博士論文

Learning High-dimensional Models with the Minimum Description Length Principle
(記述長最小化原理による高次元モデル学習)

Kohei Miyaguchi (宮口航平)

Abstract

High-dimensional models that have hundreds of thousands of parameters such as deep neural networks and sparse models are effective in machine learning and data mining tasks. Controlling the complexity of such high-dimensional models is necessary for attaining appropriate inductive inference, e.g., preventing overfit and making it easier to interpret the results.

On the contrary, there are numerous different principles for measuring the complexity of models. The minimum description length (MDL) principle is an information-theoretic principle proposed by Rissanen (1978), of which one salient feature is offering a unified framework of inductive inference without imposing any assumptions on the distribution of data. According to the MDL principle, the complexity of models is quantified depending on the minimax-regret code length. However, for high-dimensional models, the computation of the exact code length is intractable and no analytic approximation method has been implemented till date to resolve this issue. This is problematic in terms of two basic tasks of inductive inference, namely model selection and prediction: (i) High-dimensional model selection is difficult since the code length of each candidate model is intractable. (ii) Even if it is numerically tractable, designing high-dimensional prediction algorithms is difficult as the code length cannot be analytically evaluated.

Considering this, in this thesis, we propose three approaches to the problems of high-dimensional inductive inference under the MDL principle. (i) We address the problem of high-dimensional model selection over exponentially many candidates leveraging the continuous relaxation of the minimax code lengths. The proposed algorithm overcomes the computational difficulty minimizing code lengths without computing them but sampling the stochastic gradients. (ii) We study the minimax code length of smooth models to derive a new analytic approximation. We demonstrate its effectiveness through the problem of hyperparameter selection. (iii) We study a novel complexity measure, namely the envelope complexity, that provides a more general framework for the analytic approximation of the minimax code length. Its power is demonstrated by deriving an adaptive minimax predictor over high-dimensional ℓ_1-balls and systematic upper bounds on predictive risks.

These three approaches provide the tools and foundations for the MDL principle to deal with high-dimensional modeling and prediction.

Contents

Chapter 1

Introduction

Herein, we formulate the problems that we have considered in this thesis. To this end, we first present the notions of inductive inference and the minimax regret principle. Then, we introduce the minimum description length (MDL) principle, which is an information-theoretic instance of the minimax regret principle. Finally, we list our research questions and summarize our contributions with regard to the high-dimensional learning problems based on the MDL principle.

1.1 Inductive Inference and Minimax Regret Principle

We are interested in methods for learning laws and regularities in data, i.e., inference by induction. For example, one may conclude by induction that all crows are black after observing that 100 random crows are black. The conclusion may be incorrect owing to some unobserved exceptions, but it enables us to learn from our experience and generalize them as a piece of knowledge. Therefore, induction forms an essential building block of any intelligent systems that can learn from their experience and has been one of the central subjects of research in many fields, including statistics, machine learning, data mining, and artificial intelligence.

1.1.1 Elements of Inductive Inference

Inductive inference systems may be characterized with four elements interacting with each other (see Figure 1.1). The goal of inductive inference is to infer the nature of *sources* $\mathcal{S}$ on the basis of *data* $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ generated by $\mathcal{S}$ itself. In doing so, there are *models* $\mathcal{H}$ in our mind that describe $\mathcal{S}$ and they derive the corresponding *predictors* (or algorithms) A, which are designed in the light of $\mathcal{H}$ to make the best prediction on X_{i+1}, given X^i $(0 \leq i \leq n)$.

