}}) - \mathcal{R}(h^*)}_{\text{(C) Additional model misspecification error}}.\end{aligned}$$

Applying Lemma C.1 to (A), Lemma C.2 to (B), and Assumption C.8 to (C), we obtain the assertion. □

As it can be seen from the proof above, Theorem C.1 is proved in two parts, each corresponding to the two lemmas below. The first lemma evaluates the *approximation error* which reflects the fact that we are approximating $\boldsymbol{F}$ by $\hat{\boldsymbol{F}}$.

**Lemma C.1** (Approximation error bound)**.** *Given Assumptions C.2, C.3, C.4, C.5, and C.6. we have*

$$(\mathcal{R} - \bar{\bar{\mathcal{R}}})(\check{h}) \leq C \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + dB_\ell B_q \kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$$

*where $C$ and $\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$ are*

$$C := B_q L_{\ell_{\mathcal{H}}} + dB_\ell (L_q L_{\boldsymbol{F}^{-1}} + B_q d C_1'),$$
$$\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}}) := \sum_{m=2}^{d} \binom{d}{m} C_m' \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,m}}^{m}.$$

*and $C_m'(m = 1, \ldots, d)$ are constants determined in Lemma C.11.*

The second lemma evaluates the *pseudo estimation error* which reflects the fact that we rely on a finite sample to approximate the underlying distribution.

**Lemma C.2** (Pseudo estimation error bound)**.** *Assume that Assumptions C.2 and C.7 hold. Let the Rademacher complexity be defined as Definition C.1. Then for any $\delta, \delta' \in (0, 1)$, we have with probability at least $1 - (\delta + \delta')$ that*

$$\bar{\bar{\mathcal{R}}}(\check{h}) - \bar{\bar{\mathcal{R}}}(\bar{\bar{h}}) \leq 4dw_d \mathfrak{R}(\mathcal{H}) + 2dB_\ell w_d \sqrt{\frac{\log 2/\delta}{2n}} + \underbrace{2w_d(d-1) \sum_{j=2}^{d} \frac{C_j}{\delta'} n^{-j/2} + 4B_\ell \sum_{j=1}^{d-1} w_j}_{\mathcal{O}(n^{-1})}$$

*where $\{w_j\}_{j=1}^{d}$ are universal constants determined in Lemma C.3, and $\{C_j\}_{j=2}^{d}$ are constants determined in Lemma C.6. Note that $w_j = \mathcal{O}(n^{-(d-j)})$ and $w_d = \frac{n(n-1)\cdots(n-d+1)}{n^d} < 1$.*

In what follows, we first present some basic facts in Section C.4.5 and provide the proofs for the lemmas. We provide the proof of Lemma C.1 in Section C.4.7, and that of Lemma C.2 in Section C.4.6.

### C.4.5 V-statistic and U-statistic

The theoretical analysis is performed by interpreting the proposed risk estimator Equation (C.2) as a *V-statistic* (explained shortly). The proofs will be based on applying the following facts in order:

1. V-statistic can be represented as a weighted average of *U-statistics* with degrees from 1 to $d$, and only the degree-$d$) term is the leading term.
2. The degree-$d$ term is again decomposed into a degree-1 U-statistic and a set of *degenerate* U-statistics.
3. The degree-1 *U-statistic* is an i.i.d. sum admitting a Rademacher complexity bound.
4. The degenerate terms concentrate around zero following an exponential inequality under appropriate entropy conditions.

To consolidate the strategy given above, we describe what are V- and U-statistics, and how they relate to each other. These estimators emerge when we allow re-using the same data point repeatedly from a single sample to estimate a function which takes multiple data points.

**V-statistic.** For a given regular statistical functional of degree $d$ [162]:

$$\check{Q}^d\tilde{\ell} := \int \tilde{\ell}(s_1, \cdots, s_d)\check{q}(s_1)\cdots\check{q}(s_d)\mathrm{d}s_1\cdots\mathrm{d}s_d, \tag{C.3}$$

its associated von-Mises statistic (V-statistic) is the following quantity [162]:

$$V_n^d\tilde{\ell} := \frac{1}{n^d}\sum_{i_1=1}^{n}\cdots\sum_{i_d=1}^{n}\tilde{\ell}(S_{i_1}, \ldots, S_{i_d}).$$

Note that Equation (C.3) does not coincide with the expectation of $V_n^d\tilde{\ell}$ in general, i.e., the V-statistic is generally not an unbiased estimator. However, it is known to be a consistent estimator of Equation (C.3) [162].

**U-statistic.** Similarly, for a $j$-variate function $h(x_1, \ldots, x_j)$ that is symmetric and integrable, its corresponding U-statistic [162] of degree $j$ is

$$U_n^j h := \overline{\sum_{\rho\in\mathfrak{I}_j^n}} h(s_{\rho(1)}, \ldots, s_{\rho(j)}).$$

The V- and U-statistics are generalizations of the sample mean (which is the U- and V-statistics of degree 1). The important difference from the sample mean in higher degrees is that the summands may not be independent. To deal with the dependence, the following standard decompositions have been developed [162].

**Lemma C.3** (Decomposition of a V-statistic [162])**.** *A V-statistic can be expressed as a sum of U-statistics of degrees from* 1 *to* $d$ *[162, Section 4.2, Theorem 1]:*

$$V_n^d\tilde{\ell} = \sum_{j=1}^{d} w_j U_n^j\tilde{\ell}^{(j)}$$

*where the weights* $w_j$ *and* $j$*-variate functions* $\tilde{\ell}^{(j)}$ *are*

$$w_j := \frac{1}{n^d}\left|\mathfrak{G}_j^d\right|\binom{n}{j}, \quad \tilde{\ell}^{(j)}(s_1, \ldots, s_j) := \overline{\sum_{\tau\in\mathfrak{G}_j^d}}\tilde{\ell}(s_{\tau(1)}, \ldots, s_{\tau(d)}).$$

*Proof.* See [162, Section 4.2, Theorem 1 (p.183)]. □

*Remark* C.1. The weights $\{w_j\}_{j=1}^d$ satisfy $\sum_j w_j = 1$ [162, Section 4.2, Theorem 1 (p.183)]. We can also find the order of $w_j$ with respect to $n$ as:

$$w_d = \frac{1}{n^d}\underbrace{\left|\mathfrak{G}_j^d\right|}_{d!}\binom{n}{d} = \frac{n(n-1)\cdots(n-d+1)}{n^d} = \mathcal{O}(1),$$
$$w_j = \mathcal{O}(n^{-(d-j)}), \quad \tilde{\ell}^{(d)} = \tilde{\ell}.$$

**Lemma C.4** (Hoeffding decomposition of a U-statistic [237, p.449])**.** *A U-statistic with a symmetric kernel $\psi$ can be decomposed as a sum of U-statistics of degrees from 1 to $d$ as*

$$U_n^d\psi - \mathbb{E}_{\check{\mathcal{D}}}U_n^d\psi = \sum_{j=1}^{d} U_n^j\psi_j = \hat{\mathbb{E}}_n\psi_1 + \sum_{j=1}^{d} U_n^j\psi_j \tag{C.4}$$

*where $\{\psi_j\}_{j=1}^d$ are $j$-variate, symmetric and degenerate functions. Note that $\mathbb{E}_{\check{\mathcal{D}}}U_n^d\psi = \check{Q}^d\psi$. Here, a $j$-variate symmetric function $\psi_j$ is said* degenerate *when*

$$\forall s_2, \ldots, s_j, \quad \psi_j(\check{Q}, s_2, \ldots, s_j) = 0.$$

*Specifically, $\psi_1$ is*

$$\begin{aligned}\psi_1(s) &= \psi(s, \check{Q}, \ldots, \check{Q}) + \cdots + \psi(\check{Q}, \ldots, \check{Q}, s) - d\check{Q}^d\psi \\ &= d \cdot (\psi(s, \check{Q}, \ldots, \check{Q}) - \check{Q}^d\psi) \quad (\textit{by symmetry}).\end{aligned} \tag{C.5}$$

For further details, see [237, p.449]. Note that in [237, p.449], Equation (C.4) is written using $\check{Q}^d\psi$ in place of $\mathbb{E}_{\check{\mathcal{D}}}U_n^d\psi$. This is because

$$\mathbb{E}_{\check{\mathcal{D}}}U_n^d\psi = U_n^d\mathbb{E}_{\check{\mathcal{D}}}\psi = U_n^d\check{Q}^d\psi = \check{Q}^d\psi$$

holds by linearity and symmetry.

*Remark* C.2 (Connecting the lemmas to Section C.4.6). It can be easily checked by definition that the proposed risk estimator Equation (C.2) takes the form of a V-statistic: $\check{\mathcal{R}}(h) = V_n^d\tilde{\ell}$ for each $h \in \mathcal{H}$. Let us denote $\tilde{\ell}^*(s) := \tilde{\ell}(s, \check{Q}, \ldots, \check{Q})$. Then $\mathbb{E}_{\check{\mathcal{D}}}\tilde{\ell}^* = \check{Q}^d\tilde{\ell}$ holds by definition. Substituting these into Equation (C.5), we have that Equation (C.4) applied to $\psi = \tilde{\ell}$ is equivalent to

$$U_n^d\tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}}U_n^d\tilde{\ell} = d \cdot (\hat{\mathbb{E}}_n\tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}}\tilde{\ell}^*) + \sum_{j=2}^{d} U_n^j\tilde{\ell}_j.$$

where $\{\tilde{\ell}_j\}_{j=2}^d$ are symmetric degenerate functions. In Section C.4.6, we first decompose $\check{\mathcal{R}}(h)$ into a sum of U-statistics. After such conversion, we take a closer look at the leading term, $\hat{\mathbb{E}}_n\tilde{\ell}^*$.

### C.4.6 Proof of Pseudo Estimation Error Bound

*(Proof of Lemma C.2).* First, we have

$$\begin{aligned}\bar{\check{\mathcal{R}}}(\check{h}) - \bar{\check{\mathcal{R}}}(\bar{\check{h}}) &= \bar{\check{\mathcal{R}}}(\check{h}) - \check{\mathcal{R}}(\check{h}) + \check{\mathcal{R}}(\check{h}) - \bar{\check{\mathcal{R}}}(\bar{\check{h}}) \\ &\leq \bar{\check{\mathcal{R}}}(\check{h}) - \check{\mathcal{R}}(\check{h}) + \check{\mathcal{R}}(\bar{\check{h}}) - \bar{\check{\mathcal{R}}}(\bar{\check{h}}) \leq 2\sup_{h\in\mathcal{H}}\left|\check{\mathcal{R}}(h) - \bar{\check{\mathcal{R}}}(h)\right|.\end{aligned}$$

Now the right-most expression can be decomposed as

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}} |\check{\mathcal{R}}(h) - \bar{\check{\mathcal{R}}}(h)| = \sup_{h\in\mathcal{H}} |V_n^d \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} V_n^d \tilde{\ell}| \\
&\le w_d \sup_{h\in\mathcal{H}} \left| U_n^d \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} U_n^d \tilde{\ell} \right| + \sum_{j=1}^{d-1} w_j \sup_{h\in\mathcal{H}} \left| U_n^j \tilde{\ell}^{(j)} - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)} \right| \quad (\because \text{Lemma C.3}) \\
&\le w_d \sup_{h\in\mathcal{H}} |U_n^d \tilde{\ell} - \mathbb{E}_{\check{\mathcal{D}}} U_n^d \tilde{\ell}| + 2B_\ell \sum_{j=1}^{d-1} w_j \\
&\le w_d \left( \sup_{h\in\mathcal{H}} |\hat{\mathbb{E}}_n \tilde{\ell}_1| + \sum_{j=2}^{d} \sup_{h\in\mathcal{H}} |U_n^j \tilde{\ell}_j| \right) + 2B_\ell \sum_{j=1}^{d-1} w_j \quad (\because \text{Lemma C.4}) \\
&= w_d \left( \underbrace{\sup_{h\in\mathcal{H}} |\hat{\mathbb{E}}_n d(\tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^*)|}_{\text{Addressed in Lemma C.5}} + \underbrace{\sum_{j=2}^{d} \sup_{h\in\mathcal{H}} |U_n^j \tilde{\ell}_j|}_{\text{Addressed in Lemma C.6}} \right) + 2B_\ell \sum_{j=1}^{d-1} w_j.
\end{aligned}
$$

where $\tilde{\ell}_j$ are symmetric degenerate functions and $\tilde{\ell}^*$ is defined as in Remark C.2. Applying Lemma C.5 to the first term and Lemma C.6 to the second term with the union bound, we obtain the assertion. □

In the last part of the proof we used the following lemmas. Because the leading term is an i.i.d. sum, the following Rademacher complexity bound can be proved.

**Lemma C.5** (U-process bound: the leading term)**.** *Assume Assumption C.2 holds. Then, we have with probability at least* $1-\delta$*,*

$$
\sup_{h\in\mathcal{H}} |\hat{\mathbb{E}}_n (\tilde{\ell}^* - \mathbb{E}_{\check{\mathcal{D}}} \tilde{\ell}^*)| \le 2\mathfrak{R}(\mathcal{H}) + B_\ell \sqrt{\frac{\log(2/\delta)}{2n}},
$$

*where* $\mathfrak{R}$ *is defined in Definition C.1.*

*Proof.* Applying the standard one-sided Rademacher complexity bound based on McDiarmid's inequality [184, Theorem 3.3] twice with the union bound, we obtain the lemma. □

The other terms than the leading term are degenerate U-statistics, hence the following holds under appropriate entropy assumptions.

**Lemma C.6** (U-process bound: degenerate terms [237, Corollary 7])**.** *Assume Assumption C.7. Then for each* $j = 2, \dots, d$*, there exist constants* $C_j$ *such that for any* $\delta \in (0,1)$*, we have with probability at least* $1 - \delta'/(d-1)$*,*

$$
\sup_{h\in\mathcal{H}} |U_n^j \tilde{\ell}_j| \le \frac{(d-1)}{\delta'} C_j n^{-j/2}
$$

*where* $C_j$ *depends only on* $A$*,* $V$*, and* $B_\ell$*.*

*Proof.* The proof follows a similar path as that of [237, Corollary 7], but we provide more explicit expressions to inspect the order with respect to $n$. Let $\Phi^{(j)}_{\mathcal{H},\hat{\boldsymbol{F}}} := \{\tilde{\ell}_j : h \in \mathcal{H}\}$. Then $\Phi^{(j)}_{\mathcal{H},\hat{\boldsymbol{F}}}$ is Euclidean for an envelope $F_j$ satisfying $\check{Q}^j F_j^2 < \infty$ by Lemma 6 in Sherman [237] and Assumption C.7. In addition, $\Phi^{(j)}_{\mathcal{H},\hat{\boldsymbol{F}}}$ is a set of functions degenerate with respect to $\check{Q}$. Without loss of generality,

we can take $F_j$ such that $F_j \leq B_\ell$. Similarly to the proof of [237, Corollary 4], we can apply [237, Main Corollary] with $p = 1$ in their notation to obtain

$$\mathbb{E}_{\check{\mathcal{D}}} \sup_{h \in \mathcal{H}} |n^{j/2} U_n^j \tilde{\ell}_j| \leq \Gamma A^{1/2m} (\check{Q}^j F_j^2)^{(\epsilon+\alpha)/2} \leq \underbrace{\Gamma A^{1/2m} (B_\ell)^{\epsilon+\alpha}}_{=: C_j}$$

where $\Gamma$ is a universal constant [237, Main Corollary], $\epsilon \in (0, 1)$ and $m$ are chosen to satisfy $1 - V/2m > 1 - \epsilon$, and $\alpha = 1 - V/2m$. By applying Markov inequality, we have for arbitrary $u > 0$,

$$\mathbb{P}_{\check{\mathcal{D}}} \left( \sup_{h \in \mathcal{H}} |n^{j/2} U_n^j \tilde{\ell}_j| > u \right) \leq \frac{C_j}{u},$$

where $\mathbb{P}_{\check{\mathcal{D}}}(E)$ denotes the probability of the event $E$ with respect to $\check{\mathcal{D}}$. Equating the right hand side with $\delta'/(d-1)$ and solving for $u$, we obtain the result. □

### C.4.7 Proof of Approximation Error Bound

*(Proof of Lemma C.1).* Due to Lemma C.3, we have

$$\begin{aligned}
\sup_{h \in \mathcal{H}} \left( \mathcal{R}(h) - \bar{\bar{\mathcal{R}}}(h) \right) &= \sup_{h \in \mathcal{H}} \left( \mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} V_n^d \tilde{\ell} \right) \\
&= \sup_{h \in \mathcal{H}} \left( \sum_{j=1}^{d} w_j (\mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)}) \right) \\
&\leq w_d \sup_{h \in \mathcal{H}} \left( \mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U_n^d \tilde{\ell}^{(d)} \right) + 2B_\ell \sum_{j=1}^{d-1} \underbrace{w_j}_{\mathcal{O}(n^{-(d-j)})} \\
&\leq w_d \sup_{h \in \mathcal{H}} \left( \mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U_n^d \tilde{\ell}^{(d)} \right) + 2B_\ell \mathcal{O}(n^{-1})
\end{aligned}$$

By applying Lemmas C.7 (with $j = d$), we obtain

$$\sup_{h \in \mathcal{H}} \left( \mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U_n^d \tilde{\ell}^{(d)} \right) \leq \sup_{h \in \mathcal{H}} \left\| \ell(\boldsymbol{F}(h, \cdot)) - \ell(h, \hat{\boldsymbol{F}}(\cdot)) \right\|_{L^1(q)} + dB_\ell \left\| q - \check{q} \right\|_{L^1}.$$

The right-hand side can be further bounded by applying Lemmas C.9 and C.8 by

$$\begin{aligned}
&B_q L_{\ell_{\mathcal{H}}} \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + dB_\ell \left( (L_q L_{\boldsymbol{F}^{-1}} + B_q d C_1') \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + B_q \kappa_2 (\boldsymbol{F} - \hat{\boldsymbol{F}}) \right) \\
&\leq (B_q L_{\ell_{\mathcal{H}}} + dB_\ell (L_q L_{\boldsymbol{F}^{-1}} + B_q d C_1')) \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + dB_\ell B_q \kappa_2 (\boldsymbol{F} - \hat{\boldsymbol{F}})
\end{aligned}$$

and hence the assertion of the lemma. □

The above proof combined three approximation bounds, which are shown in the following lemmas. The following lemma reduces the difference in the expectation of U-statistic into the differences in the loss function and the density function. Although we apply the following Lemma C.7 only with $j = d$, we prove its general form for $j \in [d]$.

**Lemma C.7** (Approximation bound for U-statistic of degree-$j$)**.** *Fix $j \in [d]$. Assume Assumption C.2. Then we have for any $h \in \mathcal{H}$,*

$$\mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U_n^j \tilde{\ell}^{(j)} \leq \left\| \ell(h, \boldsymbol{F}(\cdot)) - \ell(h, \hat{\boldsymbol{F}}(\cdot)) \right\|_{L^1(q)} + jB_\ell \left\| q - \check{q} \right\|_{L^1}$$

*Proof.* Let us define a $d$-variate function $\ell^\dagger$ and a $j$-variate function $\ell^{\dagger(j)}$ (similarly to $\tilde{\ell}$ and $\tilde{\ell}^{(j)}$, respectively) by

$$\ell^\dagger(s_1,\dots,s_d) := \overline{\sum_{\pi\in\mathfrak{S}_d}} \ell(h, \boldsymbol{F}(s^{(1)}_{\pi(1)},\dots,s^{(d)}_{\pi(d)})),$$
$$\ell^{\dagger(j)}(s_1,\dots,s_j) := \overline{\sum_{\tau\in\mathfrak{G}^d_j}} \ell^\dagger(s_{\tau(1)},\dots,s_{\tau(d)}).$$

Then, recalling $Q \in \mathcal{Q}$, we can show $\mathcal{R}(h) = Q^n(U^j_n \ell^{\dagger(j)})$ because

$$\begin{aligned}
Q^n(U^j_n\ell^{\dagger(j)}) &= Q^n(\overline{\sum_{\rho\in\mathfrak{I}^n_j}} \ell^{\dagger(j)}(S_{\rho(1)},\dots,S_{\rho(j)})) \\
&= Q^n(\overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}} \ell^\dagger(S_{\rho\circ\tau(1)},\dots,S_{\rho\circ\tau(d)})) \\
&= Q^n(\overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} \ell(h, \boldsymbol{F}(S^{(1)}_{\rho\circ\tau\circ\pi(1)},\dots,S^{(d)}_{\rho\circ\tau\circ\pi(d)}))) \\
&= \overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} Q^n \ell(h, \boldsymbol{F}(S^{(1)}_{\rho\circ\tau\circ\pi(1)},\dots,S^{(d)}_{\rho\circ\tau\circ\pi(d)})) \\
&= \overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} Q[\ell(h, \boldsymbol{F}(S^{(1)},\dots,S^{(d)}))] \quad (\because Q \in \mathcal{Q}) \\
&= \overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} \mathcal{R}(h) = \mathcal{R}(h).
\end{aligned}$$

Combining this expression with Lemma C.3,

$$\begin{aligned}
\mathcal{R}(h) - \mathbb{E}_{\check{\mathcal{D}}} U^j_n \tilde{\ell}^{(j)} &= Q^n(U^j_n\ell^{\dagger(j)}) - \check{Q}^n(U^j_n\tilde{\ell}^{(j)}) \\
&= \underbrace{Q^n(U^j_n\ell^{\dagger(j)} - U^j_n\tilde{\ell}^{(j)})}_{A} + \underbrace{(Q^n - \check{Q}^n)(U^j_n\tilde{\ell}^{(j)})}_{B}
\end{aligned}$$

Now, $A$ can be bounded from above as

$$\begin{aligned}
A &= Q^n(U^j_n\ell^{\dagger(j)} - U^j_n\tilde{\ell}^{(j)}) \\
&= \overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} Q^n(\ell(h, \boldsymbol{F}(S^{(1)}_{\rho\circ\tau\circ\pi(1)},\dots,S^{(d)}_{\rho\circ\tau\circ\pi(d)})) - \ell(h, \hat{\boldsymbol{F}}(S^{(1)}_{\rho\circ\tau\circ\pi(1)},\dots,S^{(d)}_{\rho\circ\tau\circ\pi(d)}))) \\
&= \overline{\sum_{\rho\in\mathfrak{I}^n_j}}\,\overline{\sum_{\tau\in\mathfrak{G}^d_j}}\,\overline{\sum_{\pi\in\mathfrak{S}_d}} Q(\ell(h, \boldsymbol{F}(S^{(1)},\dots,S^{(d)})) - \ell(h, \hat{\boldsymbol{F}}(S^{(1)},\dots,S^{(d)}))) \quad (\because Q \in \mathcal{Q}) \\
&\le \left\| \ell(h, \boldsymbol{F}(\cdot)) - \ell(h, \hat{\boldsymbol{F}}(\cdot)) \right\|_{L^1(q)}
\end{aligned}$$

Then recalling Assumption C.2, we can bound $B$ from above as

$$\begin{aligned}
B &= (Q^n - \check{Q}^n)(U_n^j \tilde{\ell}^{(j)}) = (Q^n - \check{Q}^n)\left(\overline{\sum_{\rho \in \mathfrak{I}_j^n}} \tilde{\ell}^{(j)}(S_{\rho(1)}, \dots, S_{\rho(j)})\right) \\
&= \overline{\sum_{\rho \in \mathfrak{I}_j^n}} (Q^n - \check{Q}^n)\left(\tilde{\ell}^{(j)}(S_{\rho(1)}, \dots, S_{\rho(j)})\right) \\
&= (Q^j - \check{Q}^j)(\tilde{\ell}^{(j)}(S_1, \dots, S_j)) \quad (\because \text{ symmetry}) \\
&\le B_\ell \int \left| \prod_{i=1}^j q(s_i) - \prod_{i=1}^j \check{q}(s_i) \right| \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&= B_\ell \int \left| \sum_{i=1}^j q(s_1) \cdots q(s_{i-1}) \cdot (q(s_i) - \check{q}(s_i)) \cdot \check{q}(s_{i+1}) \cdots \check{q}(s_j) \right| \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&\le B_\ell \sum_{i=1}^j \int q(s_1) \cdots q(s_{i-1}) \cdot |q(s_i) - \check{q}(s_i)| \cdot \check{q}(s_{i+1}) \cdots \check{q}(s_j) \mathrm{d}s_1 \cdots \mathrm{d}s_j \\
&= B_\ell \sum_{i=1}^j \int |q(s_i) - \check{q}(s_i)| \mathrm{d}s_i = B_\ell \cdot j \left\| q - \check{q} \right\|_{L^1},
\end{aligned}$$

which proves the assertion. □

Now the following lemmas bound each approximation terms in terms of the difference between $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$.

**Lemma C.8** (Loss difference evaluation)**.** *Assume Assumption C.5. Then we have for any $h \in \mathcal{H}$,*

$$\left\| \ell(h, \boldsymbol{F}(\cdot)) - \ell(h, \hat{\boldsymbol{F}}(\cdot)) \right\|_{L^1(q)} \le B_q L_{\ell_{\mathcal{H}}} \sum_{j=1}^d \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}}$$

*Proof.*

$$\begin{aligned}
\left\| \ell(h, \boldsymbol{F}(\cdot)) - \ell(h, \hat{\boldsymbol{F}}(\cdot)) \right\|_{L^1(q)} &= \int |\ell(h, \boldsymbol{F}(s)) - \ell(h, \hat{\boldsymbol{F}}(s))| q(s) \mathrm{d}s \\
&\le B_q \int L_{\ell_{\mathcal{H}}} \left\| \boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s) \right\|_{\ell^2} \mathrm{d}s \\
&\le B_q L_{\ell_{\mathcal{H}}} \int \left\| \boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s) \right\|_{\ell^1} \mathrm{d}s \le B_q L_{\ell_{\mathcal{H}}} \sum_{j=1}^d \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}}.
\end{aligned}$$

□

**Lemma C.9** (Density difference evaluation)**.** *Assume Assumptions C.2, C.3, and C.4. Then we have*

$$\left\| q - \check{q} \right\|_{L^1} \le (L_q L_{\boldsymbol{F}^{-1}} + B_q d C_1') \sum_{j=1}^d \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + B_q \kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$$

*where $C_1'$ and $\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$ are defined as in Lemma C.11.*

*Proof.* Since $\check{q}(s) = q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s)) \left|(\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right|$, we have

$$\begin{aligned}
\|q - \check{q}\|_{L^1} &= \int \left|q(s) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s)) \left|(\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right|\right| \mathrm{d}s \\
&\leq \int |q(s) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s))| \mathrm{d}s + \int q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s)) \left|1 - \left|(\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right|\right| \mathrm{d}s \\
&\leq \underbrace{\int |q(s) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s))| \mathrm{d}s}_{\text{(A)}} + B_q \underbrace{\int \left|1 - (\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right| \mathrm{d}s}_{\text{(B)}}
\end{aligned}$$

where the last line follows from the triangle inequality. Applying Lemma C.10 to (A) and Lemma C.11 to (B) yields the assertion. □

**Lemma C.10.** *Assume Assumptions C.2 and C.3. Then,*

$$\int |q(s) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s))| \mathrm{d}s \leq L_q L_{\boldsymbol{F}^{-1}} \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,1}}$$

*Proof.* We have

$$\begin{aligned}
\int |q(s) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s))| \mathrm{d}s &= \int |q(\boldsymbol{F}^{-1} \circ \boldsymbol{F}(s)) - q(\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}(s))| \mathrm{d}s \\
&\leq L_q L_{\boldsymbol{F}^{-1}} \int \left\|\boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s)\right\|_{\ell^2} \mathrm{d}s \leq L_q L_{\boldsymbol{F}^{-1}} \int \left\|\boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s)\right\|_{\ell^1} \mathrm{d}s \\
&\leq L_q L_{\boldsymbol{F}^{-1}} \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,1}}
\end{aligned}$$

□

**Lemma C.11** (Jacobian difference evaluation)**.** *Assume Assumptions C.2 and C.4. Then,*

$$\int \left|1 - (\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right| \mathrm{d}s \leq d C_1' \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,1}} + \kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}}),$$

*where*

$$\begin{aligned}
C_m' &:= (d+1)^{\frac{7}{2}m-2} \left( (B_{\partial \boldsymbol{F}}^{\infty})^m \left( \sum_{k=1}^{d} \left\|\boldsymbol{F}_k^{-1}\right\|_{C^{1,1}} \right)^m + (B_{\partial \boldsymbol{F}^{-1}}^{\infty})^m \right), \\
\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}}) &:= \sum_{m=2}^{d} \binom{d}{m} C_m' \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,m}}^m .
\end{aligned}$$

*Proof.* Applying Lemma C.12 with $A := (\mathrm{J}\boldsymbol{F}^{-1} \circ \boldsymbol{F})(s) = I$, we obtain

$$\begin{aligned}
\int \left|1 - (\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right| \mathrm{d}s &= \int \left|(\mathrm{J}\boldsymbol{F}^{-1} \circ \boldsymbol{F})(s) - (\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s)\right| \mathrm{d}s \\
&\leq \int \sum_{m=1}^{d} \binom{d}{m} \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^m \mathrm{d}s.
\end{aligned}$$

Now, each term in the integrand can be bounded from above as

$$
\begin{aligned}
&\left\| \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&= \left\| \left( \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) \right) \left( \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) \right) - \left( \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right) \left( \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} \\
&\leq \left\| \left( \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) - \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right) \left( \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} + \left\| \left( \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right) \left( \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right) \right\|_{\mathrm{op}} \\
&\leq \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) - \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\mathrm{op}} \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} + \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\mathrm{op}} \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\qquad (\because \text{ submultiplicativity [89, Section 2.3.2]}) \\
&\leq \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) - \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\mathrm{op}} \left( d \cdot \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) \right\|_{\infty} \right) + \left( d \cdot \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\infty} \right) \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\qquad (\because \|\cdot\|_{\mathrm{op}} \leq d \, \|\cdot\|_{\infty} \text{ [89, Section 2.3.2]}) \\
&\leq \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) - \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\mathrm{op}} \cdot (dB^{\infty}_{\partial \boldsymbol{F}}) \; + \; (dB^{\infty}_{\partial \boldsymbol{F}^{-1}}) \cdot \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}} \\
&\leq dB^{\infty}_{\partial \boldsymbol{F}} \sqrt{d} \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\boldsymbol{F}(s)) - \frac{\mathrm{d}\boldsymbol{F}^{-1}}{\mathrm{d}z}(\hat{\boldsymbol{F}}(s)) \right\|_{\mathrm{op}(1)} + dB^{\infty}_{\partial \boldsymbol{F}^{-1}} \sqrt{d} \left\| \frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}s} - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}s} \right\|_{\mathrm{op}(1)} \\
&\qquad (\because \|\cdot\|_{\mathrm{op}} \leq \sqrt{d} \, \|\cdot\|_{\mathrm{op}(1)} \text{ [89, Section 2.3.1]}) \\
&= d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}} \max_{k \in [d]} \sum_{j=1}^{d} \left| \frac{\partial \boldsymbol{F}_j^{-1}}{\partial z_k}(\boldsymbol{F}(s)) - \frac{\partial \boldsymbol{F}_j^{-1}}{\partial z_k}(\hat{\boldsymbol{F}}(s)) \right| + d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}^{-1}} \max_{k \in [d]} \sum_{j=1}^{d} \left| \frac{\partial \boldsymbol{F}_j}{\partial s_k}(s) - \frac{\partial \hat{\boldsymbol{F}}_j}{\partial s_k}(s) \right| \\
&\leq d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}} \max_{k \in [d]} \sum_{j=1}^{d} \left\| \boldsymbol{F}_j^{-1} \right\|_{C^{1,1}} \left\| \boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s) \right\|_{\ell^2} + d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}^{-1}} \sum_{k=1}^{d} \sum_{j=1}^{d} \left| \frac{\partial \boldsymbol{F}_j}{\partial s_k}(s) - \frac{\partial \hat{\boldsymbol{F}}_j}{\partial s_k}(s) \right| \\
&\leq d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}} \left( \sum_{j=1}^{d} \left\| \boldsymbol{F}_j^{-1} \right\|_{C^{1,1}} \right) \left\| \boldsymbol{F}(s) - \hat{\boldsymbol{F}}(s) \right\|_{\ell^1} + d^{\frac{3}{2}} B^{\infty}_{\partial \boldsymbol{F}^{-1}} \sum_{k=1}^{d} \sum_{j=1}^{d} \left| \frac{\partial \boldsymbol{F}_j}{\partial s_k}(s) - \frac{\partial \hat{\boldsymbol{F}}_j}{\partial s_k}(s) \right| \\
&\qquad (\because \|\cdot\|_{\ell^2} \leq \|\cdot\|_{\ell^1} \text{ [89, Section 2.2.2]}).
\end{aligned}
$$

When powered to $m$, this yields

$$
\begin{aligned}
&\left\| \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{m} \\
&\leq (d^2 + d)^{m-1} \left[ \sum_{j=1}^{d} \left( d^{3/2} B^{\infty}_{\partial \boldsymbol{F}} \left( \sum_{k=1}^{d} \left\| \boldsymbol{F}_k^{-1} \right\|_{C^{1,1}} \right) |\boldsymbol{F}_j(s) - \hat{\boldsymbol{F}}_j(s)| \right)^{m} \right. \\
&\qquad \left. + \sum_{k=1}^{d} \sum_{j=1}^{d} \left( d^{3/2} B^{\infty}_{\partial \boldsymbol{F}^{-1}} \left| \frac{\partial \boldsymbol{F}_j}{\partial s_k}(s) - \frac{\partial \hat{\boldsymbol{F}}_j}{\partial s_k}(s) \right| \right)^{m} \right]
\end{aligned}
$$

where we used $(\sum_{i=1}^{L} a_i)^m \leq L^{m-1}(\sum_{i=1}^{L} a_i^m)$ for $a_i \geq 0$, which follows from Hölder inequality.

Hence,

$$
\begin{aligned}
&\int \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{m} \mathrm{d}s \\
&\leq d^{\frac{5}{2}m-1}(d+1)^{m-1} \Bigg[ \left( B_{\partial \boldsymbol{F}}^{\infty} \sum_{k=1}^{d} \left\| \boldsymbol{F}_k^{-1} \right\|_{C^{1,1}} \right)^{m} \sum_{j=1}^{d} \int |\boldsymbol{F}_j(s) - \hat{\boldsymbol{F}}_j(s)|^m \mathrm{d}s \\
&\qquad\qquad + (B_{\partial \boldsymbol{F}^{-1}}^{\infty})^m \sum_{j=1}^{d} \left( \sum_{k=1}^{d} \int \left| \frac{\partial \boldsymbol{F}_j}{\partial s_k}(s) - \frac{\partial \hat{\boldsymbol{F}}_j}{\partial s_k}(s) \right|^m \mathrm{d}s \right) \Bigg] \\
&\leq (d+1)^{\frac{7}{2}m-2} \Bigg( (B_{\partial \boldsymbol{F}}^{\infty})^m \left( \sum_{k=1}^{d} \left\| \boldsymbol{F}_k^{-1} \right\|_{C^{1,1}} \right)^{m} \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,m}}^{m} \\
&\qquad\qquad + (B_{\partial \boldsymbol{F}^{-1}}^{\infty})^m \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,m}}^{m} \Bigg) \\
&\leq C'_m \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,m}}^{m} .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\int \left| 1 - (\mathrm{J}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}})(s) \right| \mathrm{d}s \\
&\leq \sum_{m=1}^{d} \binom{d}{m} \int \left\| \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \boldsymbol{F}}{\mathrm{d}s}(s) - \frac{\mathrm{d}\boldsymbol{F}^{-1} \circ \hat{\boldsymbol{F}}}{\mathrm{d}s}(s) \right\|_{\mathrm{op}}^{m} \mathrm{d}s \\
&\leq dC'_1 \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,1}} + \underbrace{\sum_{m=2}^{d} \binom{d}{m} C'_m \sum_{j=1}^{d} \left\| \boldsymbol{F}_j - \hat{\boldsymbol{F}}_j \right\|_{W^{1,m}}^{m}}_{\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})}
\end{aligned}
$$

□

Lemma C.11 used the following lemma to bound the difference in Jacobian determinants.

**Lemma C.12** (Determinant perturbation bound [129, Corollary 2.11])**.** *Let $A$ and $E$ be $d \times d$ complex matrices. Then,*

$$
|\det(A) - \det(A+E)| \leq \sum_{m=1}^{d} \binom{d}{m} \|A\|_{\mathrm{op}}^{d-m} \|E\|_{\mathrm{op}}^{m} .
$$

### C.4.8 Comparison of Rademacher Complexities

The following consideration demonstrates how the effective complexity measure $\mathfrak{R}$ in Theorem C.1 resulting from the proposed method may enjoy a relaxed dependence on the input dimensionality compared to the ordinary empirical risk minimization. To do so, we first recall the definition of the ordinary Rademacher complexity and a standard performance guarantee derived based on it.

**Definition C.2** (Ordinary Rademacher complexity)**.** *The ordinary empirical risk minimization finds the candidate hypothesis by*

$$
\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\mathcal{R}}(h),
$$

*where*

$$\hat{\mathcal{R}}(h) := \frac{1}{n}\sum_{i=1}^{n}\ell(h, \boldsymbol{Z}_i) = \frac{1}{n}\sum_{i=1}^{n}\ell(h, \hat{\boldsymbol{F}}(S_i^{(1)}, \ldots, S_i^{(d)}))$$

*and the corresponding ordinary Rademacher complexity* $\mathfrak{R}_{\mathrm{ord}}(\mathcal{H})$ *is*

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) := \frac{1}{n}\mathbb{E}_{\tilde{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i\ell(S_i^{(1)}, \ldots, S_i^{(d)})\right|\right]$$

*where* $\{\sigma_i\}_{i=1}^n$ *are independent uniform sign variables and we denoted*

$$\ell(s^{(1)}, \ldots, s^{(d)}) = \ell(h, \hat{\boldsymbol{F}}(s^{(1)}, \ldots, s_i^{(d)}))$$

*by abuse of notation. This yields the standard Rademacher complexity based bound. Applying Lemma C.5 and using the same proof technique, we have that with probability at least* $1-\delta$*,*

$$\mathcal{R}(\hat{h}) - \mathcal{R}(h^*) \le 2\sup_{h\in\mathcal{H}}|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \le 4\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) + 2B_\ell\sqrt{\frac{\log(2/\delta)}{2n}}.$$

*Therefore, we the corresponding complexity terms are* $\mathfrak{R}_{\mathrm{ord}}(\mathcal{H})$ *and* $d\mathfrak{R}(\mathcal{H})$*. In Remark C.3, we make a comparison of these two complexity measures by taking an example. To recall, the effective Rademacher complexity can be written as, in terms of the notation in this section,*

$$\begin{aligned}\mathfrak{R}(\mathcal{H}) &= \frac{1}{n}\mathbb{E}_{\tilde{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S_2',\ldots,S_d'}\tilde{\ell}(S_i, S_2', \ldots, S_d')\right|\right]\\ &= \frac{1}{n}\mathbb{E}_{\tilde{\mathcal{D}}}\mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i\mathbb{E}_{S_1',\ldots,S_d'}\frac{1}{d}\left(\ell(S_i^{(1)}, S_2'^{(2)}, \ldots, S_d'^{(d)}) + \cdots + \ell(S_1'^{(1)}, S_2'^{(2)}, \ldots, S_i^{(d)})\right)\right|\right]\end{aligned}$$

*Remark* C.3 (Comparison of Radmacher complexities). As an example, consider $\mathcal{H}$, the set of $L$-Lipschitz functions (with respect to infinity norm) on the unit cube $[0,1]^m$. It is well-known that there exists a constant $C > 0$ such that the following holds [286, Example 5.10, p.129] for sufficiently small $t > 0$:

$$\log\mathcal{N}(t, \mathcal{H}, \|\cdot\|_\infty) \asymp (C/t)^m. \tag{C.6}$$

Here, $a(t) \asymp b(t)$ indicates that there exist $k_1, k_2 > 0$ such that, for sufficiently small $t$, it holds that $k_1 b(t) \le a(t) \le k_2 b(t)$. On the other hand, the well-known discretization argument implies that there exist constants $c$ and $B$ such that for any $t \in (0, B]$, the following relation between the Rademacher complexity and the metric entropy holds:

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \le t + c\sqrt{\frac{\log\mathcal{N}(t, \mathcal{H}, \|\cdot\|_\infty)}{n}}. \tag{C.7}$$

Substituting Equation (C.6) into Equation (C.7), we can find that, for large enough $n$, the right hand side is minimized at $t = (c\cdot C^{\frac{m}{2}}\cdot\frac{m}{2})^{\frac{2}{2+m}}\cdot n^{-\frac{1}{2+m}}$. This yields

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \le \tilde{C}\cdot n^{-\frac{1}{2+m}} \tag{C.8}$$

with a new constant $\tilde{C} = \left(c\cdot C^{\frac{m}{2}}\cdot\frac{m}{2}\right)^{\frac{2}{2+m}} + c\cdot C^{\frac{m}{2}}\left(c\cdot C^{\frac{m}{2}}\cdot\frac{m}{2}\right)^{-\frac{m}{2+m}}$. Therefore, by substituting $m = d$ in Equation (C.8), the metric-entropy based bound on the ordinary Rademacher complexity

exhibits exponential dependence on the input dimension as

$$\mathfrak{R}_{\mathrm{ord}}(\mathcal{H}) \leq \mathcal{O}\left(n^{-\frac{1}{2+d}}\right),$$

which is a manifestation of the curse of dimensionality. On the other hand, by following a similar calculation, we can see that the effective Rademacher complexity $\mathfrak{R}(\mathcal{H})$ avoids an exponential dependence on the input dimension $d$. By substituting $m = 1$ in Equation (C.8), we can see

$$d\mathfrak{R}(\mathcal{H}) \leq \mathfrak{R}_{\mathrm{ord}}(\mathcal{H}_1) + \cdots + \mathfrak{R}_{\mathrm{ord}}(\mathcal{H}_d) \leq \mathcal{O}\left(n^{-\frac{1}{3}}\right),$$

where $\mathcal{H}_j := \{\mathbb{E}_{S'_1,\ldots,S'_d} h(S'_1{}^{(1)}, \ldots, S'_{j-1}{}^{(j-1)}, (\cdot)^{(j)}, S'_{j+1}{}^{(j+1)}, \ldots, S'_d{}^{(d)}) : h \in \mathcal{H}\}$. This is because the Lipschitz constant of functions in $\mathcal{H}_j$ is at most $L$ (i.e., the Lipschitz constant does not increase by the marginalization procedure) because for any $h \in \mathcal{H}_j$,

$$\begin{aligned}
&|h(x) - h(y)| \\
&= |\mathbb{E}_{S'_1,\ldots,S'_d}[h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d) - h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d)]| \\
&\leq \mathbb{E}_{S'_1,\ldots,S'_d}|h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d) - h(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d)| \\
&\leq \mathbb{E}_{S'_1,\ldots,S'_d} L \cdot \|(S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, x, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d) - (S'^{(1)}_1, \ldots, S'^{(j-1)}_{j-1}, y, S'^{(j+1)}_{j+1}, \ldots, S'^{(d)}_d)\| \\
&= \mathbb{E}_{S'_1,\ldots,S'_d} L \cdot \|(0, \ldots, 0, x - y, 0, \ldots, 0)\| \\
&= L \cdot |x - y|.
\end{aligned}$$

### C.4.9 Remark on Higher-order Sobolev Norms

Here, we comment on how the term $\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$ is treated as a higher order term of $\boldsymbol{F} - \hat{\boldsymbol{F}}$.

*Remark* C.4 (Higher order Sobolev norms). Let us assume that $\mathrm{supp}(q) \cup \mathrm{supp}(\check{q})$ is contained in a compact set $\tilde{\mathcal{S}}$ for all $\hat{\boldsymbol{F}}$ considered. Note that for $m \in [d]$,

$$\int_{\tilde{\mathcal{S}}} |h(s)|^m \mathrm{d}s \leq (V_{\tilde{\mathcal{S}}})^{\frac{m}{m-d}} \left(\int_{\tilde{\mathcal{S}}} |h(s)|^d \mathrm{d}s\right)^{m/d}$$

by Hölder 's inequality, where we defined $V_{\tilde{\mathcal{S}}} := \int_{\tilde{\mathcal{S}}} 1 \mathrm{d}s$, hence we have $\|\cdot\|_{L^m(\tilde{\mathcal{S}})} \leq (V_{\tilde{\mathcal{S}}})^{\frac{1}{m-d}} \|\cdot\|_{L^d(\tilde{\mathcal{S}})}$. By applying the relation to each term in the definition of $\|\cdot\|_{W^{1,m}}$, we obtain

$$\|f\|^m_{W^{1,m}} \leq (V_{\tilde{\mathcal{S}}})^{\frac{m}{m-d}} \|f\|^m_{W^{1,d}}$$

Thus we obtain

$$\begin{aligned}
\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}}) &= \sum_{m=2}^{d} \binom{d}{m} C'_m \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|^m_{W^{1,m}} \\
&\leq \sum_{m=2}^{d} \binom{d}{m} (V_{\tilde{\mathcal{S}}})^{\frac{m}{m-d}} C'_m \sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|^m_{W^{1,d}} \\
&\leq \mathcal{O}\left(\sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|^2_{W^{1,d}}\right) \quad (\hat{\boldsymbol{F}} \to \boldsymbol{F}).
\end{aligned}$$

By also replacing $\sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,1}}$ with $\sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,d}}$ in Theorem C.1, we can see more clearly that $\kappa_2(\boldsymbol{F} - \hat{\boldsymbol{F}})$ is a higher order term of $\sum_{j=1}^{d} \left\|\boldsymbol{F}_j - \hat{\boldsymbol{F}}_j\right\|_{W^{1,d}}$.

## C.5 Details and Proofs of Theorem 4.1

Here, we provide the proof of Theorem 4.1. We reuse the notation and terminology from Section C.4. We prove the uniformly minimum variance property of the proposed risk estimator under the ideal situation of $\hat{\boldsymbol{F}} = \boldsymbol{F}$.

**Theorem C.2** (Known causal mechanism case)**.** *Assume $\hat{\boldsymbol{F}} = \boldsymbol{F}$. Then, for all $h \in \mathcal{H}$, we have that $\check{\mathcal{R}}(h)$ is the uniformly minimum variance unbiased estimator of $\mathcal{R}(h)$. As a special case, it has a smaller variance than the ordinary empirical risk estimator: $\forall q \in \mathcal{Q}, \forall h \in \mathcal{H}, \mathrm{Var}(\check{\mathcal{R}}(h)) \leq \mathrm{Var}(\hat{\mathcal{R}}(h))$.*

*Proof.* The proof is a result of the following two facts. When $\check{q} \in \mathcal{Q}$, the estimator $\check{\mathcal{R}}(h)$ becomes the generalized U-statistic of the statistical functional Equation (C.3). Furthermore, when $\hat{\boldsymbol{F}} = \boldsymbol{F}$, Equation (C.3) coincides with $\mathcal{R}(h)$ because the approximation error is zero. Since we assume $\hat{\boldsymbol{F}} = \boldsymbol{F}$ we have $\check{q} = q \in \mathcal{Q}$ and hence both of the statements above hold. Therefore, by Lemma C.13, the first assertion of the theorem follows. The last assertion of the theorem follows as a special case as $\hat{\mathcal{R}}(h)$ is an unbiased estimator of $\mathcal{R}(h)$ for $q \in \mathcal{Q}$.