For instance, each datum $X \in \mathcal{X}$ may be a single feature vector (i.e., unsupervised learning setting) or a pair of a feature vector $U \in \mathcal{U}$ and a label/response variable $Y \in \mathcal{Y}$ (i.e., supervised learning setting). Models are assumed to be represented with sets of hypotheses, $\mathcal{H} = \{h = h_\theta \mid \theta \in \Omega\}$, that express prior beliefs about the possible laws behind the source $\mathcal{S}$. Here, each hypothesis h is either a probability distribution over $\mathcal{X}$ $P(X|h)$ (unsupervised learning) or a regressor or classifier $h : \mathcal{U} \to \mathcal{Y}$ (supervised learning). Correspondingly, predictors A are defined as mappings from the data to the hypotheses, $A : \mathcal{X}^* \to \mathcal{H}_0$, where $\mathcal{H}_0$ is the universal set of hypotheses, including the model $\mathcal{H}$. The quality of the outputs of predictors is measured through some extended real-valued loss function $f : \mathcal{H}_0 \times \mathcal{X} \to \overline{\mathbb{R}}(= \mathbb{R} \cup \{\infty\})$.

These four elements of inductive inference is distinguished through two functional attributes as illustrated in the diagram: The subjects to the left belong to environments and others that we do not have control over, while the ones on the right belong to self and what we have control over. The subjects to the bottom are materialistic and quantifiable, whereas the ones on the top represent the corresponding mental models where inductive inference occurs.

$$\begin{tikzcd}
\mathcal{S} \arrow[d, squiggly, "\text{Generate}"'] & \mathcal{H} \arrow[l, dashed, "\text{Describe}"'] \arrow[d, squiggly, "\text{Derive}"] \\
X^n & A \arrow[l, "\text{Predict}"]
\end{tikzcd}$$

Fig. 1.1: Elements of inductive inference

The goal of inductive inference is usually either modeling or prediction. In modeling, one wants to construct good models $\mathcal{H}$ that represent knowledge and truths on $\mathcal{S}$, which is only validated through the performance of A with respect to X^n. On the contrary, in prediction, one wants to make a good prediction on unseen data X_{n+1} generated from the same source $\mathcal{S}$. As no predictor performs universally well (e.g., see the no-free-lunch theorem in Shalev-Shwartz and Ben-David (2014), Chapter 5), good predictors A cannot be designed without the help of good models $\mathcal{H}$ on the nature of $\mathcal{S}$.

1.1.2 Principles of Inductive Inference

Despite the predictors and the models being two sides of the same coin, different predictors can be derived from the same model; conversely, the same predictor can be evaluated in different ways. In terms of prediction, this is because the models often have uncertainty (remember that models are just sets of possible hypotheses) and the derivation of predictors is dependent on how we deal with the uncertainty. In contrast, in terms of modeling, this is because the goodness of algorithms depends on the purpose of modeling even if the same predictor A and the same data X^n are given.

These uncertainties are solved using the principles of inductive inference. Each principle is categorized as either one for modeling or one for prediction. The principles for modeling provide a mapping $score : (\mathcal{H}, X^n) \to \mathbb{R}$ from models and data to the scores that quantify the goodness of models (smaller is better). The mapping is often written as a function of predictors and data, $score : (A, X^n) \mapsto \mathbb{R}$, when a predictor has been already associated with the given model $\mathcal{H}$. On the contrary, the principles for prediction give a mapping $derive : \mathcal{H} \mapsto A$ from models to predictors.

Below, we review some of the common principles of inductive inference. First, we start with the principles for modeling.

Validation

One of the most common and simple principles for modeling is based on validation (Geisser, 2017). In validation, the given data are split into training sets X_{train} and validation sets $X_{\text{validation}}$. The validation sets must be held unseen and used solely for validating the performance of predictors. Thus, with some principle for prediction, the goodness of models is measured through the validation error given by $\mathit{score} : (A, X^n) = f(A(X_{\mathrm{train}}), X_{\mathrm{validation}})$ where f denotes some loss function.

One of the largest advantages of validation is that one can obtain unbiased estimates for predictive performance under mild assumptions on the data distributions. A major drawback is that a part of the given data cannot be used for prediction as it is sacrificed for validation. This is problematic because if we take a large portion of data as the validation set, additional costs of data collection to feed a sufficient number of training samples to the predictor must be paid or the quality of the prediction degenerates; On the contrary, if we take a small portion of data as the validation set, the variance of the validation errors increases and it is unreliable as a measure of the goodness. The cross-validation technique can be used to solve this trade-off by reusing validation sets as training sets. However, cross-validation is computationally expensive and difficult to be used with large-scale models and datasets.