From here, we confirm the above statements by calculation. We first show that $\check{\mathcal{R}}(h)$ is the generalized U-statistic. To see this, observe that the statistical functional Equation (C.3) allows the following expression given $\check{q} \in \mathcal{Q}$:

$$
\begin{aligned}
&\int \tilde{\ell}(s_1, \ldots, s_d) \check{q}(s_1) \cdots \check{q}(s_d) \mathrm{d}s_1 \cdots \mathrm{d}s_d \\
&= \int \overline{\sum_{\pi \in \mathfrak{S}_d}} \ell(h, \hat{\boldsymbol{F}}(s^{(1)}_{\pi(1)}, \ldots, s^{(d)}_{\pi(d)})) \check{q}(s_1) \cdots \check{q}(s_d) \mathrm{d}s_1 \cdots \mathrm{d}s_d \\
&= \int \overline{\sum_{\pi \in \mathfrak{S}_d}} \ell(h, \hat{\boldsymbol{F}}(s^{(1)}_{\pi(1)}, \ldots, s^{(d)}_{\pi(d)})) \prod_j \check{q}^{(j)}(s^{(j)}_1) \cdots \prod_j \check{q}^{(j)}(s^{(j)}_d) \mathrm{d}s_1 \cdots \mathrm{d}s_d \\
&= \int \overline{\sum_{\pi \in \mathfrak{S}_d}} \ell(h, \hat{\boldsymbol{F}}(s^{(1)}_1, \ldots, s^{(d)}_1)) \prod_j \check{q}^{(j)}(s_1) \mathrm{d}s_1 \\
&= \int \ell(h, \hat{\boldsymbol{F}}(s^{(1)}, \ldots, s^{(d)})) \check{q}_1(s^{(1)}) \cdots \check{q}_d(s^{(d)}) \mathrm{d}s^{(1)} \cdots \mathrm{d}s^{(d)}.
\end{aligned}
$$

This is a regular statistical functional of degrees $(1, \ldots, 1)$, where the kernel is $\ell(h, \hat{\boldsymbol{F}}(\cdot, \ldots, \cdot))$. On the other hand, we have

$$
\check{\mathcal{R}}(h) = \frac{1}{n^d} \sum_{(i_1, \ldots, i_d) \in [n]^d} \tilde{\ell}(S_{i_1}, \ldots, S_{i_d}) = \frac{1}{n^d} \sum_{(i_1, \ldots, i_d) \in [n]^d} \ell(h, \hat{\boldsymbol{F}}(S^{(1)}_{i_1}, \ldots, S^{(d)}_{i_d}))
$$

because the summations run through all combinations with replacement. This combined with the fact that $\{S^{(d)}_i\}_{i,d}$ are jointly independent when $\check{q} \in \mathcal{Q}$ yields that $\check{\mathcal{R}}(h)$ is the generalized U-statistic for Equation (C.3).

Now we show that Equation (C.3) coincides $\mathcal{R}(h)$. Given $\hat{\boldsymbol{F}} = \boldsymbol{F}$, we have

$$\begin{aligned}
\mathcal{R}(h) &= \int q(s)\ell(h, \boldsymbol{F}(s))\mathrm{d}s \\
&= \int q(s)\ell(h, \hat{\boldsymbol{F}}(s))\mathrm{d}s \quad (\text{By } \boldsymbol{F} = \hat{\boldsymbol{F}}.) \\
&= \int q_1(s^{(1)}) \cdots q_d(s^{(d)})\ell(h, \hat{\boldsymbol{F}}(s^{(1)}, \ldots, s^{(d)}))\mathrm{d}s^{(1)} \cdots \mathrm{d}s^{(d)} \quad (\text{by } q \in \mathcal{Q}) \\
&= \int \check{q}_1(s^{(1)}) \cdots \check{q}_d(s^{(d)})\ell(h, \hat{\boldsymbol{F}}(s^{(1)}, \ldots, s^{(d)}))\mathrm{d}s^{(1)} \cdots \mathrm{d}s^{(d)} \quad (\text{by } q = \check{q}) \\
&= \int \tilde{\ell}(s_1, \ldots, s_d)\check{q}(s_1) \cdots \check{q}(s_d)\mathrm{d}s_1 \cdots \mathrm{d}s_d. \quad (\because \text{ symmetry})
\end{aligned}$$

□

The following well-known lemma states that a generalized U-statistic is a uniformly minimum variance unbiased estimator.

**Lemma C.13** (Uniformly minimum variance property of a generalized U-statistic)**.** *Let $\theta : \mathcal{Q} \to \mathbb{R}$ be a regular statistical functional with kernel $\psi : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L} \to \mathbb{R}$ [48], i.e.,*

$$\theta(q) = \int \psi((x_1^{(1)}, \ldots, x_{k_1}^{(1)}), \ldots, (x_1^{(L)}, \ldots, x_{k_L}^{(L)})) \prod_{j=1}^{k_1} q_1(x_j^{(1)})\mathrm{d}x_j^{(1)} \cdots \prod_{j=1}^{k_L} q_L(x_j^{(L)})\mathrm{d}x_j^{(L)}.$$

*Given samples $\{x_i^{(l)}\}_{i=1}^{n_l} \overset{i.i.d.}{\sim} q_l (n_l \geq k_l \text{ and } l = 1, \ldots, L)$, let $\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi$ be the corresponding generalized U-statistic*

$$\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi := \frac{1}{\prod_l \binom{n_l}{k_l}} \sum \psi\left(\left(x_{i_1^{(1)}}^{(1)}, \ldots, x_{i_{k_1}^{(1)}}^{(1)}\right), \ldots, \left(x_{i_1^{(L)}}^{(L)}, \ldots, x_{i_{k_L}^{(L)}}^{(L)}\right)\right).$$

*where $\sum$ is a summation over all possible combinations (without replacement) of the indices. Then, $\mathrm{GU}^{(k_1,\ldots,k_L)}_{(n_1,\ldots,n_L)}\psi$ is the uniformly minimum variance unbiased estimator of $\theta$ on $\mathcal{Q}$.*

*Proof.* The assertion can be proved in a parallel manner as the proof of [162, Section 1.1, Lemma B] □

*Remark* C.5 (Relation to the UMVUE property of $\hat{\mathcal{R}}(h)$). The result in Theorem C.2 is not contradictory to the fact that the sample average $\hat{\mathcal{R}}(h)$ is a U-statistic of degree-1 and hence the minimum variance among all unbiased estimator of $\mathcal{R}(h)$ on $\mathcal{P}$, where $\mathcal{P}$ is a set of distributions containing all absolutely continuous distributions [162]. Specifically, $\check{\mathcal{R}}(h)$ is not generally an unbiased estimator of $\mathcal{R}(h)$ on $\mathcal{P} \setminus \mathcal{Q}$, even if $\hat{\boldsymbol{F}} = \boldsymbol{F}$. While $\check{\mathcal{R}}(h)$ satisfies the $d$-sample symmetry condition, the same does not hold for $\hat{\mathcal{R}}(h)$. By restricting the attention to $\mathcal{Q}$, the estimator $\check{\mathcal{R}}(h)$ achieves a smaller variance than $\hat{\mathcal{R}}(h)$.

## C.6 Further Comparison with Related Work

Here, we provide an additional detailed comparison with the related work to complement Section 4.6 of the main text.

### C.6.1 Comparison with Magliacane et al. [176]

Magliacane et al. [176] considered domain adaptation among different interventional states by using SCMs. Their problem setting and ours do not strictly include each other (the two settings

are somewhat complementary), and their assumption may be more suitable for application fields with interventional experiments such as genomics, while ours may be more suited for fields with observational data such as health record analysis [304] or economics [253]. At the methodological level, Magliacane et al. [176] takes a variable selection approach to find a subset so that the conditional distribution is invariant, whereas this chapter takes a data augmentation approach via the estimation of the SEMs (in the reduced form).

The essential assumptions of Magliacane et al. [176] are the existence of a separating set (with small "incomplete information bias") and the identifiability of such a set (yielded from Proposition 1, Assumption 1, and Assumption 2 (iii) in Magliacane et al. [176]). A particularly plausible application conforming to the assumptions is, for example (but not limited to), genomics experiments. Part of the reason is that Assumption 2 (ii) and (iii) of Magliacane et al. [176] are likely to hold for well-targeted experiments. The following is a detailed comparison.

**(1) Modeling assumption and problem setup.** The two problem settings do not strictly include one another, and they are of complementing relations where ours corresponds to the intervention-free case and Magliacane et al. [176] corresponds to the intervention case. If we try to express the problem setting of Magliacane et al. [176] within our formulation, we would be expressing the interventions as alterations to the SEMs. We assume that such alterations do not occur in our setting since our focus is on observational data; therefore, the problem formulation of Magliacane et al. [176] is not a subset of ours. On the other hand, if we try to express our problem setting within the formulation of Magliacane et al. [176], our problem setup would only have $C_1$ as the context variable, and $C_1$ would be a parent of all observed variables, e.g., $C_1$ switches the distribution of $S$ by switching different quantile functions to perform inverse transform. This potentially allows the existence of the effect $C_1 \to Y$ and diverges from Assumption 2 (iii) in Magliacane et al. [176]. Also, even if such an edge does not exist, it is acceptable that there are no separating sets (in the extreme case) if $Y$ is a parent of all $X_i$'s. In this case, conditioning on any of $X_i$'s would result in making $C_1$ and $Y$ dependent. From this consideration, our problem setting is not a subset of that of Magliacane et al. [176], either.

**(2) Plausible applications.** The problem setup of Magliacane et al. [176] is suitable especially for applications in which various experiments are conducted such as genomics [176], whereas our problem setting may be more suitable for some fields with observational data such as health record analysis [304] or economics [253].

**(3) Methodology.** Our proposed method actually estimates the SEMs (though in the reduced-form) and exploits the estimated SEMs in the domain adaptation algorithm. In fact, directly using the estimated SEMs as a tool to realize domain adaptation can be seen as the first attempt to fully leverage the structural causal models in the DA algorithm. On the other hand, Magliacane et al. [176] approaches the problem of domain adaptation via variable selection to find a subset so that the conditional distribution is invariant.

### C.6.2 Comparison with Gong et al. [90]

In the present paper, we assumed an invariance of structural equations between domains. Here, we clarify the difference from a related but different assumption considered by Causal Generative Domain Adaptation Network (CG-DAN; [90]).

**(1) Problem setup.** Gong et al. [90] presumes the *anticausal* scenario (i.e., $Y$ is the cause of $X$) and that $X$ given $Y$ follows a structural equation model, whereas this chapter considers more

general SEMs of $X$ and $Y$.

**(2) Theoretical justification.** The approach of Gong et al. [90] does not have a theoretical guarantee in terms of the identifiability of $\boldsymbol{F}$, i.e., there has been no known theoretical condition under which the learned generator is applicable across different domains. On the other hand, our method enjoys a strong theoretical justification of nonlinear ICA including the identifiability of $\boldsymbol{F}$ under known theoretical conditions.

**(3) Methodology.** The method of Gong et al. [90] estimates the GCM of $X$ given $Y$ using source domain data and uses it to design a generator neural network. On the other hand, we more directly exploit the estimated reduced-form SEM in the method.

### C.6.3 Comparison with Arjovsky et al. [8]

Arjovsky et al. [8] proposed *invariant risk minimization* (IRM) for the *out-of-distribution (OOD) generalization* problem. The IRM approach tries to learn a feature extractor that makes the optimal predictor invariant across domains, and its theoretical validity is argued based on SCMs. Here, we compare it with the present work in terms of the problem setup, theoretical justification, and the methodology.

**(1) Basic assumption and problem setup.** The OOD generalization problem tackled in Arjovsky et al. [8] assumes no access to the target domain data. In this respect, the problem is different and intrinsically more difficult than the one considered in this paper, where a small labeled sample from the target domain is assumed to be available. In order to solve the OOD generalization problem, in a nutshell, Arjovsky et al. [8] essentially assumes the existence of a feature extractor that *elicits an invariant predictor*, i.e., one that makes the optimal predictors of the different domains to be identical after the feature transformation. This can be seen as a variant of the representation learning approach for domain adaptation where we assume there exists $\mathcal{T}$ such that $p(Y|\mathcal{T}(X))$ is invariant across domains. Indeed, for example, when the loss function is the cross-entropy, the condition corresponds to the invariance of $P(Y|\mathcal{T}(X))$ across domains [8]. More technically, in addition, [8, Definition 7(ii)] requires the condition $\mathbb{E}_1[Y|\mathrm{Pa}(Y)] = \mathbb{E}_2[Y|\mathrm{Pa}(Y)]$, which can be violated when the latent factors corresponding to $Y$ have different distributions across domains. On the other hand, our assumption can be seen as the existence of a feature extractor that can simultaneously estimate the independent components in all domains, which does not necessarily imply the existence of a common feature transformer that induces a unique optimal predictor.

**(2) Theoretical justification.** Arjovsky et al. [8] formulated a condition under which the IRM principle leads to an appropriate predictor for OOD generalization, but only under a certain linearity assumption which is essentially a relaxation of linear SEMs. Furthermore, in the theoretical guarantee, the feature extractor is restricted to be linear. In addition, Arjovsky et al. [8] only provides the population-level analysis that the solution of the IRM objective formulated using the underlying distributions enjoys OOD generalization, and it does not discuss the condition under which the ideal feature extractor can be properly estimated by the empirical IRM. The requirement for the strong assumption of linearity likely stems from the intrinsic difficulty of the OOD problem in Arjovsky et al. [8], namely, its formulation does not assume specific types of interventions. On the other hand, our method enjoys a stronger theoretical guarantee of an excess risk bound without such parametric assumptions on the models or the data-generating process, by focusing on the case that the causal mechanisms are indifferent across the domains.

**(3) Methodology.** The methodology of IRM estimates a single predictor that generalizes well to all domains by finding a feature extractor that makes the predictor optimal in all domains. The approach shares the same spirit as the representation learning approaches to domain adaptation, which try to find a feature extractor that induces invariant conditional distributions, such as transfer component analysis [196]. On the other hand, our method estimates the SEMs (in the reduced-form) and exploits it to make the training on the few target domain data more efficient through data augmentation.

# Appendix D

# Appendices for Chapter 5

Table D.1 summarizes the abbreviations and the symbols used in the chapter. Figure D.1 depicts the relations among the notions of universalities appearing in Chapter 5 and how they are connected by the sections in this Appendix.

Table D.1: List of abbreviations and symbols used in the chapter.

| Abbreviation/Notation | Meaning |
|---|---|
| CF-INN | Invertible neural networks based on coupling flows |
| IAF | Inverse autoregressive flow |
| DSF | Deep sigmoidal flow |
| SoS | Sum-of-squares polynomial flow |
| MLP | Multi-layer perceptron |
| CF, $h_{k,\tau,\theta}$ | Coupling flow |
| ACF, $\Psi_{k,s,t}$ | Affine coupling flow |
| $\mathcal{H}$ | Set of functions from $\mathbb{R}^{d-1}$ to $\mathbb{R}$ |
| $\mathcal{H}$-ACF, $\Psi_{d-1,s,t}$ | $\mathcal{H}$-single-coordinate ACFs ($s, t \in \mathcal{H}$) |
| Aff | Set of invertible affine transformations |
| GL | Set of invertible linear transformations |
| $\mathcal{G}$ | Generic notation for a set of invertible functions |
| $\mathrm{INN}_{\mathcal{G}}$ | Set of invertible neural networks based on $\mathcal{G}$ |
| $\mathcal{D}^2$ | $C^2$-diffeomorphisms with $C^2$-diffeomorphic domains |
| $\mathcal{T}^\infty$ | $C^\infty$-increasing triangular maps |
| $\mathcal{S}_{\mathrm{c}}^r$ | $C^r$-single-coordinate transformations |
| $\mathrm{Diff}_{\mathrm{c}}^2$ | Group of compactly-supported $C^2$-diffeomorphisms (on $\mathbb{R}^d$) |
| $\|\cdot\|$ | Euclidean norm |
| $\|\cdot\|_{\mathrm{op}}$ | Operator norm |
| $\|\cdot\|_{p,K}$ | $L^p$-norm ($p \in [1,\infty)$) on a subset $K \subset \mathbb{R}^d$ |
| $\|\cdot\|_{\mathrm{sup},K}$ | Supremum norm on a subset $K \subset \mathbb{R}^d$ |
| $\mathbf{1}_A(\cdot)$ | Indicator (characteristic) function of $A$ |

## D.1 Proof of Lemma 5.1: From $L^p$-universality to Distributional Universality

Here, we prove Lemma D.1, which corresponds to Lemma 5.1 in the main text.

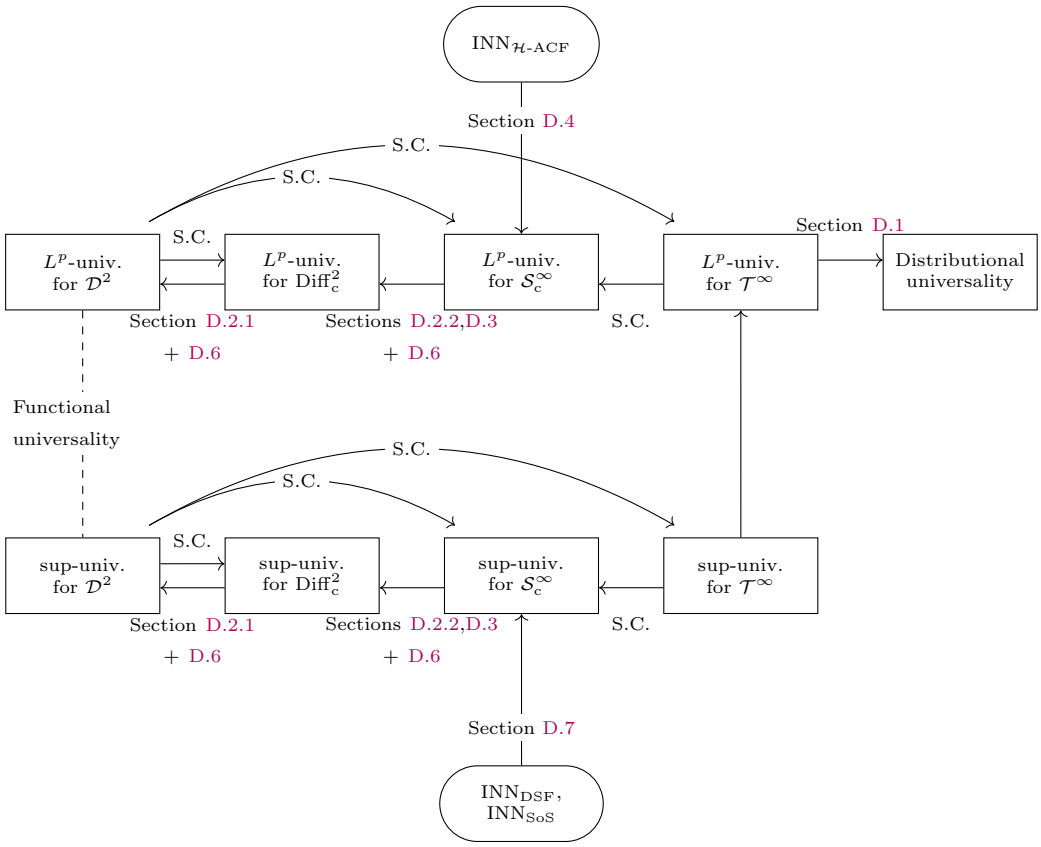



Figure D.1: Informal diagram of the relations among propositions and lemmas connecting them. Here, $p \in [1, \infty)$. *S.C.* stands for "special case" and indicates that the notion of universality implies the other as a special case. *DSF* stands for *deep sigmoidal flow*, and *SoS* stands for *sum-of-squares polynomial flow*.

First, note that the larger $p$, the stronger the notion of $L^p$-universality: if a model $\mathcal{M}$ is an $L^p$-universal approximator for $\mathcal{F}$, it is also an $L^q$-universal approximator for $\mathcal{F}$ for all $1 \leq q \leq p$. In particular, we use this fact with $q = 1$ in the following proof.

**Lemma D.1** (Lemma 5.1 in the main text)**.** *Let $p \in [1, \infty)$. Suppose $\mathcal{M}$ is an $L^p$-universal approximator for $\mathcal{T}^\infty$. Then $\mathcal{M}$ is a distributional universal approximator.*

*Proof.* We denote by $\mathrm{BL}_1$ the set of bounded Lipschitz functions $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $\|f\|_{\sup,\mathbb{R}^d} + L_f \leq 1$, where $L_f$ denotes the Lipschitz constant of $f$. Let $\mu, \nu$ be absolutely continuous probability measures, and take any $\varepsilon > 0$. By Theorem 11.3.3 in [65], it suffices to show that there exists $g \in \mathcal{M}$ such that

$$\beta(g_*\mu, \nu) := \sup_{f \in \mathrm{BL}_1} \left| \int_{\mathbb{R}^d} f \, dg_*\mu - f \, d\nu \right| < \varepsilon.$$

Let $p, q \in L^1(\mathbb{R}^d)$ be the density functions of $\mu$ and $\nu$ respectively. Let $\phi \in L^1(\mathbb{R}^d)$ be a positive $C^\infty$-function such that $\int_{\mathbb{R}^d} \phi(x)dx = 1$ (for example, Gaussian distribution), and for $t > 0$, put $\phi_t(x) := t^{-d}\phi(x/t)$. We define $\mu_t := \phi_t * pdx$ and $\nu_t := \phi_t * qdx$. Since both $\|\phi_t * p - p\|_{1,\mathbb{R}^d}$ and $\|\phi_t * q - q\|_{1,\mathbb{R}^d}$ converge to 0 as $t \to 0$, there exists $t_0 > 0$ such that for any continuous map $G : \mathbb{R}^d \to \mathbb{R}^d$,

$$\left| \int_{\mathbb{R}^d} f \, dG_*\mu_{t_0} - f \, dG_*\mu \right| < \frac{\|f\|_{\sup,\mathbb{R}^d}\, \varepsilon}{5}, \quad \left| \int_{\mathbb{R}^d} f \, d\nu_{t_0} - f \, d\nu \right| < \frac{\|f\|_{\sup,\mathbb{R}^d}\, \varepsilon}{5}.$$

By using Lemma D.2 below, there exists $T \in \mathcal{T}^\infty$ such that $T_*\mu_{t_0} = \nu_{t_0}$. Let $K \subset \mathbb{R}^d$ be a compact subset such that

$$1 - \mu_{t_0}(K) < \frac{\varepsilon}{5}.$$

By the assumption, there exists $g \in \mathcal{M}$ such that

$$\int_K |T(x) - g(x)|dx < \frac{\varepsilon}{5 \sup_{x \in K} |\phi_{t_0} * p(x)|}.$$

Thus for any $f \in \mathrm{BL}_1$, we have

$$\begin{aligned}
&\left| \int_{\mathbb{R}^d} f \, dg_*\mu - f \, d\nu \right| \\
&\leq \left| \int_{\mathbb{R}^d} f \, dg_*\mu_{t_0} - f \, dg_*\mu \right| + \left| \int_{\mathbb{R}^d} f \, d\nu_{t_0} - f \, d\nu \right| \\
&\quad + \left| \int_{\mathbb{R}^d \setminus K} f \circ T \, d\mu_{t_0} \right| + \left| \int_{\mathbb{R}^d \setminus K} f \circ g \, d\mu_{t_0} \right| + \int_K |f(T(x)) - f(g(x))| \, d\mu_{t_0}(x) \\
&< \frac{\|f\|_{\sup,\mathbb{R}^d}\, \varepsilon}{5} + \frac{\|f\|_{\sup,\mathbb{R}^d}}{5} + \frac{\|f\|_{\sup,\mathbb{R}^d}\, \varepsilon}{5} + \frac{\|f\|_{\sup,\mathbb{R}^d}\, \varepsilon}{5} + \frac{L_f \varepsilon}{5} \\
&\leq \varepsilon,
\end{aligned}$$

where $L_f$ is the Lipschitz constant of $f$. Here we used $\|f\|_{\sup,\mathbb{R}^d} + L_f \leq 1$. Therefore, we have $\beta(g_*\mu, \nu) < \varepsilon$. □

The following lemma is essentially due to [120].