Akaike Information Criterion (AIC)

Akaike's information criterion (AIC) (Akaike, 1974) can be seen as the asymptotic approximation of validation. Under stronger conditions on the data source $\mathcal{S}$ and the model $\mathcal{H}$ including the i.i.d.ness of data distributions, AIC computes an asymptotically unbiased estimate of the predictive performance of models *without* sacrificing any portions of data for validation. The scoring with AIC is given by

$$AIC(\mathcal{H}, X^n) = \min_{h \in \mathcal{H}} f(h, X^n) + d$$

where d is the dimensionality of the models $\mathcal{H}$.

Bayesian Information Criterion (BIC)

Bayesian information criterion (BIC) (Schwarz et al., 1978) was introduced as an asymptotic approximation of the Bayesian marginal likelihood, which is considered as a decision criteria under some circumstances. The scoring with BIC is given by

$$BIC(\mathcal{H}, X^n) = \min_{h \in \mathcal{H}} f(h, X^n) + \frac{d}{2} \ln n$$

Under some regularity conditions, the models chosen with BIC is known to be *consistent*, i.e., if $\mathcal{H}_1$ contains the correct hypothesis and $\mathcal{H}_2$ does not, then the BIC of $\mathcal{H}_1$ is strictly smaller than that of $\mathcal{H}_2$ with arbitrarily high probability if n is sufficiently large.

Prequential Principle

The prequential principle is one of the most general principles of modeling proposed by Dawid (1984), which is applicable without any assumptions on data distributions $\mathcal{S}$ (though it is developed in a statistical context). According to this principle, the goodness of models should be measured only with what they actually predict and the actual outcomes. More specifically, it is proposed that the goodness of predictors be measured with the prequential error, or cumulative loss,

$$f(A, X^n) \overset{\mathrm{def}}{=} \sum_{i=1}^{n} f(A(X^{i-1}), X_i).$$

We note that this measure can be seen as a generalization of the Bayesian marginal likelihood. Thus, the empirical Bayes method is an instance of the prequential principle; moreover, it is connected with BIC in an asymptotic manner.

Probably Approximately Correct (PAC) Risk Bounds

Probably approximately correct (PAC) risk bounds are often used in theoretical studies under the PAC-learning regime (Valiant, 1984) to provide the upper bounds on the predictive performance of given models or predictors. Popular instances of such bounds are the Rademacher complexity (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2002) and the PAC–Bayes bound (McAllester, 1999; Catoni, 2007). With either bounds, under some assumptions on the data distribution, one can bound above the predictive performance of given predictors on unseen data X_{n+1}. For example, one may have bounds in the form of

$$f(A(X^n), X_{n+1}) \leq \frac{1}{n} f(A(X^n), X^n) + C(n, \delta),$$

with probability $1 - \delta$ over the draw of the data X^{n+1}, where C denotes some complexity. Therefore, the right-hand side may be used as a goodness measure.

Next, we provide common examples of the principles for prediction. Those principles often come with some specific principles for modeling introduced above.

Empirical Risk Minimization Principle

The empirical risk minimization (ERM) principle (Vapnik, 1992) is one of the simplest principles for prediction. It dictates that one should pick the hypothesis $h \in \mathcal{H}$ that performs best on X^n,

$$A_{\mathrm{ERM}}(X^n) = \operatorname*{argmin}_{h \in \mathcal{H}} f(h, X^n).$$

In the statistical context, it is also known as the maximum likelihood principle (RA Fisher, 1922). If the generation mechanism of X^n and X_{n+1} is homogeneous (e.g., i.i.d.), then A_{ERM} is expected to perform well with high probability or in expectation.