**Lemma D.2.** *Let $\mu$ be a probability measure on $\mathbb{R}^d$ with a $C^\infty$ density function $p$. Let $U := \{x \in \mathbb{R}^d : p(x) > 0\}$. Then there exists a diffeomorphism $T : U \to (0,1)^d$ such that its Jacobian is upper triangular matrix with positive diagonal, and $T_*\mu = \mathrm{U}(0,1)^d$. Here, $\mathrm{U}(0,1)^d$ is the uniform distribution on $[0,1]^d$.*

*Proof.* Let $q_i(x_1,\dots,x_i) := \int_{\mathbb{R}^{d-i}} p(x_1,\dots,x_{i+1},\dots,x_d)\,dx_{i+1}\dots dx_d$. Then we define $T : U \to (0,1)^d$ by

$$T(x_1,\dots,x_d) := \left(\int_{-\infty}^{x_i} \frac{q_i(x_1,\dots,x_{i-1},y)}{q_{i-1}(x_1,\dots,x_{i-1})}dy\right)_i.$$

Then we see that $T$ is a diffeomorphism and its Jacobian is upper triangular with positive diagonal elements. Moreover, by a direct computation, we have $T_* d\mu = U(0,1)$. □

We include a proof for the statement that that any probability measure on $\mathbb{R}^m$ is arbitrarily approximated by an absolutely continuous probability measure in the weak convergence topology:

**Lemma D.3.** *Let $\mu$ be an arbitrary probability measure of $\mathbb{R}^m$. Then there exists a sequence $\{\mu_n\}_{n=1}^\infty$ of absolutely continuous probability measures such that $\mu_n$ weakly converges to $\mu$.*

*Proof.* Let $\phi$ be a positive bounded $C^\infty$ function such that $\int_{\mathbb{R}^m}\phi(x)dx = 1$. For $t > 0$, put $\phi_t(x) := t^{-m}\phi(x/t)$. We define

$$w_t(x) = \int_{\mathbb{R}^m} \phi_t(x-y)d\mu(y).$$

We prove the absolutely continuous measure $w_t dx$ weakly converges to $\mu$ as $t \to 0$. In fact, for any bounded continuous function $f$, we have

$$\begin{aligned}\left|\int_{\mathbb{R}^m} f w_t dx - \int f d\mu\right| &= \left|\int\int_{\mathbb{R}^m} (f(y+tx)-f(y))\,\phi(x)dxd\mu(y)\right| \\ &\le \int\int_{\mathbb{R}^m} |f(y+tx)-f(y)|\phi(x)dxd\mu(y).\end{aligned}$$

Since $f$ is bounded and $\phi$ is absolutely integrable, by the dominated convergence theorem, as $t \to 0$, we have

$$\int_{\mathbb{R}^m} f w_t dx \to \int f d\mu,$$

namely, $w_t dx$ weakly converges to $\mu$. □

## D.2 Proof of Theorem 5.1: Equivalence of Universality Properties

In this section, we provide the proof details of Theorem 5.1 in the main text. Section D.2.1 explains the reduction from $\mathcal{D}^2$ to $\mathrm{Diff}^2_{\mathrm{c}}$, and Section D.2.2 explains the reduction from $\mathrm{Diff}^2_{\mathrm{c}}$ to $\mathcal{S}^\infty_{\mathrm{c}}$ and permutations of variables.

Here, we formally repost the proof of Theorem 5.1 which has been essentially completed in Section 5.4.1.

*Proof of Theorem 5.1.* Since we have $\mathcal{S}^\infty_{\mathrm{c}} \subset \mathcal{T}^\infty \subset \mathcal{D}^2$, it is sufficient to prove that the universal approximation properties for $S^\infty$ imply those for $\mathcal{D}^2$. Therefore, we focus on describing the reduction from $\mathcal{D}^2$ to $\mathcal{S}^\infty_{\mathrm{c}}$. First, by combining Lemma D.9 with the $L^p$-universality (in the case A) or the sup-universality (in the case B) of $\mathrm{INN}_{\mathcal{G}}$ for $\mathcal{S}^\infty_{\mathrm{c}}$, we obtain the $L^p$-universal (resp. sup-universal) approximation property for $\mathcal{S}^2_{\mathrm{c}}$. Now, in light of Lemma D.4 and Theorem D.1, we obtain the assertion of Theorem 5.4 in the main text, i.e., for any $f \in \mathcal{D}^2$ and compact subset $K \subset U_f$, there exist $W_1,\dots,W_r \in \mathrm{Aff}$ and $\tau_1,\dots,\tau_r \in \mathcal{S}^2_{\mathrm{c}}$ and $b \in \mathbb{R}^d$ such that $f(x) = W_1 \circ \tau_1 \circ \cdots \circ W_r \circ \tau_r(x)$ for all $x \in K$. Given this decomposition, we combine the $L^p$-universality (in the case A) or the sup-universality (in the case B) of $\mathrm{INN}_{\mathcal{G}}$ for $\mathcal{S}^2_{\mathrm{c}}$ with Proposition D.3 to obtain the assertion of Theorem 5.1. □

### D.2.1 From $\mathcal{D}^2$ to $\mathrm{Diff}^2_{\mathrm{c}}$

In this section, we describe how the approximation of $\mathcal{D}^2$ is reduced to that of $\mathrm{Diff}^2_{\mathrm{c}}$ when we are only concerned with its approximation on a compact set.

**Lemma D.4.** *Let $f\colon U \to \mathbb{R}^d$ be an element of $\mathcal{D}^2$, and let $K \subset U$ be a compact set. Then, there exists $h \in \mathrm{Diff}^2_{\mathrm{c}}$ and an affine transform $W \in \mathrm{Aff}$ such that*

$$W \circ h|_K = f|_K.$$

*Proof of Lemma D.4.* We denote the injections of $U$ and $f(U)$ into $\mathbb{R}^d$ by $\iota_1\colon U \hookrightarrow \mathbb{R}^d$ and $\iota_2\colon f(U) \hookrightarrow \mathbb{R}^d$, respectively. Since $U$ is $C^2$-diffeomorphic to $\mathbb{R}^d$ and $f$ is $C^2$-diffeomorphic, $f(U)$ is also $C^2$-diffeomorphic to $\mathbb{R}^d$. By applying Theorem 3.3 in [23] to $\iota_1 \circ f^{-1}|_{f(U)}\colon f(U) \to \mathbb{R}^d$ and the injection $\iota_2$, we can obtain diffeomorphisms $F_1\colon f(U) \to \mathbb{R}^d$ and $F_2\colon f(U) \to \mathbb{R}^d$ such that $F_1|_{f(K)} = f^{-1}|_{f(K)}$ and $F_2|_{f(K)} = \mathrm{Id}_{f(K)}$, where $\mathrm{Id}_{f(K)}$ denotes the identity map on $f(K)$. Let $F := F_2 \circ F_1^{-1}\colon \mathbb{R}^d \to \mathbb{R}^d$. By definition, we have $F|_K = f|_K$.

Take a sufficiently large open ball $B$ centered at 0 such that $K \subset B$. Let $W \in \mathrm{Aff}$ such that $W(x) = DF^{-1}(0)(x - F(0))$. Then by Lemma D.5 below, we conclude that there exists a compactly supported diffeomorphism $h\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $W \circ h|_K = F|_K = f|_K$. □

**Lemma D.5.** *Let $B_r \subset \mathbb{R}^d$ be an open ball of radius $r$ with origin 0, and let $f : B_r \to f(B_r) \subset \mathbb{R}^d$ be a $C^2$-diffeomorphism onto its image such that $f(0) = 0$ and $Df(0) = I$. Let $\varepsilon \in (0, r/2)$. Then there exists $h \in \mathrm{Diff}^2_{\mathrm{c}}$ such that $f(x) = h(x)$ for any $x \in B_{r-\varepsilon}$.*

*Proof.* Put $\delta := \varepsilon/(2r - \varepsilon)$, and define $I_\delta := (-1 - \delta, 1 + \delta)$. We define $F : B_{r-\varepsilon/2} \times I_\delta \to \mathbb{R}^d$ by

$$F(x,t) := \begin{cases} \frac{f(tx)}{t} & \text{if } t \neq 0, \\ x & \text{if } t = 0. \end{cases}$$

Let $U := F(B_{r-\varepsilon/2})$ and let $F^\dagger : U \times I_\delta \to B_{r-\varepsilon/2}$ such that $F^\dagger(F(x,t)) = x$ for any $(x,t) \in U$. Fix a compactly supported function on $\mathbb{R}^d \times I_\delta$ such that for $(x,t) \in F\big(\overline{B_{r-\varepsilon}} \times [-1,1]\big)$, $\phi(x,t) = 1$, and for $(x,t) \notin U$ $\phi = 0$. Then we define $H : \mathbb{R}^d \times I_\delta \to \mathbb{R}^d$ by

$$H(x,t) := \phi(x,t)\frac{\partial F}{\partial t}(F^\dagger(x,t),t).$$

Since $f$ is $C^2$ diffeomorphism, there exists $L > 0$ such that for any $t \in I_\delta$, $\|H(x,t) - H(y,t)\| < L\|x - y\|$ with $x, y \in \mathbb{R}^d$. Thus the differential equation

$$\frac{dz}{dt} = H(z,t), \quad z(0) = x$$

has a unique solution $\phi_x(t)$. Then $h(x) := \phi_x(1)$ is the desired extension. □

Here, we remark that Lemma D.5 is a modified version of Lemma D.1 in Bernard et al. [23], with a correction to make it explicit that the extended diffeomorphism is compactly supported. Their Lemma D.1 does not explicitly state that it is compactly supported, but by Theorem 1.4 in Section 8 of Hirsch [104], it can be shown that the diffeomorphism is actually compactly supported.

### D.2.2 From $\mathrm{Diff}^2_{\mathrm{c}}$ to $\mathcal{S}^\infty_{\mathrm{c}}$ and Permutations

The goal of this section is to show Theorem D.1, which reduces the approximation problem of $\mathrm{Diff}^2_{\mathrm{c}}$ to that of $\mathcal{S}^2_{\mathrm{c}}$, and Lemma D.9, which reduces from $\mathcal{S}^2_{\mathrm{c}}$ to $\mathcal{S}^\infty_{\mathrm{c}}$.

**Theorem D.1.** *Let $f \in \mathrm{Diff}^2_{\mathrm{c}}$. Then there exist $\tau_1, \dots, \tau_n \in \mathcal{S}^2_{\mathrm{c}} \cap \mathrm{Diff}^2_{\mathrm{c}}$, and permutations of variables $\sigma_1, \dots, \sigma_n \in \mathfrak{S}_d$, such that*

$$f = \tau_1 \circ \sigma_1 \circ \cdots \circ \tau_n \circ \sigma_n.$$

*Proof.* Combining Corollary D.1, Lemma D.6, and Lemma D.7, we have the assertion. □

We defer the statement and proof of Corollary D.1, which describes the key properties of $\mathrm{Diff}^2_{\mathrm{c}}$, to Section D.3. In the remainder of this section, we describe Lemma D.6, Lemma D.7, and Lemma D.9. First, Lemma D.6 claims that the nearly-Id elements necessarily satisfy the condition of Lemma D.7 below.

**Lemma D.6.** *Let $A = (a_{i,j})_{i,j=1,\dots,d}$ be a matrix. If $\|A - I_d\|_{\mathrm{op}} < 1$, then for $k = 1, \dots, d$, the $k$-th trailing principal submatrix $A_k := (a_{i+k-1,j+k-1})_{i,j=1,\dots,d-(k-1)}$ of $A$ is invertible. Here $I_d$ is a unit matrix of degree $d$.*

*Proof.* Let $v \in \mathbb{R}^{d-k+1}$ with $\|v\| = 1$, and put $w := (0, \dots, 0, v) \in \mathbb{R}^d$. Then we have $1 > \|(A - I_d)w\|^2 \geq \|(A_k - I_k)v\|^2$. Thus $\|A_k - I_k\| < 1$. Since $\sum_{r=0}^{\infty}(I_k - A_k)^r$ absolutely converges, and it is identical to the inverse of $A_k$, we have that $A_k$ is invertible. □

We apply the following lemma together with Lemma D.6 to decompose nearly-Id elements into $\mathcal{S}^2_{\mathrm{c}}$ and permutations. For $a \in \mathbb{N}$, we denote the set of $a$-by-$a$ real-valued matrices by $M(a, \mathbb{R})$.

**Lemma D.7.** *Let $r$ be a positive integer and $f \colon \mathbb{R}^d \to \mathbb{R}^d$ a compactly supported $C^r$-diffeomorphism. We write $f = (f_1, \dots, f_d)$ with $f_i \colon \mathbb{R}^d \to \mathbb{R}$. For $k \in [d]$, let $\Delta^f_k(\boldsymbol{x}) \in M(d - (k-1), \mathbb{R})$ be the $k$-th trailing principal submatrix of Jacobian matrix of $f$, whose $(i, j)$ component is given by $\left(\frac{\partial f_{i+k-1}}{\partial x_{j+k-1}}(\boldsymbol{x})\right)$ $(i, j = 1, \cdots, d - (k-1))$. We assume*

$$\det \Delta^f_k(x) \neq 0 \text{ for any } k \in [d] \text{ and } x \in \mathbb{R}^d.$$

*Then there exist compactly supported $C^r$-diffeomorphisms $F_1, \dots, F_d : \mathbb{R}^d \to \mathbb{R}^d$ in the forms of*

$$F_i(\boldsymbol{x}) := (x_1, \dots, x_{i-1}, h_i(\boldsymbol{x}), x_{i+1}, \dots, x_d)$$

*for some $h_i \colon \mathbb{R}^d \to \mathbb{R}$ such that the identity holds:*

$$f = F_1 \circ \cdots \circ F_d.$$

*Proof.* The proof is based on induction. Suppose that $f$ is in the form of $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}), x_{m+1}, \dots, x_d)$. By means of induction with respect to $m$, we prove that there exist compactly supported $C^r$-diffeomorphisms $F_1, \dots, F_m : \mathbb{R}^d \to \mathbb{R}^d$ in the forms of $F_i(\boldsymbol{x}) := (x_1, \dots, x_{i-1}, h_i(\boldsymbol{x}), x_{i+1}, \dots, x_d)$ for some $h_i : \mathbb{R}^d \to \mathbb{R}$ such that $f = F_1 \circ \cdots \circ F_m$.

In the case of $m = 1$, the above is clear. Assume that the statement is true in the case of any $k < m$. Define

$$F(x_1, \dots, x_d) := (x_1, \dots, x_{m-1}, f_m(\boldsymbol{x}), x_{m+1}, \dots, x_d),$$
$$\tilde{f} := f \circ F^{-1}.$$

Note that $F$ is a compactly supported $C^r$-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. In fact, compactly supportedness and surjectivity of $F$ comes from the compactly supportedness of $f$. Moreover, since we have $\det DF_x = \frac{\partial f_m}{\partial x_m}(x) \neq 0$ for any $x \in \mathbb{R}^d$ by the assumption on $f$, $F$ is injective and is a $C^r$-diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$ by inverse function theorem. Therefore, $\tilde{f}$ is also a $C^r$-diffeomorphism

from $\mathbb{R}^d$ to $\mathbb{R}^d$. We show that $\tilde{f}$ is of the form $\tilde{f}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), x_m, \cdots, x_d)$ for some $C^r$-functions $g_i \colon \mathbb{R}^d \to \mathbb{R}$ $(i = 1, \cdots, m-1)$ satisfying $\det \Delta_k^{\tilde{f}}(x) \neq 0$ for any $x \in \mathbb{R}^d$ and $k \in [d]$. From Lemma D.8, there exist $g_i, h \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m)$ such that

$$f^{-1}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_m(\boldsymbol{x}), x_{m+1}, \cdots, x_d)$$
$$F^{-1}(\boldsymbol{x}) = (x_1, \cdots, x_{m-1}, h(\boldsymbol{x}), x_{m+1}, \cdots, x_d).$$

Then we have

$$\begin{aligned}\tilde{f}^{-1}(\boldsymbol{x}) = F \circ f^{-1}(\boldsymbol{x}) &= (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), f_m(f^{-1}(\boldsymbol{x})), x_{m+1}, \cdots, x_d) \\ &= (g_1(\boldsymbol{x}), \cdots, g_{m-1}(\boldsymbol{x}), x_m, \cdots, x_d).\end{aligned}$$

Therefore, from Lemma D.8, $\tilde{f}$ is of the following form

$$\tilde{f}(x) = f \circ F^{-1}(x) = (f_1 \circ F^{-1}(x), \cdots, f_{m-1} \circ F^{-1}(x), x_m, \cdots, x_d).$$

Moreover, by the form of $F^{-1}$ and $f$, we have $D\tilde{f}(x) = Df(F^{-1}(x)) \circ DF^{-1}(x)$ and

$$Df = \begin{pmatrix} A & \\ & I \end{pmatrix}, \quad D(F^{-1}) = \begin{pmatrix} I_{m-1} & & \\ \frac{\partial h}{\partial x_1} & \cdots & \frac{\partial h}{\partial x_d} \\ & & I_{d-m} \end{pmatrix}$$

for some $A \in M(m, \mathbb{R})$ with all the trailing principal minors nonzero. Therefore, we obtain $\det \Delta_k^f(x) \neq 0$ for any $x \in \mathbb{R}^d$ and $k \in [d]$. Here, by the assumption of the induction, there exist compactly supported $C^r$-diffeomorphisms $F_i \colon \mathbb{R}^d \to \mathbb{R}^d$ and $h_i \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m-1)$ such that

$$\tilde{f} = F_1 \circ \cdots \circ F_{m-1}, \ F_i(\boldsymbol{x}) = (x_1, \cdots x_{i-1}, h_i(x), x_{i+1}, \cdots, x_d).$$

Thus $f = \tilde{f} \circ F$ has a desired form. □

**Lemma D.8.** *Let $r$ be a positive integer and $f \colon \mathbb{R}^d \to \mathbb{R}^d$ $C^r$-diffeomorphism of the form*

$$f(\boldsymbol{x}) := (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x}), x_{m+1}, \cdots, x_d),$$

*where $f_i \colon \mathbb{R}^d \to \mathbb{R}$ belongs to $C^r(\mathbb{R}^d)$ $(i = 1, \cdots, m)$. Then the inverse map $f^{-1}$ becomes of the form*

$$f^{-1}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_m(\boldsymbol{x}), x_{m+1}, \cdots x_d),$$

*where $g_i : \mathbb{R}^d \to \mathbb{R}$ belongs to $C^r(\mathbb{R}^d)$ for $i = 1, \cdots, m$.*

*Proof.* We write $f^{-1}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \cdots, h_d(\boldsymbol{x}))$, where $h_i \in C^r(\mathbb{R}^d)$ $(i = 1, \cdots, d)$. Then by the definition of the inverse map, the identity

$$(x_1, \cdots, x_d) = f \circ f^{-1}(\boldsymbol{x}) = (f_1(h_1(\boldsymbol{x})), \cdots, f_m(h_m(\boldsymbol{x})), h_{m+1}(\boldsymbol{x}), \cdots, h_d(\boldsymbol{x}))$$

holds for any $\boldsymbol{x} \in \mathbb{R}^d$, which implies that we obtain $h_i(x) = x_i$ $(i = m+1, \cdots, d)$. This completes the proof of the lemma. □

The following Lemma D.9 is used in the main text in reducing the approximation problem from $\mathcal{S}_c^2$ to $\mathcal{S}_c^\infty$. We say that $f \colon \mathbb{R}^d \to \mathbb{R}$ is a *locally $L^p$-function* if $\int_K |f(x)|^p dx < \infty$ holds for any compact set $K \subset \mathbb{R}^d$.

**Definition D.1** (Last-increasing)**.** *We say that a map $f : \mathbb{R}^d \to \mathbb{R}$ is* last-increasing *if, for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the function $f(a_1, \dots, a_{d-1}, x)$ is strictly increasing with respect to $x$.*

**Lemma D.9.** *Let $\tau \colon \mathbb{R}^d \to \mathbb{R}$ be a last-increasing locally $L^p$-function. Then for any compact subset $K \subset \mathbb{R}^d$ and any $\varepsilon > 0$, there exists a last-increasing $C^\infty$-function $\tilde{\tau} \colon \mathbb{R}^d \to \mathbb{R}$ satisfying*

$$\|\tau - \tilde{\tau}\|_{p,K} < \varepsilon.$$

*Moreover, if $\tau$ is continuous, there exists a last-increasing $C^\infty$-function $\tilde{\tau}$ such that*

$$\|\tau - \tilde{\tau}\|_{\sup,K} < \varepsilon.$$

*Proof.* Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a compactly supported non-negative $C^\infty$-function with $\int |\phi(x)|dx = 1$ such that for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the function $\phi(a_1, \dots, a_{d-1}, x)$ of $x$ is even and decreasing on $\{x > 0 : \phi(a_1, \dots, a_{d-1}, x) > 0\}$. For $t > 0$, we define $\phi_t(x) := t^{-d}\phi(x/t)$. Then we see that $\tau_t := \phi_t * \tau$ is a $C^\infty$-function. We take any $\boldsymbol{a} \in \mathbb{R}^{d-1}$. We verify that $\tau_t(\boldsymbol{a}, x_d)$ is strictly increasing with respect to $x_d$. Take any $x_d, x_d' \in \mathbb{R}$ satisfying $x_d > x_d'$. Since $\tau$ is strictly increasing, we have

$$\tau_t(\boldsymbol{a}, x_d) - \tau_t(\boldsymbol{a}, x_d') = \int_{\mathbb{R}^d} \phi_t(x)(\tau((\boldsymbol{a}, x_d) - x) - \tau((\boldsymbol{a}, x_d') - x))dx > 0.$$

Thus for any $(a_1, \dots, a_{d-1}) \in \mathbb{R}^{d-1}$, the $C^\infty$-function $\tau_t(a_1, \dots, a_{d-1}, x)$ is strictly increasing for with respect to $x$.

Next, take any compact subset $K \subset \mathbb{R}^d$. We show $\|\tau_t - \tau\|_{p,K} \to 0$ as $t \to 0$. We prove $\tau_t$ converges $\tau$ as $t \to 0$. Take $R > 0$ satisfying $K \subset B(R) := \{x \in \mathbb{R}^d : |x| \le R\}$. We assume $0 < t < 1$. Then we have $\phi_t * \tau = \phi_t * (\mathbf{1}_{B(R+1)}\tau)$. Since we have $\mathbf{1}_{B(R+1)}\tau \in L^p(\mathbb{R}^d)$, we obtain

$$\begin{aligned}\|\phi_t * \tau - \tau\|_{p,K} &= \|\phi_t * (\mathbf{1}_{B(R+1)}\tau) - \mathbf{1}_{B(R+1)}\tau\|_{p,K} \\ &\le \|\phi_t * (\mathbf{1}_{B(R+1)}\tau) - \mathbf{1}_{B(R+1)}\tau\|_{p,\mathbb{R}^d} \to 0 \quad (t \to 0).\end{aligned}$$

Here, we used a property of mollifier $\phi_t$ (see Theorem 8.14 in [79] for example).

Next, we consider the sup-approximation when $\tau$ is continuous. By direct computation, we have

$$\begin{aligned}\sup_{y \in K} |\tau_t(y) - \tau(y)| &\le \sup_{y \in K} \int_{\mathbb{R}^d} |\phi(x)| \cdot |\tau(y - tx) - \tau(y)|dx \\ &\le C \sup_{(x,y) \in \mathrm{supp}(\phi) \times K} |\tau(y - tx) - \tau(y)| \to 0 \quad (t \to 0).\end{aligned}$$

Here $C := \sup_{x \in \mathbb{R}^d} |\phi(x)|$. Thus in both cases above, By taking sufficiently small $t$, we obtain the desired $C^\infty$-function $\tilde{\tau} = \tau_t$. □

## D.3 Properties of Diffeomorphisms on $\mathbb{R}^d$: From $\mathrm{Diff}^2_{\mathrm{c}}$ to Nearly-Id

This section explains the reduction of the universality for $\mathrm{Diff}^2_{\mathrm{c}}$ to Nearly-Id elements. The reduction involves a structure theorem from the field of differential geometry. The results of this section are used as a building block for the proofs in Section D.2.2.

**Definition D.2** (Compactly supported diffeomorphism)**.** *The diffeomorphism $f$ on $\mathbb{R}^d$ is* compactly supported *if there exists a compact subset $K \subset \mathbb{R}^d$ such that for any $x \notin K$, $f(x) = x$. We denote by $\mathrm{Diff}^2_{\mathrm{c}}$ the space of compactly supported $C^2$-diffeomorpshisms.*

The set $\mathrm{Diff}^2_c$ constitutes a group whose group operation is the function composition. Moreover, $\mathrm{Diff}^2_c$ is a topological group with respect to the *Whitney topology* [95, Proposition 1.7.(9)]. Then there is a crucial structure theorem of $\mathrm{Diff}^2_c$ attributed to Herman, Thurston [270], Epstein [71], and Mather [179, 180]:

**Fact D.1.** *The group* $\mathrm{Diff}^2_c$ *is simple, i.e., any normal subgroup* $H \subset \mathrm{Diff}^2_c$ *is either* $\{\mathrm{Id}\}$ *or* $\mathrm{Diff}^2_c$.

The assertion is proven in Mather [180] for the connected component containing Id, instead of the entire set of compactly-supported $C^2$-diffeomorphisms when the domain space is a general manifold instead of $\mathbb{R}^d$. In the special case of $\mathbb{R}^d$, the connected component containing Id is shown to be $\mathrm{Diff}^2_c$ itself [95, Example 1.15], hence Fact D.1 follows. For details, see [95, Corollary 3.5 and Example 1.15].

As a side note, the assertion of Theorem D.1 is proved to hold generally for $C^r$-diffeomorphisms only except for $r = d + 1$ [95]. Nevertheless, this exception does not cause any problem in our proof, because we apply it with $r = 2$ and $d \geq 2$. The limitation only means that the structure of $C^2$-diffeomorphisms is better understood than that of $C^{d+1}$-diffeomorphisms. Also note that this exception does not affect the approximation capability for $C^{d+1}$-diffeomorphisms either as they are contained in $C^2$ where we perform our theoretical analyses. For the details of mathematical ingredients, see [13].

Here, we provide a precise definition of the *flow endpoints* introduced in Section 5.4.1.

**Definition D.3** (Flow endpoints)**.** *A* flow endpoint *is an element of* $\mathrm{Diff}^2_c$ *which can be represented as* $\phi(1)$*, where* $\phi : [0,1] \to \mathrm{Diff}^2_c$ *is a continuous map such that* $\phi(0) = \mathrm{Id}$ *and that* $\phi$ *is additive, namely,* $\phi(s) \circ \phi(t) = \phi(s+t)$ *for any* $s, t \in [0,1]$ *with* $s + t \in [0,1]$.

We use Fact D.1 to prove that a compactly supported diffeomorphism can be represented as a composition of flow endpoints in $\mathrm{Diff}^2_c$. The following lemma is a restatement of Lemma 5.2 in the main text.

**Lemma D.10.** *Let* $S \subset \mathrm{Diff}^2_c$ *be the set of all flow endpoints. Then,* $\mathrm{Diff}^2_c$ *coincides with the set of finite compositions of elements in* $S$ *defined by*

$$H := \{g_1 \circ \cdots \circ g_n : n \geq 1, g_1, \ldots, g_n \in S\}.$$

*Proof.* In view of Fact D.1, it is enough to show that $H$ forms a subgroup, that it is normal, and that it is non-trivial.

First, we prove the $H$ consists a subgroup of $\mathrm{Diff}^2_c$. By definition, for any $g, h \in H$, it is immediate to show that $g \circ h \in H$. We prove that $H$ is closed under inversion. For this, it suffices to show that $S$ is closed under inversion. Let $g = \phi(1) \in S$. Consider the map $\varphi : [0,1] \to \mathrm{Diff}^2_c$ defined by $\varphi(t) := (\phi(t))^{-1}$. Since $\mathrm{Diff}^2_c$ is a topological group [95, Proposition 1.7.(9)], $\varphi$ is continuous. Moreover, it is immediate to show that $\varphi$ is additive in the sense of Definition D.3, and that $\varphi(0) = \mathrm{Id}$. Thus, $g^{-1} = \varphi(1)$ is an element of $S$.

Next, we prove $H$ is normal. It suffice to show that $S$ is closed under conjugation since the conjugation $g \mapsto hgh^{-1}$ is a group homomorphism on $\mathrm{Diff}^2_c$. Let $g = \phi(1) \in S$, where $\phi : [0,1] \to \mathrm{Diff}^2_c$ is a continuous map associated to $g$. Then, we define a $\Phi : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ by $\Phi(x,t) = \phi(t)(x)$. We call $\Phi$ a flow associated with $g$. We take arbitrary $h \in \mathrm{Diff}^2_c$. Then, the function $\Phi' : \mathbb{R}^d \times [0,1]$ defined by $\Phi'(\cdot, s) := h^{-1} \circ \Phi(\cdot, s) \circ h$ is a flow associated with $h^{-1}gh$, which means $h^{-1}gh \in S$, i.e., $S$ is closed under conjugation.

Finally, we show $H$ is nontrivial. It suffices to show that $S$ includes a non-identity element. Let $\psi : \mathbb{R} \to \mathrm{O}(d)$ be a nontrivial homomorphism of Lie groups, where $\mathrm{O}(d)$ is a orthogonal group of degree $d$. Such $\psi$ exists, for example, let $\psi(t) := \exp(tA)$ for some nonzero skew-symmetric matrix $A$, namely, $A^\top = -A$. Let $u : [0, \infty) \to \mathbb{R}$ be a compactly supported $C^\infty$ function such that its

support does not include 0. Then, We define $\Phi : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ by $\Phi(x,t) := \psi(u(|x|)t)x$. Then, $\Phi$ is the flow associated with $\Phi(\cdot,1) \in S$, that is a non-identity element.

□

**Definition D.4** (Nearly-Id elements)**.** *Let $f \in \mathrm{Diff}^2_{\mathrm{c}}$. We say $f$ is* nearly-Id *if, for any $x \in \mathbb{R}^d$, the Jacobian $Df$ of $f$ at $x$ satisfies*

$$\|Df(x) - I\|_{\mathrm{op}} < 1,$$

*where $I$ is the unit matrix.*

**Corollary D.1.** *For any $f \in \mathrm{Diff}^2_{\mathrm{c}}$, there exist finite elements $g_1, \dots, g_r \in \mathrm{Diff}^2_{\mathrm{c}}$ such that $f = g_r \circ \cdots \circ g_1$ and $g_i$ is nearly-*Id *for any $i \in [r]$.*

*Proof.* Let $S$ be the subset of $\mathrm{Diff}^2_{\mathrm{c}}$ as defined above. Therefore, by Lemma D.10, there exist $h_1, \dots, h_m \in S$ such that $f = h_m \circ \cdots \circ h_1$. For $i \in [m]$, let $\phi_i$ be a flow associated with $h_i$. Since $[0,1] \ni t \mapsto \Phi_i(\cdot,t) \in \mathrm{Diff}^2_{\mathrm{c}}$ is continuous with respect to Whitney topology and $\Phi_i(\cdot,0)$ is the identity function, we can take a sufficiently large $n$ such that $\tilde{h}_i := \Phi_i(\cdot,1/n)$ is nearly-Id. By the additive property of $\Phi_i$, we have

$$f = h_m \circ \cdots \circ h_1 = \underbrace{\tilde{h}_m \circ \cdots \circ \tilde{h}_m}_{n \text{ times}} \circ \cdots \circ \underbrace{\tilde{h}_1 \circ \cdots \circ \tilde{h}_1}_{n \text{ times}},$$

which completes the proof of the corollary. □

## D.4 Proof of Theorem 5.2: $L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$

In this section, we provide the proof details of Theorem 5.2 in the main text. The correspondence between this section and Section 5.4.2 in the main text is as follows: Steps 1, 2, 3 correspond to Section D.4.1, Step 4 corresponds to Section D.4.2, and Step 5 is justified by Proposition D.3 in Section D.6.

### D.4.1 Approximation of General Elements of $\mathcal{S}^0_{\mathrm{c}}$

In this section, we prove the following lemma to construct an approximator for an arbitrary element of $\mathcal{S}^0_{\mathrm{c}}$ (hence for $\mathcal{S}^\infty_{\mathrm{c}}$) within $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. It is based on Lemma D.12 proved in Section D.4.2, which corresponds to a special case.

Here, we rephrase Theorem 5.2 as in the following:

**Lemma D.11** ($L^p$-universality of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ for compactly supported $\mathcal{S}^\infty_{\mathrm{c}}$)**.** *Let $p \in [1,\infty)$. Assume $\mathcal{H}$ is a* sup*-universal approximator for $C^\infty_c(\mathbb{R}^{d-1})$ and that it consists of piecewise $C^1$-functions. Let $f \in \mathcal{S}^0_{\mathrm{c}}$, $\varepsilon > 0$, and $K \subset \mathbb{R}^d$ be a compact subset. Then, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|f - g\|_{p,K} < \varepsilon$.*

*Proof.* Since we can take $a > 0$, $b \in \mathbb{R}$ satisfying $aK + b \subset [0,1]^d$, it is enough to prove the assertion for the case $K = [0,1]^d$.

Next, we show that we can assume that for any $(\boldsymbol{x}, y) \in \mathbb{R}^d$, $u(\boldsymbol{x},0) = 0$ and $u(\boldsymbol{x},1) = 1$ for any $\boldsymbol{x} \in \mathbb{R}^{d-1}$. Since $u(\boldsymbol{x},\cdot)$ is a diffeomorphism, we have $u(\boldsymbol{x},0) \neq u(\boldsymbol{x},1)$ for any $x \in \mathbb{R}$. By the continuity of $f$, either of $u(\boldsymbol{x},0) > u(\boldsymbol{x},1)$ for all $\boldsymbol{x} \in [0,1]^{d-1}$ or $u(\boldsymbol{x},0) < u(\boldsymbol{x},1)$ for all $x \in [0,1]^{d-1}$ holds. Without loss of generality, we assume the latter case holds (if the former

one holds, we just switch $u(\boldsymbol{x},0)$ and $u(\boldsymbol{x},1)$). We define $s(\boldsymbol{x}) = -\log(u(\boldsymbol{x},1) - u(\boldsymbol{x},0))$ and $t(\boldsymbol{x}) = -u(\boldsymbol{x},0)(u(\boldsymbol{x},1) - u(\boldsymbol{x},0))^{-1}$. By a direct computation, we have

$$\Psi_{d-1,s,t} \circ f(\boldsymbol{x},y) = \left(\boldsymbol{x}, \frac{u(\boldsymbol{x},y) - u(\boldsymbol{x},0)}{u(\boldsymbol{x},1) - u(\boldsymbol{x},0)}\right) =: (\boldsymbol{x}, u_0(\boldsymbol{x},y)).$$

In particular, $\Psi_{s,t} \circ f(\boldsymbol{x},0) = (\boldsymbol{x},0)$ and $\Psi_{s,t} \circ s(\boldsymbol{x},1) = (\boldsymbol{x},1)$ hold. , and the map $y \mapsto u_0(\boldsymbol{x},y)$ is a diffeomorphism for each $\boldsymbol{x}$. Thus if we prove the existence of an approximator for $\Psi_{s,t} \circ f$, by Proposition D.3, we can arbitrarily approximate $f$ itself.

For $\underline{k} := (k_1, \ldots, k_{d-1}) \in \mathbb{Z}^{d-1}$ and $n \in \mathbb{N}$, we define $(\underline{k})_n := \sum_{i=1}^{d} k_i n^{i-1} \in \{0, \ldots, n^d - 1\}$, that is, $\underline{k}$ is the $n$-adic expansion of $(\underline{k})_n$. For any $n \in \mathbb{N}$, define the following discontinuous ACF: $\psi_n \colon [0,1]^d \to [0,1]^{d-1} \times [0, n^d]$ by

$$\psi_n(\boldsymbol{x},y) := \left(\boldsymbol{x}, y + \sum_{k_1,\cdots,k_{d-1}=0}^{n-1} (\underline{k})_n 1_{\Delta^n_{\underline{k}+1}}(\boldsymbol{x})\right),$$

where $\underline{k} := (k_1, \ldots, k_d)$ and $\underline{k}+1 := (k_1+1, \ldots, k_d+1)$. We take an increasing function $v_n \colon \mathbb{R} \to \mathbb{R}$ that is smooth outside finite points such that

$$v_n(z) := \begin{cases} u\left(\frac{k_1}{n}, \cdots, \frac{k_{d-1}}{n}, z - (\underline{k})_n\right) + (\underline{k})_n & \text{if } z \in [(\underline{k})_n, (\underline{k})_n + 1) \\ z & \text{if } z \notin [0, n^d). \end{cases}$$

We consider maps $h_n$ on $[0,1]^{d-1} \times [0, n^d]$ and $f_n : [0,1]^d \to [0,1]^d$ defined by

$$\begin{aligned} h_n(\boldsymbol{x},z) &:= (\boldsymbol{x}, v_n(z)), \\ f_n &:= \psi_n^{-1} \circ h_n \circ \psi_n. \end{aligned}$$

Then we have the following claim.

**Claim.** For all $k_1, \cdots, k_{d-1} = 0, \cdots, n-1$, we have

$$f_n(\boldsymbol{x},y) = \left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right)\right)$$

on $\prod_{i=1}^{d-1} [\frac{k_i}{n}, \frac{k_i+1}{n}) \times [0,1)$.

In fact, we have

$$\begin{aligned} f_n(\boldsymbol{x},y) &= \psi_n^{-1} \circ h_n \circ \psi_n(\boldsymbol{x},y) \\ &= \psi_n^{-1} \circ h_n(\boldsymbol{x}, y + (\underline{k})_n) \\ &= \psi_n^{-1}(\boldsymbol{x}, v_n(y + (\underline{k})_n)) \\ &= \psi_n^{-1}\left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right) + (\underline{k})_n\right) \\ &= \left(\boldsymbol{x}, u\left(\frac{k_1}{n}, \ldots, \frac{k_{d-1}}{n}, y\right)\right). \end{aligned}$$

Therefore, the claim above has been proved. Hence we see that $\|f - f_n\|_{\sup,K} \to 0$ as $n \to \infty$. By Lemma D.12 below and the universal approximation property of $\mathcal{H}$, for any compact subset $K$ and $\varepsilon > 0$, there exist $g_1, g_2, g_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\left\|g_1 - \psi_n^{-1}\right\|_{p,K} < \varepsilon$, $\|g_2 - h_n\|_{p,K} < \varepsilon$, and $\|g_3 - \psi_n\|_{p,K} < \varepsilon$. Thus by Proposition D.3, for any compact $K$ and $\varepsilon > 0$, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|g - f\|_{p,K} < \varepsilon$. □

### D.4.2 Special Case: Approximation of Coordinate-wise Independent Transformation

In this section, we show the lemma claiming that special cases of single-coordinate transformations, namely coordinate-wise independent transformations, can be approximated by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$ given sufficient representational power of $\mathcal{H}$.

**Lemma D.12.** *Let $p \in [1, \infty)$. Assume $\mathcal{H}$ is a* sup*-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$ and that it consists of piecewise $C^1$-functions. Let $u : \mathbb{R} \to \mathbb{R}$ be a continuous increasing function. Let $f : \mathbb{R}^d \to \mathbb{R}^d; (\boldsymbol{x}, y) \mapsto (\boldsymbol{x}, u(y))$ where $\boldsymbol{x} \in \mathbb{R}^{d-1}$ and $y \in \mathbb{R}$. For any compact subset $K \subset \mathbb{R}^d$ and $\varepsilon > 0$, there exists $g \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|f - g\|_{p,K} < \varepsilon$.*

*Proof.* We may assume without loss of generality, in light of Lemma D.9, that $u$ is a $C^\infty$-diffeomorphism on $\mathbb{R}$ and that the inequality $u'(y) > 0$ holds for any $y \in \mathbb{R}$. Furthermore, we may assume that $u$ is compactly supported (i.e., $u(y) = y$ outside a compact subset of $\mathbb{R}$) without loss of generality because we can take a compactly supported diffeomorphism $\tilde{u}$ and $a, b \in \mathbb{R}$ $(a \neq 0)$ such that $a\tilde{u} + b = u$ on any compact set containing $K$ by Lemma D.4, and the scaling $a$ and the offset $b$ can be realized by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$.

Fix $\delta \in (0, 1)$. We define the following functions:

$$\begin{aligned}
\psi_0(\boldsymbol{x}, y) &:= (\boldsymbol{x}_{\leq d-2}, u'(y)x_{d-1}, y) \\
&= (\boldsymbol{x}_{\leq d-2}, \exp(\log u'(y))x_{d-1}, y), \\
\psi_1(\boldsymbol{x}, y) &:= \left(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y\right), \\
\psi_2(\boldsymbol{x}, y) &:= (\boldsymbol{x}_{\leq d-2}, x_{d-1}, y + \delta x_{d-1}), \\
\psi_3(\boldsymbol{x}, y) &:= \left(\boldsymbol{x}_{\leq d-2}, x_{d-1} - \delta^{-1}(y - u^{-1}(y)), y\right),
\end{aligned}$$

where we denote $\boldsymbol{x} = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. First, we show

$$\|f - \psi_3 \circ \psi_2 \circ \psi_1 \circ \psi_0\|_{\sup,K} \to 0 \quad (\delta \to 0).$$

By a direct computation, we have

$$\begin{aligned}
&\psi_3 \circ \psi_2 \circ \psi_1(\boldsymbol{x}, y) \\
&= \psi_3 \circ \psi_2(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y) \\
&= \psi_3(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), y + \delta(x_{d-1} + \delta^{-1}(u(y) - y))) \\
&= \psi_3(\boldsymbol{x}_{\leq d-2}, x_{d-1} + \delta^{-1}(u(y) - y), \delta x_{d-1} + u(y)) \\
&= (\boldsymbol{x}_{\leq d-2}, x_{d-1} - \delta^{-1}(\delta x_{d-1} + u(y) - u^{-1}(\delta x_{d-1} + u(y))), \delta x_{d-1} + u(y)) \\
&= (\boldsymbol{x}_{\leq d-2}, \delta^{-1}u^{-1}(\delta x_{d-2} + u(y)) - \delta^{-1}y, u(y) + \delta x_{d-1}),
\end{aligned}$$

where $\boldsymbol{x} = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. Since $u \in C^\infty([-r, r])$ where $r = \max_{(\boldsymbol{x},y)\in K} |y|$, by applying Taylor's theorem, there exists a function $R(\boldsymbol{x}, y; \delta)$ and a constant $C = C([-r, r], u) > 0$ such that

$$u^{-1}(u(y) + \delta x) = y + u'(y)^{-1}\delta x + R(\boldsymbol{x}, y; \delta)(\delta x)^2 \quad \text{and} \quad \sup_{\delta\in(0,1)} |R(\boldsymbol{x}, y; \delta)| \leq C$$

for all $(\boldsymbol{x}, y) \in K$. Therefore, we have

$$\psi_3 \circ \psi_2 \circ \psi_1 \circ \psi_0(\boldsymbol{x}, y) = (\boldsymbol{x}, u(y)) + \delta(R(\boldsymbol{x}, u'(y)x_{d-1}; \delta)\boldsymbol{x}_{\leq d-1}, u'(y)x_{d-1}).$$

For any compact subset $K$, the last term uniformly converges to 0 as $\delta \to 0$ on $K$.

Assume $\delta$ is taken to be small enough. Now, we approximate $\psi_3 \circ \cdots \circ \psi_0$ by the elements of $\mathrm{INN}_{\mathcal{H}\text{-ACF}}$. Since $u$ is a compactly-supported $C^\infty$-diffeomorphism on $\mathbb{R}$, the functions $(\boldsymbol{x}_{\leq d-2}, y) \mapsto \log u'(y)$, $(\boldsymbol{x}_{\leq d-2}, y) \mapsto u(y) - y$, and $(\boldsymbol{x}_{\leq d-2}, y) \mapsto y - u^{-1}(y)$, each appearing in $\psi_0$, $\psi_1$, $\psi_3$, respectively, belong to $C_c^\infty(\mathbb{R}^{d-1})$. On the other hand, $\psi_2$ can be realized by $\mathrm{GL} \subset \mathrm{Aff}$. Therefore, combining the above with the fact that $\mathcal{H}$ is a sup-universal approximator for $C_c^\infty(\mathbb{R}^{d-1})$, we have that for any compact subset $K' \subset \mathbb{R}^d$ and any $\varepsilon > 0$, there exist $\phi_0, \ldots, \phi_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|\psi_i - \phi_i\|_{\sup,K'} < \varepsilon$. In particular, we can find $\phi_0, \ldots, \phi_3 \in \mathrm{INN}_{\mathcal{H}\text{-ACF}}$ such that $\|\psi_i - \phi_i\|_{p,K'} < \varepsilon$.

Now, recall that $\mathcal{H}$ consists of piecewise $C^1$-functions as well as $\psi_i$ $(i = 0, \ldots, 3)$. Moreover, $\psi_0, \psi_1, \psi_3$ are compactly supported while $\psi_2 \in \mathrm{GL}$, hence they are Lipschitz continuous outside a bounded open subset. Therefore, by Proposition D.3, we have the assertion of the lemma.

□

## D.5 Locally Bounded Maps and Piecewise Diffeomorphisms

In this section, we provide the notions of locally bounded maps and piecewise $C^1$-maps. These notions are used to state the regularity conditions on the CF layers in Theorem 5.1 and to prove the results in Section D.6.

### D.5.1 Definition of Locally Bounded Maps

Here, we provide the definition of locally bounded maps. It is a very mild condition that is satisfied in most cases of practical interest, e.g., by continuous maps.

**Definition D.5** (Locally bounded maps)**.** *Let $f$ be a map from $\mathbb{R}^m$ to $\mathbb{R}^n$. We say $f$ is* locally bounded *if for each point $\boldsymbol{x} \in \mathbb{R}^m$, there exists a neighborhood $U$ of $\boldsymbol{x}$ such that $f$ is bounded on $U$.*

As a special case, continuous maps are locally bounded; take an open ball $U$ centered at $\boldsymbol{x}$ and take a compact set containing $U$ to see that $f$ is bounded on $U$.

### D.5.2 Definition and Properties of Piecewise $C^1$-maps

In this section, we give the definition of piecewise $C^1$-maps and their properties. Examples of piecewise $C^1$-diffeomorphisms appearing in the chapter include $\mathcal{H}$-ACF with $\mathcal{H}$ being MLPs with ReLU activation.

**Definition D.6** (piecewise $C^1$-maps)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a measurable map. We say $f$ is a* piecewise $C^1$-map *if there exists a mutually disjoint family of (at most countable) open subsets $\{V_i\}_{i \in I}$ such that*

- $\mathrm{vol}(\mathbb{R}^d \setminus U_f) = 0$,
- *for any $i \in I$, there exists an open subset $W_i$ containing the closure $\overline{V_i}$ of $V_i$, and $C^1$-map $\tilde{f}_i : W_i \to \mathbb{R}^d$ such that $\tilde{f}_i|_{V_i} = f|_{V_i}$, and*
- *for any compact subset $K$, $\#\{i \in I : V_i \cap K \neq \emptyset\} < \infty$.*

*where we denote $U_f := \bigsqcup_{i \in I} V_i$, and $\#(\cdot)$ denotes the cardinality of a set.*

We remark that piecewise $C^1$-maps are essentially locally bounded in the sense that for any compact set $K \subset \mathbb{R}^d$, $\mathrm{ess.sup}_K \|f\| = \|f\|_{\sup, K \cap U_f} < \infty$. Then we define a piecewise $C^1$-diffeomorphisms:

**Definition D.7** (piecewise $C^1$-diffeomorphisms)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a piecewise $C^1$-map. We say $f$ is a* piecewise $C^1$-diffeomorphism *if*

1. *the image of nullset via $f$ is also a nullset,*
2. *$f|_{U_f}$ is injective, and for $i \in I$, $\tilde{f}_i$ is a $C^1$-diffeomorphism from $W_i$ onto $\tilde{f}_i(W_i)$,*
3. $\mathrm{vol}\left(\mathbb{R}^d \setminus f(U_f)\right) = 0$, *and*
4. *for any compact subset $K$, $\#\{i \in I : f(V_i) \cap K \neq \emptyset\} < \infty$.*

We summarize the basic properties of piecewise $C^1$-diffeomorphisms in the proposition below:

**Proposition D.1.** *Let $f$ and $g$ be piecewise $C^1$-diffeomorphisms. Then, we have the following.*

1. *There exists a piecewise $C^1$-diffeomorphism $f^\dagger$ such that $f(f^\dagger(x)) = x$ for $x \in U_{f^\dagger}$ and $f^\dagger(f(y)) = y$ for $y \in U_f$.*
2. *For any $h \in L^1$, we have $\int h(x)dx = \int h(f(x))|Df(x)|dx$, where $|Df(x)|$ is the absolute value of the determinat of the Jacobian matrix of $f$ at $x$.*
3. *For any compact subset $K$, $f^{-1}(K) \cap U_f$ is a bounded subset.*
4. *For any nullset $F$, then $f^{-1}(F)$ is also a nullset.*
5. *For any measurable set $E$ and any compact set $K$, $f^{-1}(E \cap K)$ has a finite volume.*
6. *The composition $f \circ g$ is also a piecewise $C^1$-diffeomorphism.*

*Proof. Proof of 1* : Fix $a \in \mathbb{R}^d$. For $x \in \mathbb{R}^d \setminus f(U_f)$, define $f^\dagger(x) = a$, and for $x \in f(V_i)$, define $f^\dagger(x) := f|_{V_i}^{-1}(x)$. Then, $f^\dagger$ is a piecewise $C^1$-map with respect to the family of pairwise disjoint open subsets $\{f(V_i)\}_{i\in I}$, and satisfies the conditions for piecewise $C^1$-diffeomphism.