As for the modeling, the ERM principle can be coupled with most modeling principles, e.g., validation, AIC, BIC, and PAC risk bounds.

Minimax Risk Principle

The minimax risk principle (Wald, 1950; Lehmann and Casella, 2006) is a theoretical principle and is often used for showing the optimality of the other principle such as the ERM principle. Under the minimax risk principle, the hypotheses $h \in \mathcal{H}$ are assumed to be probability distributions over $\mathcal{X}$ and the data is assumed to be subject to one of them. Then, the minimax risk predictor is given by

$$A_{\mathrm{risk}}(X^n) = \operatorname*{argmin}_{h \in \mathcal{H}_0} \max_{h_0 \in \mathcal{H}} \mathbb{E}_{h_0} \left[f(h, X_{n+1}) \right].$$

Bayes Principle

Under the Bayes principle (Gelman et al., 2013), each hypothesis $h \in \mathcal{H}_0$ is assumed to be a probability distribution over $\mathcal{X}^*$ and the loss is measured with the logarithmic loss, $f(h, X) = -\ln p(X|h)$ where p is the probability density function. In addition, we have a prior probability distribution π over $\mathcal{H}$ instead of just having a set of hypotheses, which expresses a prior belief on which hypothesis is possibly true. Then, we may adopt the hypothesis that minimizes the posterior risk,

$$A_{\mathrm{Bayes}}(X^n) = \operatorname*{argmin}_{h \in \mathcal{H}_0} \mathbb{E}_{\pi} \left[f(h, X_{n+1}) | X^n \right].$$

Typically, models are evaluated either with marginal likelihood (i.e., under the prequential principle) or with BIC.

Minimax Regret Principle

Under the minimax regret principle (Savage, 1951), we construct the predictor whose worst-case regret is the minimum. Here, the regret of algorithm A is defined as the excessive loss relative to the best hindsight prediction $h \in \mathcal{H}$, $\mathrm{REG}(A|X^n, \mathcal{H}) = f(A, X^n) - \min_{h\in\mathcal{H}} f(h, X^n)$. Thus, the minimax regret predictor is given by

$$A_{\mathrm{MMR}} = \operatorname*{argmin}_{A\in\mathcal{A}} \max_{X^n\in\mathcal{X}^n} \mathrm{REG}(A|X^n, \mathcal{H}),$$

where $\mathcal{A}$ denotes the set of all feasible prediction algorithms $\mathcal{A} = \{A : \mathcal{X}^* \to \mathcal{H}_0\}$. As no assumption on $\mathcal{S}$ is made, the quality of predictors is often measured with cumulative loss under the prequential principle.

These principles differ in their assumptions about the source $\mathcal{S}$. The ERM and minimax risk principle implicitly assume that X^n and X_{n+1} is similar in their distributions. The Bayes principle utilize the prior beliefs on the true hypothesis. In contrast, the minimax regret principle uses no side information and assumes nothing about the source $\mathcal{S}$ but is only justified if the worst-case regret is justified as a performance metric. Therefore, these principles should be used appropriately depending on the problem setting.

1.1.3 Advantages of Minimax Regret Principle

As mentioned later, we adopt the minimax regret principle in this work. There are at least three reasons supporting this choice. First, it does not involve any implicit assumptions on $\mathcal{S}$ other than the model $\mathcal{H}$. Therefore, the principle clarifies where the responsibility of prediction lies even under circumstances wherein we do not know the source $\mathcal{S}$ much, which is often the case in inductive inference. In other words, if the prediction is good, then the model must be good and vice versa. Second, as will also be mentioned later, adopting the minimax regret principle is justified from an information-theoretic perspective. Finally, if the models are high-dimensional, the minimax regret also characterizes the behavior of the other inference systems that follows the other principles such as the ERM principle and Bayes principle. As a result, the predictions made with the minimax regret principle often perform reasonably well in terms of the scores of the ERM and Bayes principles.