*Proof of 2* : It follows by the following computation:

$$\begin{aligned}\int h(x)dx &= \int_{f(U_f)} h(x)dx \\ &= \sum_{i\in I} \int_{f(V_i)} h(x)dx \\ &= \sum_{i\in I} \int_{V_i} h(f(x))|Df(x)|dx = \int h(f(x))|Df(x)|dx.\end{aligned}$$

*Proof of 3* It suffices to show that $f^{-1}(K) \cap U_f$ is covered by finitely many compact subsets. In fact, we remark that only finitely many $V_i$'s intersect with $f^{-1}(K)$. If not, infinitely many $f(V_i)$ intersects $f(f^{-1}(K)) \subset K$, which contradicts the definition of piecewise $C^1$-diffeomorphisms. Let $I_0 \subset I$ be a finite subset composed of $i \in I$ such that $V_i$ intersecting with $f^{-1}(K)$. For $i \in I_0$, we define a compact subset $F_i := \tilde{f}_i^{-1}(\tilde{f}_i(\overline{V_i}) \cap K)$. Then we see that $f^{-1}(K) \cap U_f$ is contained in $\cup_{i\in I_0} F_i$.

*Proof of 4* : It suffices to show that for any compact subset $K$, the volume of $f^{-1}(F) \cap K$ is zero. By applying 2 to the case $h = \mathbf{1}_F$, we see that

$$\int_{f^{-1}(F)} |Df(x)|dx = 0.$$

For $n > 0$, let $E_n := f^{-1}(F) \cap K \cap \{x \in \mathbb{R}^d : |Df(x)| \geq 1/n\}$. Then we have

$$\frac{\mathrm{vol}(E_n)}{n} \leq \int_{E_n} |Df(x)|dx \leq \int_{f^{-1}(F)} |Df(x)|dx = 0,$$

thus $\mathrm{vol}(K \cap f^{-1}(F)) = \lim_{n\to\infty} \mathrm{vol}(E_n) = 0$

*Proof of 5* : By applying 2 to the case $h = \mathbf{1}_{E\cap K}$, we see that

$$\int_{f^{-1}(E\cap K)} |Df(x)|dx = \mathrm{vol}(E \cap K).$$

Let $F$ be a closure of $f^{-1}(K) \cap U_f$. By 3, $F$ is a compact subset. Let $I_0 := \{i \in I : F \cap V_i \neq \emptyset\}$ be a finite subset. Then we have

$$\begin{aligned} C &:= \inf_{f^{-1}(K)\cap U_f} |Df| \\ &\geq \inf_{i\in I_0} \inf_{F\cap \overline{V_i}} |D\tilde{f}_i| > 0. \end{aligned}$$

Thus,

$$\int_{f^{-1}(E\cap K)\cap U_f} |Df(x)|dx \geq C\mathrm{vol}(f^{-1}(E \cap K)),$$

where the last equality follows from $\mathrm{vol}(f^{-1}(E\cap K) \setminus U_f) = 0$. Thus we have $\mathrm{vol}(f^{-1}(E\cap K)) < \infty$

*Proof of 6* : We denote by $\{V_i\}_{i\in I}$, $\{V'_j\}_{j\in J}$ the disjoint open families associated with $f$ and $g$, respectively. At first, we prove $f \circ g$ is a piecewise $C^1$-map. Let $V_{ij} := g^{-1}(V_i \cap g(V'_j)) \cap U_g$ and define $\mathcal{U}_{f\circ g} := \{V_{ij}\}_{(i,j)\in I\times J}$. Let $U_{f\circ g} := \cup_{i,j} V_{ij} = g^{-1}(U_f \cap g(U_g)) \cap U_g$. By 4, the volume of $\mathbb{R}^d \setminus U_{f\circ g}$ is zero. On each $V_{ij}$, $\tilde{f}_i \circ \tilde{g}_j$ is an extension of $f \circ g|_{V_{ij}}$. For any compact subset $K$, $\#\{(i,j) \in I \times J : K \cap V_{ij} \neq \emptyset\} < \infty$. In fact, suppose the number is infinite. Then $g(U_f \cap K)$ intersects with an infinite number of open subsets in the form of $g(U_f \cap K) \cap V_i \cap g(V'_j)$. On the other hand $g(U_f \cap K)$ is a bounded subset, thus by definition, the number of $(i,j) \in I \times J$ satisfying $\overline{g(U_f \cap K)} \cap V_i \cap g(V'_j) \neq \emptyset$ is finite. It is a contradiction. Therefore, $g \circ f$ is a piecewise $C^1$-map.

Next, we prove $f \circ g$ is a piecewise $C^1$-diffeomorphism. The first and second condition follows by definition. For the third condition, since $\mathbb{R}^d \setminus f \circ g(U_{f\circ g}) = \big(\mathbb{R}^d \setminus f(U_f)\big) \cup \big(\mathbb{R}^d \setminus f\big(g(U_g)\big) \subset \mathbb{R}^d \setminus f(g(U_g) \cap U_f)$, it suffices to show that the volue of $\mathbb{R}^d \setminus f(g(U_g) \cap U_f)$ is zero. In fact, by the injectivity of $f$ on $U_f$, we have $f(g(U_g) \cap U_f) = f(U_f) \setminus f(U_f \setminus g(U_g))$. Thus $\mathbb{R}^d \setminus f(g(U_g) \cap U_f) = (\mathbb{R}^d \setminus f(U_f)) \cup f(U_f \setminus g(U_g))$. By definition of $C^1$-diffeomorphism, we conclude $\mathbb{R}^d \setminus f(g(U_g) \cap U_f)$ is a nullset. For the fourth condition, let $K$ be a compact subset. Assume the $\{(i,j) \in I \times J : f \circ g(V_{ij}) \cap K \neq \emptyset\} = \infty$. Since $f$ is a piecewise $C^1$-diffeomorphism, there exist infinitely many elements in $j \in J$ such that $f \circ g(V'_j) \cap f(U_f) \cap K \neq \emptyset$. On the other hand, $f^{-1}(K \cap f(U_f)) \cap U_f$ is bounded, and its closure intersects with only finitely many $g(V'_j)$'s, thus $K \cap f(U_f)$ intersects with only finitely many $f \circ g(V'_j)$, which is a contradiction.

□

For a measurable map $f : \mathbb{R}^m \to \mathbb{R}^n$ and any $R > 0$, we define a measurable set

$$\mathcal{L}(R; f) := \{x \in \mathbb{R}^m : \|f(x) - f(y)\| > R\|x - y\| \text{ for some } y \in U_f\}.$$

Then we have the following proposition:

**Proposition D.2.** *Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a piecewise $C^1$-map. Assume $f$ is linearly increasing, namely, there exists $a, b > 0$ such that $\|f(x)\| < a\|x\| + b$ for any $x \in \mathbb{R}^m$. Then for any compact subset $K'$, $\mathrm{vol}(\mathcal{L}(R; f) \cap K') \to 0$ as $R \to \infty$.*

*Proof.* Let $B$ be an open ball containing $K'$ of radius $r$. Fix an arbitrary $\varepsilon > 0$. We note that the linearly increasing condition implies the locally boundedness of $f$. Let $C := \sup_{\overline{B}} \|f\|$. For $\delta > 0$, we define

$$V_\delta := \{x \in \overline{B} : \mathrm{dist}\,(x, \partial U_f \cup \partial B)) < \delta\},$$

where $\mathrm{dist}(x, S) := \inf_{y\in S}\{\|x-y\|\}$. Set $\delta$ to be $\mathrm{vol}(V_\delta) < \varepsilon$. We claim that

$$L := \sup_{(x,y)\in K'\times\mathbb{R}^m\setminus B} \frac{\|f(x)-f(y)\|}{\|x-y\|}$$

is finite. In fact, let $r' := \inf_{x\in K', y\notin B}\|x-y\|$. Then for $x \in K'$ and $y \notin B$, we have

$$\begin{aligned}
\frac{\|f(x)-f(y)\|}{\|x-y\|} &\le \frac{\|f(x)\|+\|f(y)\|}{\|x-y\|} \\
&\le \frac{a\|x\|+a\|y\|+2b}{\|x-y\|} \\
&\le \frac{a\|x\|+a(\|x-y\|+\|x\|)+2b}{\|x-y\|} \\
&\le a + \frac{2a\|x\|+2b}{\|x-y\|} \\
&< a + \frac{2ar+2b}{r'}.
\end{aligned}$$

Thus, $L$ is finite. Since $\overline{B}$ intersects with finitely many $V_i$'s, $f|_{B\setminus V_{\delta/2}}$ is a Lipschitz function. Put $L_\delta > 0$ as the Lipschitz constant of $f|_{B\setminus V_{\delta/2}}$. Then for any $R > \max(L, L_\delta, 4C/\delta)$, we see that $\mathcal{L}(R; f) \cap K'$ is contained in $V_\delta$. Actually, we should prove that $x \notin \mathcal{L}(R; f)$ when $x \in K' \setminus V_\delta$. Take arbitrary $y \in \mathbb{R}^m$. When $y \notin B$, since $x \in K'$, we have $\frac{\|f(x)-f(y)\|}{\|x-y\|} \le L$ by the definition of $L$. When $y \in B \setminus V_{\delta/2}$, since $x \in K' \setminus V_\delta \subset B \setminus V_{\delta/2}$, we have $\frac{\|f(x)-f(y)\|}{\|x-y\|} \le L_\delta$ by the definition of $L_\delta$. When $y \in V_{\delta/2}$, we have $\|x-y\| \ge \frac{\delta}{2}$ because $x \notin V_\delta$. Thus,

$$\frac{\|f(x)-f(y)\|}{\|x-y\|} \le \frac{\|f(x)\|+\|f(y)\|}{\delta/2} \le \frac{C+C}{\delta/2} \le \frac{4C}{\delta}.$$

Combining these three cases, we conclude that $x \notin \mathcal{L}(R; f)$. Thus we have $\mathrm{vol}(\mathcal{L}(R; f) \cap K') < \varepsilon$, namely, we conclude $\mathrm{vol}(\mathcal{L}(R; f) \cap K') \to 0$ as $R \to \infty$. □

*Remark* D.1. The linearly increasing condition is important to prove our main theorem. Our approximation targets are compactly supported diffeomorphisms, affine transformations, and the discontinuous ACFs appeared in Section 5.4.2 or Section D.4.1, all of which satisfy the linearly increasing condition.

## D.6 Compatibility of Approximation and Composition

In this section, we prove the following proposition. It enables the component-wise approximation, i.e., given a transformation that is represented by a composition of some transformations, we can approximate it by approximating each constituent and composing them. The justification of this procedure is not trivial and requires a fine mathematical argument. The results here build on the terminologies and the propositions for piecewise $C^1$-diffeomorphisms presented in Section D.5.

**Proposition D.3.** *Let $\mathcal{M}$ be a set of piecewise $C^1$-diffeomorphisms (resp. locally bounded maps) from $\mathbb{R}^d$ to $\mathbb{R}^d$, and $F_1, \dots, F_r$ be linearly increasing piecewise $C^1$-diffeomorphisms (resp. continuous maps) from $\mathbb{R}^d$ to $\mathbb{R}^d$ ($r \ge 2$). Assume for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, there exists $\widetilde{G}_1, \dots, \widetilde{G}_r \in \mathcal{M}$ such that for $i \in [r]$, $\left\|F_i - \widetilde{G}_i\right\|_{p,K} < \varepsilon$ (resp. $\left\|F_i - \widetilde{G}_i\right\|_{\sup,K} < \varepsilon$). Then for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, there exists $G_1, \dots, G_r \in \mathcal{M}$, such that*

$$\|F_r \circ \cdots \circ F_1 - G_r \circ \cdots \circ G_1\|_{p,K} < \varepsilon$$

$$\left(\textit{resp.}\ \ \|F_r \circ \cdots \circ F_1 - G_r \circ \cdots \circ G_1\|_{\sup,K} < \varepsilon\right)$$

*Proof.* We prove by induction. In the case of $r = 2$, it follows by Lemma D.13 (for $L^p$-norm) or Lemma D.14 (for sup-norm) below in the case of $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}$. In the general case, let $\widetilde{F}_2 := F_r \circ \cdots F_2$. Then by the induction hypothesis, for any compact set $K$ and $\varepsilon > 0$, there exists $\widetilde{G}_2 = G_r \circ \cdots \circ G_2$ for some $G_i \in \mathcal{M}$ such that $\left\|\widetilde{F}_2 - \widetilde{G}_2\right\|_{?,K} < \varepsilon$, where $? = p$ or sup. By applying Lemma D.13 or Lemma D.14 with $\mathcal{M}_1 = \mathcal{M}$ and $\mathcal{M}_2 = \mathcal{M} \circ \cdots \circ \mathcal{M}$ (the set of compositions of $r-1$ elements of $\mathcal{M}$) below, we conclude the proof. □

**Lemma D.13.** *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be sets of piecewise $C^1$-diffeomorphisms from $\mathbb{R}^d$ to $\mathbb{R}^d$. Let $F_1, F_2 : \mathbb{R}^d \to \mathbb{R}^d$ be linearly increasing piecewise $C^1$-diffeomorphisms. Assume for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, for $i = 1, 2$, there exists $\widetilde{G}_i \in \mathcal{M}_i$ such that $\left\|F_i - \widetilde{G}_i\right\|_{p,K} < \varepsilon$. Then for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, for $i = 1, 2$, there exists $G_i \in \mathcal{M}_i$, such that*

$$\|F_2 \circ F_1 - G_2 \circ G_1\|_{p,K} < \varepsilon.$$

*Proof.* Fix arbitrary $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$. Put $K' := \overline{F_1(K \cap U_{F_1})}$. Then, since $F_1(K \cap U_{F_1})$ is bounded (see the remark under Definition D.6), $K'$ is compact. We claim that there exists $R > 0$ such that

$$\mathrm{vol}(F_1^{-1}\left(\mathcal{L}(R; F_2) \cap K'\right))^{1/p} < \frac{\varepsilon}{3\underset{K'}{\mathrm{ess.sup}}\|F_2\|},$$

which can be confirmed as follows. Take an increasing sequence $R_n > 0$ $(n \geq 1)$ satisfying $\lim_{n\to\infty} R_n = \infty$. Let $B_n := \mathcal{L}(R_n; f) \cap K'$ and $A_n := F_1^{-1}(B_n)$. Then, from Proposition D.2, we have $\mathrm{vol}(B_n) \to 0$, which implies $\mathrm{vol}(\bigcap_{n=1}^{\infty} B_n) = 0$. By Proposition D.1 (4), we have $\mathrm{vol}(\bigcap_{n=1}^{\infty} A_n) = \mathrm{vol}(F_1^{-1}(\bigcap_{n=1}^{\infty} B_n)) = 0$. By Proposition D.1 (5), we have $\mathrm{vol}(A_1) = \mathrm{vol}(F_1^{-1}(B_1)) < \infty$. Recall that if a decreasing sequence $\{S_n\}_{n=1}^{\infty}$ of measurable sets satisfies $\mathrm{vol}(S_1) < \infty$ and $\mathrm{vol}(\bigcap_{n=1}^{\infty} S_n) = 0$, then $\lim_{n\to\infty} \mathrm{vol}(S_n) = 0$. Therefore, we obtain $\lim_{n\to\infty} \mathrm{vol}(A_n) = 0$ and we have the assertion of the claim.

Take $G_1 \in \mathcal{M}_1$ such that

$$\|F_1 - G_1\|_{p,K} < \frac{\varepsilon}{3R}.$$

Let $S := F_1^{-1}\left(\mathcal{L}(R; F_2) \cap K'\right)$, and define $K'' := \overline{(G_1^{\dagger})^{-1}(K) \cap U_{G_1^{\dagger}}}$. Then, the compactness of $K''$ follows from Proposition D.1 (3). Next, we take $G_2 \in \mathcal{M}_2$ such that

$$\|F_2 - G_2\|_{p,K''} < \frac{\varepsilon}{3 \underset{(G_1^{\dagger})^{-1}(K)}{\mathrm{ess.sup}} |\det(DG_1^{\dagger})|}$$

where $G_1^{\dagger}$ is a piecewise $C^1$-diffeomorphism defined by Proposition D.1 (1). Then we have

$$\begin{aligned}
&\|F_2 \circ F_1 - G_2 \circ G_1\|_{p,K} \\
&\leq \|F_2 \circ F_1 - F_2 \circ G_1\|_{p,K} + \|F_2 \circ G_1 - G_2 \circ G_1\|_{p,K} \\
&\leq \|(F_2 \circ F_1 - F_2 \circ G_1)\mathbf{1}_S\|_{p,K} + \left\|(F_2 \circ F_1 - F_2 \circ G_1)\mathbf{1}_{K\setminus S}\right\|_{p,K} \\
&\quad + \underset{(G_1^{\dagger})^{-1}(K)}{\mathrm{ess.sup}} |\det(DG_1^{\dagger})| \|F_2 - G_2\|_{p,K''} \\
&< \varepsilon.
\end{aligned}$$

□

**Lemma D.14** (compatibility of composition)**.** *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be sets of locally bounded maps from $\mathbb{R}^d$ to $\mathbb{R}^d$. Let $F_1, F_2 : \mathbb{R}^d \to \mathbb{R}^d$ be continuous maps. Assume for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, for $i = 1, 2$, there exists $\widetilde{G}_i \in \mathcal{M}_i$ such that $\left\|F_i - \widetilde{G}_i\right\|_{\sup,K} < \varepsilon$. Then for any $\varepsilon > 0$ and compact set $K \subset \mathbb{R}^d$, for $i = 1, 2$, there exists $G_i \in \mathcal{M}_i$, such that*

$$\|F_2 \circ F_1 - G_2 \circ G_1\|_{\sup,K} < \varepsilon.$$

*Proof.* Take any positive number $\epsilon > 0$ and compact set $K \subset \mathbb{R}^d$. Put $r := \max_{k\in K} |F_1(k)|$ and $K' := \{x \in \mathbb{R}^d : |x| \leq r + 1\}$. Let $G_2 \in \mathcal{M}_2$ satisfying

$$\sup_{x\in K'} |F_2(x) - G_2(x)| \leq \frac{\epsilon}{2}.$$

Since any continuous map is uniformly continuous on a compact set, we can take a positive number $\delta > 0$ such that for any $x, y \in K'$ with $|x - y| < \delta$,

$$|F_2(x) - F_2(y)| < \frac{\varepsilon}{2}.$$

From the assumption, we can take $G_1 \in \mathcal{M}_1$ satisfying

$$\sup_{x\in K} |F_1(x) - G_1(x)| \leq \min\{1, \delta\}.$$

Then, it is clear that $F_1(K) \subset K'$ by the definition of $K'$. Moreover, we have $G_1(K) \subset K'$. In fact, we have

$$|G_1(k)| \leq \sup_{x\in K} |F_1(x) - G_1(x)| + |F_2(k)| \leq 1 + r \quad (k \in K).$$

Then for any $x \in K$, we have

$$\begin{aligned} |F_2 \circ F_1(x) - G_2 \circ G_1(x)| &\leq |F_2(F_1(x)) - F_2(G_1(x))| + |F_2(G_1(x)) - G_2(G_1(x))| \\ &< \epsilon. \end{aligned}$$

□

## D.7 Examples of Flow Architectures Covered in Chapter 5

Here, we provide the proofs for the universal approximation properties of certain CF-INNs.

### D.7.1 Neural Autoregressive Flows (NAFs)

In this section, we prove that *neural autoregressive flows* [115] yield sup-universal approximators for $\mathcal{S}^1_{\mathrm{c}}$ (hence for $\mathcal{S}^\infty_{\mathrm{c}}$). The proof is not merely an application of a known result in Huang et al. [115] but it requires additional non-trivial consideration to enable the adoption of Lemma 3 in Huang et al. [115] as it is applicable only for those smooth maps that match certain boundary conditions.

**Definition D.8.** *A* deep sigmoidal flow *(DSF; a special case of neural autoregressive flows) [115, Equation (8)] is a flow layer $g = (g_1, \ldots, g_d) \colon \mathbb{R}^d \to \mathbb{R}^d$ of the following form:*

$$g_k(\boldsymbol{x}) := \sigma^{-1}\left(\sum_{j=1}^{n} w_{k,j}(\boldsymbol{x}_{\leq k-1}) \cdot \sigma\left(\frac{x_k - b_{k,j}(\boldsymbol{x}_{\leq k-1})}{\tau_j(\boldsymbol{x}_{\leq k-1})}\right)\right),$$

*where $\sigma$ is the sigmoid function, $n \in \mathbb{N}$, $w_j, b_j, \tau_j \colon \mathbb{R}^{k-1} \to \mathbb{R}$ $(j \in [n])$ are neural networks such that $b_j(\cdot) \in (r_0, r_1)$, $\tau_j(\cdot) \in (0, r_2)$, $w_j(\cdot) > 0$, and $\sum_{j=1}^n w_j(\cdot) = 1$ $(r_0, r_1 \in \mathbb{R},\ r_2 > 0)$. We define* DSF *to be the set of all possible DSFs.*

**Proposition D.4.** *The elements of* DSF *are locally bounded, and* $\mathrm{INN}_{\mathrm{DSF}}$ *is a* sup*-universal approximator for $\mathcal{S}^1_{\mathrm{c}}$.*

*Proof.* The elements of DSF are continuous, hence locally bounded. Let $s = (s_1, \cdots, s_d) \in \mathcal{S}^1_{\mathrm{c}}$. Take any compact set $K \subset \mathbb{R}^d$ and $\epsilon > 0$. Since $K$ is compact, there exist $r_0, r_1 \in \mathbb{R}$ such that $K \subset [r_0, r_1]^d$. Put $r_0' = r_0 - 1$, $r_1' = r_1 + 1$. We take a $C^1$-function $b \colon (r_0', r_1') \to \mathbb{R}$ satisfying

1. $b|_{[r_0, r_1]} = 0$,
2. $b|_{(r_0', r_0)}$ and $b|_{(r_1, r_1')}$ are strictly increasing,
3. $\lim_{x \to r_0' + 0} b(x) = -\infty$ and $\lim_{x \to r_1' - 0} b(x) = \infty$,
4. $\lim_{x \to r_0' + 0} \frac{d(\sigma \circ b)}{dx}(x)$ and $\lim_{x \to r_1' - 0} \frac{d(\sigma \circ b)}{dx}(x)$ exist in $\mathbb{R}$,

where $\sigma$ is the sigmoid function. For each $k \in [d]$, we define a $C^1$-function $\tilde{s}_k \colon [r_0', r_1']^{k-1} \times (r_0', r_1') \times [r_0', r_1']^{d-k} \to \mathbb{R}$, which is strictly increasing with respect to $x_k$, by

$$\tilde{s}_k(x) := s_k(x) + b(x_k) \quad (x = (x_1, \cdots, x_d)).$$

Moreover, we define a map $S \colon [r_0', r_1']^d \to [0,1]^d$ by

$$\begin{aligned} S_k|_{[r_0', r_1']^{k-1} \times (r_0', r_1') \times [r_0', r_1']^{d-k}} &= \sigma \circ \tilde{s}_k, \\ S_k(x_1, \cdots, x_{k-1}, r_0', x_{k+1}, \cdots, x_d) &= 0, \\ S_k(x_1, \cdots, x_{k-1}, r_1', x_{k+1}, \cdots, x_d) &= 1, \end{aligned}$$

where we write $S = (S_1, \cdots, S_d)$. Then, by Lemma D.15, $S$ satisfies the assumptions of Lemma 3 in [115]. Since $S([r_0, r_1]^d) \subset (0,1)^d$ is compact, there exists a positive number $\delta > 0$ such that

$$S([r_0, r_1]^d) + B(\delta) := \{S(x) + v \ : \ x \in [r_0, r_1]^d, v \in B(\delta)\} \subset [\delta, 1-\delta]^d,$$

where $B(\delta) := \{x \in \mathbb{R}^d : |x| \leq \delta\}$. Let $L > 0$ be a Lipschitz constant of $\sigma^{-1} \colon (0,1)^d \to \mathbb{R}^d$ on $[\delta, 1-\delta]^d$. By Lemma 3 in [115], there exists $g \in \mathrm{INN}_{\mathrm{DSF}}$ such that

$$\|S - \sigma \circ g\|_{\sup, [r_0', r_1']^d} < \min\left\{\delta, \frac{\epsilon}{L}\right\}.$$

As a result, $\sigma \circ g([r_0, r_1]^d) \subset S([r_0, r_1]^d) + B(\delta) \subset [\delta, 1-\delta]^d$. Then we obtain

$$\begin{aligned} \|s - g\|_{\sup, K} \leq \|s - g\|_{\sup, [r_0, r_1]^d} &= \|\sigma^{-1} \circ \sigma \circ s - \sigma^{-1} \circ \sigma \circ g\|_{\sup, [r_0, r_1]^d} \\ &\leq L \|S - \sigma \circ g\|_{\sup, [r_0, r_1]^d} \\ &< \epsilon. \end{aligned}$$

□

**Lemma D.15.** *We denote by $\mathcal{T}^1$ the set of all $C^1$-increasing triangular maps from $\mathbb{R}^d$ to $\mathbb{R}^d$. For $s = (s_1, \cdots, s_d) \in \mathcal{T}^1$, we define a map $S \colon [r_0', r_1']^d \to [0,1]^d$ as in the proof of Proposition D.4. Then $S$ is a $C^1$-map.*

*Proof.* It is enough to show that $S_d\colon [r_0', r_1']^d \to [0,1]$ is a $C^1$-function. We prove that for any $i \in [d]$, the $i$-th partial derivative of $S_d$ exists and that it is continuous on $[r_0', r_1']^d$. First, for $i \in [d-1]$, we consider the $i$-th partial derivative.

**Claim 1**.

$$\frac{\partial S_d}{\partial x_i}(x) = \begin{cases} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) & (x \in [r_0', r_1']^{d-1} \times (r_0', r_1')) \\ 0 & (x_d = r_0', r_1') \end{cases}$$

In fact, for $x \in [r_0', r_1']^{d-1} \times (r_0', r_1')$, we have

$$\frac{\partial S_d}{\partial x_i}(x) = \frac{\partial(\sigma \circ \tilde{s_d})}{\partial x_i}(x) = \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \left( \frac{\partial s_d}{\partial x_i}(x) + 0 \right).$$

For $x = (x_{\leq d-1}, r_0')$, we have

$$\begin{aligned} \frac{\partial S_d}{\partial x_i}(x) &= \lim_{h \to 0} \frac{S_d(x_{\leq i-1}, x_i + h, x_{i+1}, \cdots, x_{d-1}, r_0') - S_d(x_{\leq d-1}, r_0')}{h} \\ &= \lim_{h \to 0} \frac{0 - 0}{h} = 0 \end{aligned}$$

Here, note that by the definition of $S_d$, the notation

$$S_d(x_{\leq i-1}, x_i + h, x_{i+1}, \cdots, x_{d-1}, r_0')$$

makes sense even if $x_i = r_0'$ or $x_i = r_1'$. We can verify the case $x = (x_{\leq d-1}, r_1')$ similarly.

Next, we show that $\frac{\partial S_d}{\partial x_i}$ is continuous. We take any $x_{\leq d-1} \in [r_0', r_1']^{d-1}$. Since we have $\lim_{x \to r_0'} b(x) = -\infty$, $\lim_{x \to r_1'} b(x)$, $\lim_{x \to \pm\infty} \frac{d\sigma}{dx}(x) = 0$, and $|\frac{\partial s_d}{\partial x_I}(x)| < \infty$ $(x \in [r_0', r_1']^d)$, we obtain

$$\begin{aligned} \lim_{x \to (x_{d-1}, r_0')} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) = 0, \\ \lim_{x \to (x_{d-1}, r_1')} \frac{d\sigma}{dx}(s_i(x) + b(x_d))\frac{\partial s_d}{\partial x_i}(x) = 0. \end{aligned}$$

Therefore, the partial derivative $\frac{\partial S_d}{\partial x_i}(x)$ is continuous on $[r_0', r_1']^d$ for $i \in [d-1]$.

Next, we consider the $d$-th derivative of $S_d$.

**Claim 2.**

$$\frac{\partial S_d}{\partial x_d}(x) = \begin{cases} \frac{d\sigma}{dx}(s_d(x) + b(x_d)) \left( \frac{\partial s_d}{\partial x_d}(x) + \frac{db}{dx}(x_d) \right) & (x \in [r_0', r_1']^{d-1} \times (r_0', r_1')) \\ e^{s_d(x_{\leq d-1}, r_0')} \lim_{x \to r_0' + 0} \frac{d(\sigma \circ b)}{dx}(x) & (x_d = r_0') \\ e^{-s_d(x_{\leq d-1}, r_1')} \lim_{x \to r_1' - 0} \frac{d(\sigma \circ b)}{dx}(x) & (x_d = r_1') \end{cases}$$

We verify Claim 2. Since it is clear for the case $x \in [r_0', r_1']^{d-1} \times (r_0', r_1')$ by the definition of $S_k$, we consider the case $x_d = r_0', r_1'$.

**Subclaim.** For $x'_{\leq d-1} \in [r_0', r_1']^{d-1}$,

$$\begin{aligned} \lim_{x \to (x'_{\leq d-1}, r_0')} \frac{\sigma(s_d(x) + b(x_d))}{\sigma(b(x_d))} = e^{s_d(x'_{\leq d-1}, r_0')} \\ \lim_{x \to (x'_{\leq d-1}, r_1')} \frac{\sigma(s_d(x) + b(x_d)) - 1}{\sigma(b(x_d)) - 1} = e^{-s_d(x'_{\leq d-1}, r_1')} \end{aligned}$$

We verify this subclaim. From $\lim_{x\to r_0'} b(x) = -\infty$, we have

$$\frac{\sigma(s_d(x)+b(x_d))}{\sigma(b(x_d))} = \frac{1+e^{-b(x_d)}}{1+e^{-s_d(x)-b(x_d)}} = \frac{e^{b(x_d)}+1}{e^{b(x_d)}+e^{-s_d(x)}}$$
$$\to \frac{1}{e^{-s_d(x'_{\le d-1},r'_0)}} = e^{s_d(x'_{\le d-1},r'_0)} \quad (x \to (x'_{\le d-1}, r'_0))$$

Similarly, from $\lim_{x\to r_1'} b(x) = \infty$, we have

$$\frac{\sigma(s_d(x)+b(x_d))-1}{\sigma(b(x_d))-1} = e^{-s_d(x)}\frac{1+e^{-b(x_d)}}{1+e^{-s_d(x)-b(x_d)}}$$
$$\to e^{-s_d(x_{\le d-1},r'_1)} \quad (x \to (x'_{\le d-1}, r'_1)).$$

Therefore, our subclaim has been proved. By using L'Hôpital's rule, we have

$$\lim_{h\to+0}\frac{\sigma(b(r'_0+h))}{h} = \lim_{x\to r'_0}\frac{d(\sigma\circ b)}{dx}(x), \quad \lim_{x\to r'_1}\frac{\sigma(b(r'_1+h))-1}{h} = \lim_{x\to r'_1}\frac{d(\sigma\circ b)}{dx}(x).$$

Then, from Subclaim, we obtain

$$\begin{aligned}
\frac{\partial S_d}{\partial x_d}(x_{\le d-1}, r'_0) &= \lim_{h\to+0}\frac{\sigma(s_d(x_{\le d-1}, r'_0+h)+b(r'_0+h))-0}{h}\\
&= \lim_{h\to+0}\frac{\sigma(s_d(x_{\le d-1}, r'_0+h)+b(r'_0+h))}{\sigma(b(r_0+h))}\cdot\frac{\sigma(b(r'_0+h))}{h}\\
&= e^{s_d(x_{\le d-1},r'_0)}\lim_{x\to r'_0+0}\frac{d(\sigma\circ b)}{dx}(x),\\
\frac{\partial S_d}{\partial x_d}(x_{\le d-1}, r'_1) &= \lim_{h\to-0}\frac{\sigma(s_d(x_{\le d-1}, r'_1+h)+b(r'_1+h))-1}{h}\\
&= \lim_{h\to-0}\frac{\sigma(s_d(x_{\le d-1}, r'_1+h)+b(r'_1+h))-1}{\sigma(b(r'_1+h))-1}\cdot\frac{\sigma(b(r'_1+h))-1}{h}\\
&= e^{s_d(x_{\le d-1},r'_1)}\lim_{x\to r'_1}\frac{d(\sigma\circ b)}{dx}(x).
\end{aligned}$$

Therefore, Claim 2 was proved.

Finally, we verify $\frac{\partial S_d}{\partial x_d}(x)$ is continuous on $[r'_0, r'_1]^d$. Fix $x'_{\le d-1} \in [r'_0, r'_1]^{d-1}$. Since we have $\lim_{x\to(x'_{\le d-1},r'_0)} \frac{d\sigma}{dx}(\sigma_d(x)+b(x_d))\frac{\partial s_d}{\partial x_d}(x) = 0$, from Claim 2, it is enough to show the following:

**Claim 3.**

$$\lim_{x\to(x'_{\le d-1},r'_0)}\frac{d\sigma}{dx}(s_d(x)+b(x_d))\frac{db}{dx}(x_d) = e^{s_d(x_{\le d-1},r'_0)}\lim_{x\to r'_0+0}\frac{d(\sigma\circ b)}{dx}(x),$$
$$\lim_{x\to(x'_{\le d-1},r'_1)}\frac{d\sigma}{dx}(s_d(x)+b(x_d))\frac{db}{dx}(x_d) = e^{-s_d(x_{\le d-1},r'_1)}\lim_{x\to r'_1-0}\frac{d(\sigma\circ b)}{dx}(x).$$

We verify Claim 3. We have

$$\begin{aligned}
\frac{d\sigma}{dx}(s_d(x)+b(x_d))\frac{db}{dx}(x_d) &= \frac{\frac{d\sigma}{dx}(s_d(x)+b(x_d))}{\frac{d\sigma}{dx}(b(x_d))}\frac{d\sigma}{dx}(b(x_d))\frac{db}{dx}(x_d)\\
&= \frac{\frac{d\sigma}{dx}(s_d(x)+b(x_d))}{\frac{d\sigma}{dx}(b(x_d))}\frac{d(\sigma\circ b)}{dx}(x_d).
\end{aligned}$$

Since we have $\frac{d\sigma}{dx}(x) = \sigma(x)(1-\sigma(x))$, from Subclaim above, Claim 3 follows from

$$\frac{\frac{d\sigma}{dx}(s_d(x)+b(x_d))}{\frac{d\sigma}{dx}(b(x_d))} = \frac{\sigma(s_d(x)+b(x_d))}{\sigma(b(x_d))} \cdot \frac{1-\sigma(s_d(x)+b(x_d))}{1-\sigma(b(x_d))}$$
$$\to \begin{cases} e^{s_d(x'_{\le d-1}, r'_0)} & (x \to (x'_{\le d-1}, r'_0)) \\ e^{-s_d(x'_{\le d-1}, r'_1)} & (x \to (x'_{\le d-1}, r'_1)) \end{cases}.$$

Therefore, we proved the continuity of $\frac{\partial S_d}{\partial x_d}(x)$. □

### D.7.2 Sum-of-squares Polynomial Flows (SoS flows)

In this section, we prove that *sum-of-squares polynomial flows* [131] yield CF-INNs with the sup-universal approximation property for $\mathcal{S}_{\mathrm{c}}^1$ (hence for $\mathcal{S}_{\mathrm{c}}^\infty$). Even though Jaini et al. [131] claimed the distributional universality of the SoS flows by providing a proof sketch based on the univariate Stone-Weierstrass approximation theorem, we regard the sketch to be invalid or at least incomplete as it does not discuss the smoothness of the coefficients, i.e., whether the polynomial coefficients can be realized by continuous functions. Here, we provide complete proof that takes an alternative route to prove the sup-universality of the SoS flows via the multivariate Stone-Weierstrass approximation theorem.

**Definition D.9.** *A* sum-of-squares polynomial flow *(SoS flow) [131, Equation (9)] is a flow layer* $g = (g_1, \dots, g_d) \colon \mathbb{R}^d \to \mathbb{R}^d$ *of the following form:*

$$g_k(\boldsymbol{x}) := \mathfrak{B}_{2r+1}(x_k; C_k(\boldsymbol{x}_{\le k-1})),$$
$$\mathfrak{B}_{2r+1}(z; (c, \boldsymbol{a})) := c + \int_0^z \sum_{b=1}^{B} \left( \sum_{l=0}^{r} a_{l,b} u^l \right)^2 du,$$

*where* $C_k \colon \mathbb{R}^{k-1} \to \mathbb{R}^{B(r+1)+1}$ *is a neural network,* $r \in \mathbb{N} \cup \{0\}$*, and* $B \in \mathbb{N}$*. We define* SoS *to be the set of all possible SoS flows.*

**Proposition D.5.** *The elements of* SoS *are locally bounded, and* $\mathrm{INN}_{\mathrm{SoS}}$ *is a* sup*-universal approximator for* $\mathcal{S}_{\mathrm{c}}^1$*.*

*Proof.* The elements of SoS are continuous, hence locally bounded. The sup-universality follows from the Stone-Weierstrass approximation theorem as in the below. Let $s = (s_1, \dots, s_d) \in \mathcal{S}_{\mathrm{c}}^1$, a compact subset $K \subset \mathbb{R}^d$, and $\epsilon > 0$ be given. Then, there exists $R > 0$ such that $K \subset [-R, R]^d$. Since $s_d(\boldsymbol{x})$ is strictly increasing with respect to $x_d$ and $s$ is $C^1$, we have $\eta(\boldsymbol{x}) := \frac{\partial s_d}{\partial x_d}(\boldsymbol{x}) > 0$ and $\eta$ is continuous. Therefore, we can apply the Stone-Weierstrass approximation theorem [79, Corollary 4.50] to $\sqrt{\eta(\boldsymbol{x})}$: for any $\delta > 0$, there exists a polynomial $\pi(x_1, \dots, x_d)$ such that $\left\| \sqrt{\eta} - \pi \right\|_{\sup, [-R,R]^d} < \delta$. Then, by rearranging the terms, there exist $r \in \mathbb{N}$ and polynomials $\xi_l(x_1, \dots, x_{d-1})$ such that $\pi(x_1, \dots, x_d) = \sum_{l=0}^{r} \xi_l(x_1, \dots, x_{d-1}) x_d^l$. Now, define

$$\tilde{g}_d(\boldsymbol{x}) := s_d(\boldsymbol{x}_{\le d-1}, 0) + \int_0^{x_d} (\pi(\boldsymbol{x}_{\le d-1}, u))^2 du$$
$$= s_d(\boldsymbol{x}_{\le d-1}, 0) + \int_0^{x_d} \left( \sum_{l=0}^{r} \xi_l(x_1, \dots, x_{d-1}) u^l \right)^2 du$$

and $\tilde{g}(\boldsymbol{x}) := (x_1, \ldots, x_{d-1}, \tilde{g}_d(\boldsymbol{x}))$. Then,

$$\begin{aligned}
\|s - \tilde{g}\|_{\sup,K} &= \sup_{\boldsymbol{x}\in K} |s_d(\boldsymbol{x}) - \tilde{g}_d(\boldsymbol{x})| \\
&= \sup_{\boldsymbol{x}\in K} \left| s_d(\boldsymbol{x}_{\leq d-1}, 0) + \int_0^{x_d} \eta(\boldsymbol{x}_{\leq d-1}, u) du - \tilde{g}_d(\boldsymbol{x}) \right| \\
&= \sup_{\boldsymbol{x}\in K} \left| \int_0^{x_d} (\sqrt{\eta(\boldsymbol{x}_{\leq d-1}, u)}^2 - \pi(\boldsymbol{x}_{\leq d-1}, u)^2) du \right| \\
&\leq R \cdot \sup_{\boldsymbol{x}\in[-R,R]^d} \left| \sqrt{\eta(\boldsymbol{x})}^2 - \pi(\boldsymbol{x})^2 \right| \\
&= R \cdot \sup_{\boldsymbol{x}\in[-R,R]^d} |\sqrt{\eta(\boldsymbol{x})} + \pi(\boldsymbol{x})| \cdot |\sqrt{\eta(\boldsymbol{x})} - \pi(\boldsymbol{x})| \\
&\leq R \left( \sup_{\boldsymbol{x}\in[-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})} + \delta \right) \delta,
\end{aligned}$$

where we used

$$\begin{aligned}
\sup_{\boldsymbol{x}\in[-R,R]^d} |\sqrt{\eta(\boldsymbol{x})} + \pi(\boldsymbol{x})| &\leq \sup_{\boldsymbol{x}\in[-R,R]^d} |2\sqrt{\eta(\boldsymbol{x})}| + |\sqrt{\eta(\boldsymbol{x})} - \pi(\boldsymbol{x})| \\
&\leq \sup_{\boldsymbol{x}\in[-R,R]^d} 2\sqrt{\eta(\boldsymbol{x})} + \delta.
\end{aligned}$$

It is straightforward to show that there exists $g \in \mathrm{SoS}$ such that $\|\tilde{g} - g\|_{\sup,K} < \frac{\epsilon}{2}$ by approximating each of $s_d(\boldsymbol{x}_{\leq d-1})$ and $\xi_l$ on $K$ using neural networks. Finally, taking $\delta$ to be small enough so that $\|s - \tilde{g}\|_{\sup,K} < \frac{\epsilon}{2}$ holds, the assertion is proved. □

## D.8 Using Permutation Matrices Instead of Aff in the Definition of $\mathrm{INN}_{\mathcal{G}}$

In terms of representation power, there is no essential difference between using the permutation group and using the general linear group in Definition 5.1. In fact, one can express the elementary operation matrices (hence the regular matrices) by combining affine coupling flows, permutations.

From this result, we can see that employing Aff in Definition 5.1 instead of the permutation matrices is not an essential requirement for the universal approximation properties to hold. For this reason, we believe that the empirically reported difference in the performances of Glow [145] and RealNVP [63] is mainly in the efficiency of approximation rather than the capability of approximation.

**Lemma D.16.** *We have*

$$\mathrm{INN}_{\mathcal{H}\text{-ACF}} = \{W_1 \circ g_1 \circ \cdots \circ W_n \circ g_n \ : \ g_i \in \mathcal{H}\text{-ACF}, W_i \in \mathfrak{S}_d\}, \tag{D.1}$$

*where $\mathfrak{S}_d$ is the permutation group of degree $d$.*

*Proof.* Since any translation operator (i.e., addition of a constant vector) can be easily represented by the elements of $\mathcal{H}$-ACF and permutations, it is enough to show that any element of $\mathrm{GL}(n, \mathbb{R})$ can be realized by a finite composition of elements of $\mathcal{H}$-ACF and $\mathfrak{S}_d$. To show that, it is sufficient to consider only the elementary matrices. Row switching comes from $\mathfrak{S}_d$. Moreover, element-wise sign flipping can be described by a composition of finite elements of $\mathcal{H}$-ACF. To see this, first observe

that

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

holds. Now, any lower triangular matrix with positive diagonals can be described by a composition of finite elements of $\mathcal{H}$-ACF. Therefore, any diagonal matrix whose components are $\pm 1$ can be described by a composition of elements in $\mathcal{H}$-ACF and $\mathfrak{S}_d$. Therefore, any affine transform is an element of the right hand side of Equation (D.1). □

## D.9 Other Related Work

In this section, we elaborate on the relation of the present chapter and the existing literature.

**Approach to make universal approximators by augmenting the dimensionality.** Zhang et al. [308] showed that invertible residual networks (i-ResNets) [18] and neural ordinary differential equations (NODEs) [40, 67] can be turned into universal approximators of homeomorphisms by increasing the dimensionality and padding zeros. Given that, one may wonder if we can apply a similar technique to augment CF-INN to have the universality, which can bypass the proof techniques developed in this study. However, there is a problem that the approach can undermine the exact invertibility of the model: unless the model is ideally trained so that it always outputs zeros in the zero-padded dimensions, the model can no longer represent an invertible map operating on the original dimensionality. On the other hand, we showed the universality properties of certain CF-INNs without introducing the complication arising from the dimensionality augmentation.

# Bibliography


[1] A. Abadie and M. D. Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10(1):465–503, 2018.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access : practical innovations, open solutions*, 6:52138–52160, 2018.

[3] R. A. Adams and J. J. Fournier. *Sobolev Spaces*. Academic press, 2003.

[4] C. C. Aggarwal. *Outlier Analysis*. Springer International Publishing, second edition, 2017.

[5] R. A. C. Aguilar, D. G. Mahler, and D. Newhouse. Nowcasting Global Poverty. Working Paper Prepared for the IARIW-World Bank Conference, 2019.

[6] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press, 2008.

[7] L. Ardizzone, J. Kruse, C. Rother, and U. Köthe. Analyzing inverse problems with invertible neural networks. In *7th International Conference on Learning Representations*. OpenReview.net, 2019.

[8] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893 [cs, stat]*, 2020.

[9] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[10] A. Baker. Simplicity. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[11] B. Baltagi. *Econometric Analysis of Panel Data*. New York: John Wiley and Sons, 3rd edition, 2005.

[12] B. H. Baltagi and J. M. Griffin. Gasoline demand in the OECD: An application of pooling and testing procedures. *European Economic Review*, 22(2):117–137, 1983.

[13] A. Banyaga. Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique. *Commentarii Mathematici Helvetici*, 53(1):174–227, 1978.

[14] E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, 1342–1350, 2015.

[15] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[16] M. Bauer and A. Mnih. Resampled priors for variational autoencoders. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 66–75. PMLR, 2019.

[17] H. Beebee, C. Hitchcock, and P. Menzies, editors. *The Oxford Handbook of Causation*. Oxford University Press, 2009.

[18] J. Behrmann, W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*, 573–582. PMLR, 2019.

[19] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[20] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, 137–144. MIT Press, 2007.

[21] T. N. Beran and C. Violato. Structural equation modeling in medical research: a primer. *BMC Research Notes*, 3:Article 267, 2010.

[22] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, 1985.

[23] P. Bernard, V. Andrieu, and L. Praly. Expressing an observer in preferred coordinates by transforming an injective immersion into a surjective diffeomorphism. *SIAM Journal on Control and Optimization*, 56(3):2327–2352, 2018.

[24] R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv:2003.12659 [stat.ML]*, 2020.

[25] R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2314–2322. PMLR, 2021.

[26] R. Bisiani. Beam search. *Encyclopedia of Artificial Intelligence*:56–58, 1987.

[27] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 20*, 129–136. Curran Associates, Inc., 2008.

[28] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335, 2005.

[29] S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[30] S. Bonner and F. Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 104–112. Association for Computing Machinery, 2018.

[31] G. Borboudakis and I. Tsamardinos. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. In *Proceedings of the 29th International Conference on Machine Learning*, 427–434, 2012.

[32] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013.

[33] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[34] C. H. Brase and C. P. Brase. *Understanding Basic Statistics*. Cengage Learning, 2012.

[35] N. Cartwright. *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, 1999.