1.2 Minimum Description Length Principle

In this section, we first quickly introduce the minimum description length (MDL) principle, a meta-principle of inductive inference, which combines the prequential principle and the minimax regret principle within the information theory framework.

The minimum description length (MDL) principle suggests that one must choose the hypothesis that compress the data X^n the most. The birth of the MDL principle in the most primitive form dates back in the 13th century known as Occam's razor. Later, the notion of the shortest code length is formally given by Kolmogorov (1963) as the Kolmogorov complexity, and it is extended to the context of information theory by Rissanen (1978) with the current formalization of the MDL principle.

1.2.1 Code Length, Logarithmic Loss, and MDL Criterion

In view of the MDL principle, hypotheses $h \in \mathcal{H}_0$ are seen as lossless encoders and models $\mathcal{H}$ are sets of such encoders. In particular, encoders are mappings from n-sequences of alphabets $X^n \in \mathcal{X}^n$ to binary sequences $C(X^n|h) \in \{0,1\}^*$ where $C(\cdot|h) : \mathcal{X}^n \to \{0,1\}^*$ is a bijective mapping. We denote the length of codes by $L(X^n|h) = |C(X^n|h)| \ln 2$ in nats. The lossless assumption implies the Kraft–McMillan inequality,

$$\sum_{X^n \in \mathcal{X}^n} e^{-L(X^n|h)} \leq 1,$$

and hence there exists a sub-probability mass function $P(\cdot|h)$ such that $L(X^n|h) = -\ln P(X^n|h)$. Conversely, for any sub-probability distribution P over $\mathcal{X}^n$, there exists a lossless encoding method called Shannon coding C whose code length is equal to the negative logarithm of P within only one bit, $||C(X^n)| - -\ln P(X^n)| \leq \ln 2$. As such, it is customary to identify lossless encoders with sub-probability distributions and allow them to take non-integer values ignoring the difference of one bit. In other words, models are the set of sub-probability measures and the loss of prediction is measured with logarithmic loss.

To facilitate inductive inference based on the MDL principle, a code length function $L(X^n|\mathcal{H}) = -\ln \bar{P}(X^n|\mathcal{H})$ is considered that "represents" each model $\mathcal{H}$. For example, one may consider two-part code lengths of X^n with respect to $\mathcal{H}$, which are given by

$$L(X^n|\mathcal{H}) = \min_{h \in \mathcal{H}} L(X^n|h) + L(h|\mathcal{H}) = \min_{h \in \mathcal{H}} \ln \frac{1}{P(X^n|h)} + \ln \frac{1}{P(h|\mathcal{H})},$$

where the cardinality of the model $\mathcal{H}$ is assumed to be countable. Here, the second term is arbitrarily defined before seeing data X^n, e.g., uniform coding $L(h|\mathcal{H}) = -\ln |\mathcal{H}|$.

According to the philosophy of the MDL principle, a good model $\mathcal{H}$ describes regularities in data X^n well and hence can be utilized to compress it. To illustrate this, consider the following binary sequence

$$X^{100} = \underbrace{0101010101010101010101010101010101\ldots01}_{100\text{bits}}.$$

We can describe this by just writing "repeat '01' for 50 times", which has a length of 24 characters and is much shorter than directly writing down the whole binary sequence. In this specific example, the belief that "'0' is followed by '1' for every consecutive two bits" allows us to compress the length of the description from 100 characters to 24 characters. Further expanding this argument, we consider that a model $\mathcal{H}$ is a good model if and only if the description length $L(X^n|\mathcal{H})$ is small. In particular, $L(X^n|\mathcal{H})$ is referred to as the MDL criterion (of $\mathcal{H}$).

1.2.2 Universal Coding and Minimax Regret Principle

The MDL criterion is actually a metacriterion as there is a freedom of choice on the code length $L(X^n|\mathcal{H})$ as long as it "represents" the model $\mathcal{H}$. Formally, this is defined through the notion of universality.

As a natural measure of the relative performance of code length functions L with respect to base encoders $h \in \mathcal{H}$, we define the redundancy of L with respect to X^n and h as

$$\mathrm{RED}(L|X^n, h) \overset{\text{def}}{=} L(X^n) - \ln \frac{1}{P(X^n|h)}.$$

A code length L is *universal* (in the sense of individual sequence) with respect to $\mathcal{H}$ if, for all $h \in \mathcal{H}$ and $\epsilon > 0$, there exists n_0 such that

$$\sup_{X^n \in \mathcal{X}^n} \mathrm{RED}(L|X^n, h) < \epsilon,$$

where for all $n \geq n_0$. This notion has a practically important meaning such that if L is universal with respect to $\mathcal{H}$, then for any encoders $h \in \mathcal{H}$, the code length $L(X^n)$ is asymptotically no larger than $L(X^n|h)$ no matter what sequence X^n is given. In fact, the two-part code length is also universal and the universality is the only requirement on the MDL criterion $L(\cdot|\mathcal{H})$ (Grünwald, 2007).

Taking this one step further, the worst-case regret is given as a doubly worst-case approach such that

$$\mathrm{REG}^{\star}(L|\mathcal{H}) = \sup_{X^n \in \mathcal{X}^n} \sup_{h \in \mathcal{H}} \left[\mathrm{RED}(L|X^n, h)\right].$$

This provides a concept stronger than that of universality. In particular, if the worst-case regret of L grows sublinearly, then L is universal too. Now, the minimax-regret code length is defined as the minimizer of the worst-case regret,

$$L^{\star} = \operatorname*{argmin}_{L:\text{lossless}} \sup_{X^n \in \mathcal{X}^n} \mathrm{REG}(L|X^n, \mathcal{H}),$$

whose explicit form is given by Shtarkov (1987) as

$$L^{\star}(X^n|\mathcal{H}) = \inf_{h \in \mathcal{H}} L(X^n|h) + \ln Z(\mathcal{H}),$$

where $Z(\mathcal{H}) = \sum_{X^n \in \mathcal{X}^n} \sup_{h \in \mathcal{H}} P(X^n|h)$. Rissanen (1996) later demonstrated that the minimax-regret code length $L^{\star}(X^n)$ is realizable with a smart two-part coding scheme and proposed to use it for the code length function of the MDL criterion such that $L(X^n|\mathcal{H}) = L^{\star}(X^n)$. This special instance of the MDL criterion is called the normalized maximum likelihood (NML) code length $L^{\star} = L_{\mathrm{NML}}$ since it consists of the maximum logarithmic likelihood term (the first term) and the normalizing term (the second term). Sometimes we refer to NML as the *stochastic complexity* $L_{\mathrm{NML}} = SC$ for short. In the context of probabilistic modeling, one may be interested in the probabilistic counterpart $\bar{P}_{\mathrm{NML}}(X^n|\mathcal{H}) \overset{\mathrm{def}}{=} \exp\{-L_{\mathrm{NML}}(X^n|\mathcal{H})\}$ and refer to it as the NML distribution.

When the NML code length L_{NML} is used, the MDL principle is nothing more than the minimax-regret principle with the logarithmic loss $f(h, X^n) = L(X^n|h) = -\ln P(X^n|h)$. In fact, the corresponding minimax predictor $A_{\mathrm{NML}} : \mathcal{X}^* \to \mathcal{H}_0$ defines a probability distribution

$$P(X_{i+1}|A_{\mathrm{NML}}(X^i)) = \bar{P}_{\mathrm{NML}}(X_{i+1}|X^i, \mathcal{H}) = \frac{\sum_{X_{i+2}^n \in \mathcal{X}^{n-i-1}} \sup_{h \in \mathcal{H}} P(X^n|h)}{\sum_{X_{i+1}^n \in \mathcal{X}^{n-i}} \sup_{h' \in \mathcal{H}} P(X^n|h')}, \tag{1.1}$$

which coincides with the prequential form of the NML distribution. Moreover, the score of the predictor A_{NML} with respect to the prequential principle is equivalent to the NML code length, $\mathit{score}(A_{\mathrm{NML}}, X^n) = f(A_{\mathrm{NML}}, X^n) = L_{\mathrm{NML}}(X^n|\mathcal{H})$.