[36] A. Castañeda, C. Lakner, E. B. Prydz, J. S. Lopez, R. Wu, and Q. Zhao. Povcalnet: Stata module to access world bank global poverty and inequality measures. Statistical Software Components, Boston College Department of Economics, 2015.

[37] K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt. Unsupervised discovery of El Nino using causal feature learning on microlevel climate data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 72–81, 2016.

[38] K. Chalupka, F. Eberhardt, and P. Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

[39] K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 181–190, 2015.

[40] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, 6571–6583. Curran Associates, Inc., 2018.

[41] T. Chen and C. Guestrin. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. Association for Computing Machinery, 2016.

[42] Z. Chen and B. Liu. *Lifelong Machine Learning*, number 38 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, second edition, 2018.

[43] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[44] S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7801–7808, 2019.

[45] D. M. Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2002.

[46] Y. Chikahara, S. Sakaue, A. Fujino, and H. Kashima. Learning individually fair classifier with path-specific causal-effect constraint. In *International Conference on Artificial Intelligence and Statistics*, 145–153. PMLR, 2021.

[47] E. Çinlar. *Probability and Stochastics*. Springer New York, 2011.

[48] S. Clémençon, I. Colin, and A. Bellet. Scaling-up empirical risk minimization: optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016.

[49] P. Comon. Independent component analysis, a new concept? *Signal Processing*. Higher Order Statistics, 36(3):287–314, 1994.

[50] C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*. Algorithmic Learning Theory, 519:103–126, 2014.

[51] C. Cortes, M. Mohri, and A. M. Medina. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

[52] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[53] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems 30*, 3730–3739. Curran Associates, Inc., 2017.

[54] C. Craver and J. Tabery. Mechanisms in science. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019.

[55] C. F. Craver. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Clarendon Press, 2007.

[56] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.

[57] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–31, 1979.

[58] A. P. Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617, 1980.

[59] N. De Cao, W. Aziz, and I. Titov. Block neural autoregressive flow. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, 1263–1273. PMLR, 2020.

[60] G. De Pierris and M. Friedman. Kant and Hume on causality. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018.

[61] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*:1–1, 2021.

[62] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear independent components estimation. *arXiv:1410.8516 [cs.LG]*, 2014.

[63] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, Conference Track Proceedings*. OpenReview.net, 2017.

[64] J. Dony, U. Einmahl, and D. M. Mason. Uniform in bandwidth consistency of local polynomial regression function estimators. *Austrian Journal of Statistics*, 35(2&3):105–120, 2006.

[65] R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, second edition, 2002.

[66] O. D. Duncan, D. L. Featherman, and B. Duncan. *Socioeconomic Background and Achievement*. Socioeconomic Background and Achievement. Seminar Press, 1972.

[67] E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems 32*, 3140–3150. Curran Associates, Inc., 2019.

[68] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems 32*, 7511–7522. Curran Associates, Inc., 2019.

[69] U. Einmahl and D. M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, 13(1):1–37, 2000.

[70] U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33(3):1380–1403, 2005.

[71] D. B. A. Epstein. The simplicity of certain groups of homeomorphisms. *Compositio Mathematica*, 22(2):165–173, 1970.

[72] R. Evans. Markov properties for mixed graphical models. In *Handbook of Graphical Models*. Chapman & Hall/CRC, 2019.

[73] R. Evans and T. Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, 42(4):1452–1482, 2014.

[74] R. J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

[75] R. J. Evans. Margins of discrete Bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[76] R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848–876, 2019.

[77] Y. Fan, J. Chen, G. Shirkey, R. John, S. R. Wu, H. Park, and C. Shao. Applications of structural equation modeling (SEM) in ecological studies: an updated review. *Ecological Processes*, 5:Article 19, 2016.

[78] J. A. Fodor. Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, 28(2):97–115, 1974.

[79] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*, number 125 in Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts Book. Wiley, second edition, 1999.

[80] P. Forré and J. M. Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv:1710.08775 [math, stat]*, 2017.

[81] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[82] A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems 30*, 3011–3021. Curran Associates, Inc., 2017.

[83] S. Glennan. Chapter 15: Mechanisms. In *The Oxford Handbook of Causation*. Oxford University Press, 2009.

[84] S. Glennan. Mechanisms and the nature of causation. *Erkenntnis*, 44(1):49–71, 1996.

[85] S. Glennan. Productivity, relevance and natural selection. *Biology & Philosophy*, 24(3):325–339, 2009.

[86] S. Glennan. Rethinking mechanistic explanation. *Philosophy of Science*, 69(3):S342–S353, 2002.

[87] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256. JMLR Workshop and Conference Proceedings, 2010.

[88] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:Article 524, 2019.

[89] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, fourth edition, 2013.

[90] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich. Causal generative domain adaptation networks. *arXiv:1804.04333 [cs, stat]*, 2018.

[91] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the 33rd International Conference on Machine Learning*, 2839–2848. PMLR, 2016.

[92] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[93] W. H. Greene. *Econometric Analysis*. Prentice Hall, seventh edition, 2012.

[94] T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.

[95] S. Haller. *Groups of Diffeomorphisms*. Magister, University of Vienna, Vienna, Austria, 1995.

[96] J. Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12(1):317–337, 2000.

[97] J. Y. Halpern and J. Pearl. Causes and explanations: a structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

[98] D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[99] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[100] T. Hayfield and J. S. Racine. Nonparametric econometrics: the np package. *Journal of Statistical Software*, 27(5):1–32, 2008.

[101] L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv:1907.02435 [math, stat]*, 2020.

[102] L. Henderson. The problem of induction. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

[103] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[104] M. W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics*. Springer-Verlag, first edition, 1976.

[105] C. Hitchcock. Causal models. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.

[106] C. Hitchcock. Chapter 14: Causal modelling. In *The Oxford Handbook of Causation*. Oxford University Press, 2009.

[107] C. Hitchcock. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6):273–299, 2001.

[108] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning*, 2722–2730. PMLR, 2019.

[109] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[110] J. Hoogeveen and U. Pape. *Data Collection in Fragile States: Innovations from Africa and Beyond*. Springer Nature, 2020.

[111] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[112] L. Horváth and B. S. Yandell. Asymptotics of conditional empirical processes. *Journal of Multivariate Analysis*, 26(2):184–206, 1988.

[113] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schoelkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, 689–696, 2009.

[114] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1551–1560. ACM Press, 2018.

[115] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning*, chapter Machine Learning, 2078–2087. PMLR, 2018.

[116] Y. Huang and M. Valtorta. On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, 54(4):363–408, 2008.

[117] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[118] D. Hume. *An Enquiry Concerning Human Understanding*. 1748.

[119] P. Hünermund and E. Bareinboim. Causal inference and data-fusion in econometrics. arXiv:1912.09104 [econ.EM], 2019.

[120] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[121] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[122] A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 460–469, 2017.

[123] A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29*, 3765–3773. Curran Associates, Inc., 2016.

[124] A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 859–868, 2019.

[125] A. Hyvärinen and S. M. Smith. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

[126] G. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[127] G. W. Imbens. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. Working Paper 294, National Bureau of Economic Research, 2003.

[128] N. Indurkhya and F. J. Damerau, editors. *Handbook of Natural Language Processing*. Machine Learning & Pattern Recognition Series. Chapman and Hall/CRC, second edition, 2010.

[129] I. C. F. Ipsen and R. Rehman. Perturbation bounds for determinants and characteristic polynomials. *SIAM Journal on Matrix Analysis and Applications*, 30(2):762–776, 2008.

[130] P. Izmailov, P. Kirichenko, M. Finzi, and A. G. Wilson. Semi-supervised learning with normalizing flows. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[131] P. Jaini, K. A. Selby, and Y. Yu. Sum-of-squares polynomial flow. In *Proceedings of the 36th International Conference on Machine Learning*, 3009–3018. PMLR, 2019.

[132] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[133] D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[134] Y. Jung, J. Tian, and E. Bareinboim. Estimating identifiable causal effects on Markov equivalence class through double machine learning. In *Proceedings of the 38th International Conference on Machine Learning*, 5168–5179. PMLR, 2021.

[135] Y. Jung, J. Tian, and E. Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 12113–12122, 2021.

[136] M. Kalisch and P. Bühlmann. Causal Structure Learning and Inference: A Selective Review. *Quality Technology & Quantitative Management*, 11(1):3–21, 2014.

[137] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on the Science of Modeling, the 30th Anniversary of the Information Criterion,* 261–270, 2003.

[138] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362. Association for Computing Machinery, 2021.

[139] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, 265–277. Curran Associates, Inc., 2020.

[140] H. H. Keller. The SCREEN I (Seniors in the Community: Risk Evaluation for Eating and Nutrition) index adequately represents nutritional risk. *Journal of Clinical Epidemiology*, 59(8):836–841, 2006.

[141] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ICA: a unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR, 2020.

[142] H. Kiiveri, T. P. Speed, and J. B. Carlin. Recursive causal models. *Journal of the Australian Mathematical Society*, 36(1):30–52, 1984.

[143] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon. FloWaveNet : A generative flow for raw audio. In *Proceedings of the 36th International Conference on Machine Learning*, 3370–3378. PMLR, 2019.

[144] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference for Learning Representations*, 2015.

[145] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, 10215–10224. Curran Associates, Inc., 2018.

[146] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29*, 4743–4751. Curran Associates, Inc., 2016.

[147] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer International Publishing, 2020.

[148] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker. Normalizing flows: an introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

[149] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[150] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[151] W. Kumagai. Learning bound for parameter transfer learning. In *Advances in Neural Information Processing Systems 29*, 2721–2729. Curran Associates, Inc., 2016.

[152] S. Kuroki, N. Charoenphakdee, H. Bao, J. Honda, I. Sato, and M. Sugiyama. Unsupervised domain adaptation based on source-guided discrepancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4122–4129, 2019.

[153] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, 4066–4076. Curran Associates, Inc., 2017.

[154] D. Kutach. *Causation*. Key Concepts in Philosophy. Polity, 2014.

[155] T. Kyono and M. van der Schaar. Improving model robustness using causal knowledge. *arXiv:1911.12441 [cs.LG]*, 2019.

[156] T. Kyono, Y. Zhang, and M. van der Schaar. CASTLE: Regularization via auxiliary causal graph discovery. In *Advances in Neural Information Processing Systems 33*, 1501–1512. Curran Associates, Inc., 2020.

[157] F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 29*, 1181–1189. Curran Associates, Inc., 2016.

[158] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

[159] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[160] S. L. Lauritzen. *Graphical Models*, number 17 in Oxford Statistical Science Series. Clarendon Press, 1996.

[161] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[162] A. J. Lee. *U-Statistics: Theory and Practice*. CRC Press, 1990.

[163] H. Lee, R. Raina, A. Teichman, and A. Y. Ng. Exponential family sparse coding with applications to self-taught learning. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, 1113–1119. Morgan Kaufmann Publishers Inc., 2009.

[164] S. Lee and E. Bareinboim. Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems 31*, 2568–2578. Curran Associates, Inc., 2018.

[165] D. Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

[166] M. Lewis and A. Kuerbis. An overview of causal directed acyclic graphs for substance abuse researchers. *Journal of Drug and Alcohol Research*, 5:1–8, 2016.

[167] H. Lin and S. Jegelka. ResNet with one-neuron hidden layers is a universal approximator. In *Advances in Neural Information Processing Systems 31*, 6172–6181. Curran Associates, Inc., 2018.

[168] P. Lipton. Chapter 29: causation and explanation. In *The Oxford Handbook of Causation*. Oxford University Press, 2009.

[169] M. A. Little and R. Badawy. Causal bootstrapping. *arXiv:1910.09648 [cs.LG]*, 2020.

[170] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: a survey. *International Journal of Computer Vision*, 128(2):261–318, 2020.

[171] E.-M. Løberg, H. A. Jørgensen, M. F. Green, B. R. Rund, A. Lund, A. Diseth, M. Oie, and K. Hugdahl. Positive symptoms and duration of illness predict functional laterality and attention modulation in schizophrenia. *Acta Psychiatrica Scandinavica*, 113(4):322–331, 2006.

[172] D. Lopez-paz, J. M. Hernández-lobato, and B. Schölkopf. Semi-supervised domain adaptation with non-parametric copulas. In *Advances in Neural Information Processing Systems 25*, 665–673. Curran Associates, Inc., 2012.

[173] C. Louizos and M. Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2218–2227. PMLR, 2017.

[174] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*. Bayesian Networks in Biomedicince and Health-Care, 30(3):201–214, 2004.

[175] P. Machamer, L. Darden, and C. F. Craver. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.

[176] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31*, 10846–10856. Curran Associates, Inc., 2018.

[177] D. G. Mahler, R. A. Castaneda Aguilar, and D. Newhouse. Nowcasting Global Poverty. Working Paper, World Bank, Washington, DC, 2021.

[178] C. F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, 2003.

[179] J. N. Mather. Commutators of diffeomorphisms. *Commentarii mathematici Helvetici*, 49(1):512–528, 1974.

[180] J. N. Mather. Commutators of diffeomorphisms: II. *Commentarii Mathematici Helvetici*, 50(1):33–40, 1975.

[181] P. Menzies. The causal structure of mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4):796–805, 2012.

[182] P. Menzies and H. Beebee. Counterfactual theories of causation. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition, 2020.

[183] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

[184] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, second edition, 2018.

[185] R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, 186–195, 2019.

[186] J. M. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

[187] J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

[188] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

[189] W. E. Morris and C. R. Brown. David Hume. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 edition, 2021.

[190] S. Mumford and R. L. Anjum. *Causation: A Very Short Introduction*. Very Short Introductions. Oxford University Press, 2013.

[191] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[192] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan. Hybrid models with deep and invertible features. In *Proceedings of the 36th International Conference on Machine Learning*, 4723–4732. PMLR, 2019.

[193] A. Y. Ng. Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 245–253. Morgan Kaufmann Publishers Inc., 1997.

[194] T. D. Nguyen, M. Christoffel, and M. Sugiyama. Continuous target shift adaptation in supervised learning. In *Asian Conference on Machine Learning*, 285–300. PMLR, 2016.

[195] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, chapter Machine Learning, 3918–3926. PMLR, 2018.

[196] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[197] G. Papa, S. Clémençon, and A. Bellet. SGD algorithms based on incomplete U-statistics: large-scale minimization of empirical risk. In *Advances in Neural Information Processing Systems 28*, 1027–1035. Curran Associates, Inc., 2015.

[198] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[199] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems 30*, 2338–2347. Curran Associates, Inc., 2017.

[200] D. Pardoe and P. Stone. Boosting for regression transfer. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, 863–870, 2010.

[201] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc., 2019.

[202] J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article 7, 2010.

[203] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[204] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, second edition, 2009.

[205] J. Pearl. Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[206] J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: a formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 247–254. AAAI Press, 2011.

[207] J. Perktold, S. Seabold, J. Taylor, and statsmodels-developers. Statsmodels. GitHub, 2020.

[208] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, 2017.

[209] J. Peters, J. Mooij, D. Janzing, and B. Schoelkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 589–598, 2011.

[210] J. Peters and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(June):2009–2053, 2014.

[211] S. Pitis, E. Creager, and A. Garg. Counterfactual data augmentation using locally factored dynamics. In *Advances in Neural Information Processing Systems 33*, 2020.

[212] J. R. Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, 236–243. Morgan Kaufmann Publishers Inc., 1993.

[213] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors. *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press, 2009.

[214] R Core Team. *R: A Language and Environment for Statistical Computing*. 2018.

[215] P. C. Reiss and F. A. Wolak. Structural econometric modeling: rationales and examples from industrial organization. In *Handbook of Econometrics*. Volume 6, 4277–4415. Elsevier, 2007.

[216] W. Rejchel. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(May):1373–1392, 2012.

[217] T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

[218] T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv:1701.06686 [stat.ME]*, 2017.

[219] R. M. Rifkin and R. A. Lippert. Notes on Regularized Least Squares. Technical Report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2007.

[220] S. Robbiano and J. Tressou. Maximal deviations of incomplete U-statistics with applications to empirical risk sampling. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 19–27. Society for Industrial and Applied Mathematics, 2013.

[221] J. M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[222] J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*. Oxford University Press, 2011.

[223] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[224] A. Rønn-Nielsen and E. Hansen. *Conditioning and Markov Properties*. Department of Mathematical Sciences, University of Copenhagen, 2014.

[225] A. Rotnitzky and E. Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

[226] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[227] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[228] W. C. Salmon. The uniformity of nature. *Philosophy and Phenomenological Research*, 14(1):39–48, 1953.

[229] M. K. Santos, J. R. Ferreira Júnior, D. T. Wada, A. P. M. Tenório, M. H. N. Barbosa, and P. M. d. A. Marques. Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. *Radiologia Brasileira*, 52(6):387–396, 2019.

[230] B. Schölkopf. Causality for machine learning. *arXiv:1911.10500 [cs, stat]*, 2019.

[231] B. Schölkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of the 32nd International Conference on Machine Learning*, 2218–2226. PMLR, 2015.

[232] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on Machine Learning*, 459–466. Omnipress, 2012.

[233] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[234] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, 2002.

[235] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12*, 582–588. MIT Press, 2000.

[236] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[237] R. P. Sherman. Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459, 1994.

[238] S. Shimizu. Non-Gaussian structural equation models for causal discovery. In *Statistics and Causality*, 153–184. John Wiley & Sons, Inc., 2016.

[239] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

[240] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(33):1225–1248, 2011.

[241] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[242] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):Article 60, 2019.

[243] I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[244] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 437–444. AUAI Press, 2006.

[245] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[246] I. Shpitser and J. Tian. On identifying causal effects. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, 523–543. College Publication, 2010.

[247] J. Sill. Monotonic networks. In *Advances in Neural Information Processing Systems 10*, 661–667. MIT Press, 1998.

[248] R. Silva, C. Blundell, and Y. W. Teh. Mixed cumulative distribution networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 670–678. JMLR Workshop and Conference Proceedings, 2011.

[249] R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10(41):1187–1238, 2009.

[250] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, 1st edition, 1986.

[251] E. Smucler, F. Sapienza, and A. Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *Biometrika*, 109(1):49–65, 2021.

[252] P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[253] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, 2000.

[254] P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 499–506. Morgan Kaufmann Publishers Inc., 1995.

[255] J. Splawa-Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by D. M. Dabrowska and T. P. Speed. *Statistical Science*, 5(4):465–472, 1990.

[256] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980.

[257] P. Stojanov, M. Gong, J. Carbonell, and K. Zhang. Data-driven approach to multiple-source domain adaptation. In *Proceedings of Machine Learning Research*, 3487–3496. PMLR, 2019.

[258] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[259] A. J. Storkey and M. Sugiyama. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems 19*, 1337–1344. MIT Press, 2007.

[260] W. Stute. Conditional empirical processes. *Annals of Statistics*, 14(2):638–647, 1986.

[261] N. Subramani. Pag2admg: an algorithm for the complete causal enumeration of a Markov equivalence class. In *Proceedings of the CausalML Workshop at ICML*, 2018.

[262] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012.

[263] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[264] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

[265] K. M. Takeshita, T. I. Hayashi, and H. Yokomizo. The effect of intervention in nickel concentrations on benthic macroinvertebrates: A case study of statistical causal inference in ecotoxicology. *Environmental Pollution*, 265:115059, 2020.

[266] E. Taskesen. Bnlearn - Library for Bayesian network learning and inference. GitHub, 2020.

[267] T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *Proceedings of the 37th International Conference on Machine Learning*, 9458–9469, 2020.

[268] J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

[269] P. Thagard. *How Scientists Explain Disease*. Princeton University Press, 1999.

[270] W. Thurston. Foliations and groups of diffeomorphisms. *Bulletin of the American Mathematical Society*, 80(2):304–307, 1974.

[271] J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, University of California, Los Angeles, Los Angeles, CA, 2002.

[272] J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, 567–573. American Association for Artificial Intelligence, 2002.

[273] J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 519–527. Morgan Kaufmann Publishers Inc., 2002.

[274] I. Tsamardinos and C. Aliferis. Towards principled feature selection: relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers, 2003.

[275] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[276] United Nations Statistical Commission and United Nations Economic Commission for Europe. *Terminology on Statistical Metadata*. United Nations, 2000.

[277] S. Van Buuren. *Flexible Imputation of Missing Data*. CRC press, 2012.

[278] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4*, 831–838. Morgan-Kaufmann, 1992.

[279] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, 2000.

[280] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[281] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: a review. *arXiv:2010.10596 [cs, stat]*, 2020.

[282] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, 255–270. Elsevier Science Inc., 1990.

[283] T. S. Verma. Graphical Aspects of Causal Models. Technical Report R-191, Computer Science Department, University of California, Los Angeles, 1993.

[284] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33*, 2020.

[285] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[286] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1st edition, 2019.

[287] Y. Wang, D. Liang, L. Charlin, and D. M. Blei. Causal inference for recommender systems. In *Fourteenth ACM Conference on Recommender Systems*, 426–431. ACM, 2020.

[288] P. N. Ward, A. Smofsky, and A. J. Bose. Improving exploration in soft-actor-critic with normalizing flows policies. *arXiv:1906.02771 [cs, stat]*, 2019.

[289] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.

[290] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[291] W. Wiedermann and A. von Eye. *Statistics and Causality: Methods for Applied Empirical Research*. Wiley, 1st edition, 2016.

[292] T. C. Williams, C. C. Bach, N. B. Matthiesen, T. B. Henriksen, and L. Gagliardi. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research*, 84(4):487–493, 2018.

[293] J. Witte, L. Henckel, M. H. Maathuis, and V. Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

[294] J. Woodward. Causation and manipulability. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[295] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

[296] J. Woodward. Law and explanation in biology: invariance is the kind of stability that matters. *Philosophy of Science*, 68(1):1–20, 2001.

[297] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

[298] Y. Wu, S. Srivastava, N. Hay, S. Du, and S. Russell. Discrete-continuous mixtures in probabilistic programming: generalized semantics and inference algorithms. In *International Conference on Machine Learning*, 5343–5352. PMLR, 2018.

[299] Y. Wu, L. Zhang, and X. Wu. Counterfactual fairness: unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 1438–1444. International Joint Conferences on Artificial Intelligence Organization, 2019.

[300] Y. Wu, L. Zhang, and X. Wu. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2536–2544. Association for Computing Machinery, 2018.

[301] Y. Wu, L. Zhang, X. Wu, and H. Tong. PC-Fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[302] L. Xu, T. Fan, X. Wu, K. Chen, X. Guo, J. Zhang, and L. Yao. A pooling-LiNGAM algorithm for effective connectivity analysis of fMRI data. *Frontiers in Computational Neuroscience*, 8:Article 125, 2014.

[303] O. Yadan. Hydra: a framework for elegantly configuring complex applications. GitHub, 2019.

[304] P. Yadav, M. Steinbach, V. Kumar, and G. Simon. Mining electronic health records (EHRs): a survey. *ACM Computing Surveys*, 50(6):1–40, 2018.

[305] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems 24*, 594–602. Curran Associates, Inc., 2011.

[306] D. Yu and L. Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. Springer-Verlag, 1st edition, 2015.

[307] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu. Causality-based feature selection: methods and evaluations. *ACM Computing Surveys*, 53(5):1–36, 2020.

[308] H. Zhang, X. Gao, J. Unterman, and T. Arodz. Approximation capabilities of neural ODEs and invertible residual networks. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

[309] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008.

[310] J. Zhang and E. Bareinboim. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1340–1346, 2017.

[311] K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: a causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3150–3157. AAAI Press, 2015.

[312] K. Zhang and A. Hyvärinen. Nonlinear functional causal models for distinguishing cause from effect. In *Statistics and Causality*, 185–201. John Wiley & Sons, Inc., 2016.

[313] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, 819–827, 2013.

[314] Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, 7404–7413. PMLR, 2019.

[315] C. Zhou, X. Ma, D. Wang, and G. Neubig. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1588–1598. Association for Computational Linguistics, 2019.

[316] Z. Ziegler and A. Rush. Latent normalizing flows for discrete sequences. In *Proceedings of the 36th International Conference on Machine Learning*, chapter Machine Learning, 7673–7682. PMLR, 2019.

[317] V. A. Zorich. *Mathematical Analysis I*. Universitext. Springer Berlin Heidelberg, second edition, 2015.