The advantage of considering the MDL principle instead of the general minimax-regret principle is two-fold. First, the explicit formula of the minimax optimal predictor is available. This allows us to analyze the optimal predictor in an accurate and non-asymptotic manner even if the model is so complex that the general minimax regret is difficult to analyze. Second, the logarithmic loss has an information-theoretic interpretation and thus has a wide range of applications on its own, such as classification and clustering. Moreover, general loss functions can be reframed into a form of logarithmic ones via entropification (Grünwald, 1999).

1.2.3 Tasks of Inductive Inference under the MDL principle

On the basis of the MDL principle, we focus on two common tasks of inductive inference—model selection and prediction. In both tasks, the NML code length L_{NML} plays a central role as the measure of model complexity.

Model selection

In model selection, we have candidates of models $\{\mathcal{H}_k\}_{k=1}^K$ associated with the minimax regret predictors A_k that describe the same source $\mathcal{S}$. To select the best models, given data X^n, we may choose the model $\mathcal{H}_k$ such that the corresponding A_k performs best on X^n. With the MDL principle, the optimal predictors are the NML distributions. Therefore, the problem is the evaluation of the NML code length for the given models $\mathcal{H}_k$. Specifically, most previous work in this literature has been focused on the computation or approximation of the normalizing term $Z(\mathcal{H})$ as it is analytically intractable in general.

Prediction

To make a prediction, given models $\mathcal{H}$, we derive computationally tractable algorithms A on the basis of $\mathcal{H}$. In terms of the MDL principle, the optimal predictor is also given by the NML distribution. However, owing to the computational difficulty of the normalizing term $Z(\mathcal{H})$, the corresponding optimal prediction algorithm is not tractable either. Furthermore, the numerical computation or approximation of $Z(\mathcal{H})$ does not suffice for this task. One has to obtain the prequential form $\bar{P}_{\mathrm{NML}}(X_i|X^{i-1}, \mathcal{H})$ for actually making predictions and it requires analytically intractable summation over the space of possible data as in (1.1).

1.3 Research Question: MDL principle in High Dimensions

High-dimensional models are the models parameterized with vectors $\theta \in \Omega \subset \mathbb{R}^d$, $\mathcal{H} = \{h_\theta \mid \theta \in \Omega\}$, whose dimensionality d is (much) larger than the sample size n. High-dimensional models have been extensively utilized for machine learning and data mining, partially because of recent successes in prediction tasks involving sparse modeling (Rish and Grabarnik, 2014), gradient boosting (Friedman, 2001), and deep learning (Goodfellow et al., 2016). Owing to the high-dimensionality, these models are flexible in the sense that they have large degrees of freedom. The prediction with such models is especially useful when one knows nothing much about the data-generating source $\mathcal{S}$ or one cannot rigorously formulate the knowledge for inference. Specifically, deep learning has been demonstrated to dramatically improve state-of-the-art methods in speech recognition, image recognition, object detection, and reinforcement learning, as well as drug discovery and genomics. Moreover, as the high dimensionality allows us to model the source with the sparseness condition, i.e., most coefficients of the parameter θ are zero, they make it easier to interpret the results of inference than low-dimensional dense models.

In contrast, the MDL principle in the high-dimensional context has not been extensively explored so far. Specifically, the existing studies on the NML code length largely rely on asymptotic analyses with large sample limit $n \to \infty$ where the dimensionality of the models held constant. Thus, we focus on the following two research questions in view of the MDL principle. **(Q1)** How can we evaluate the NML code length of high-dimensional models non-asymptotically? **(Q2)** Can we provide (approximately) analytic expressions of the prequential form of the NML distribution under the high-dimensional regime? As

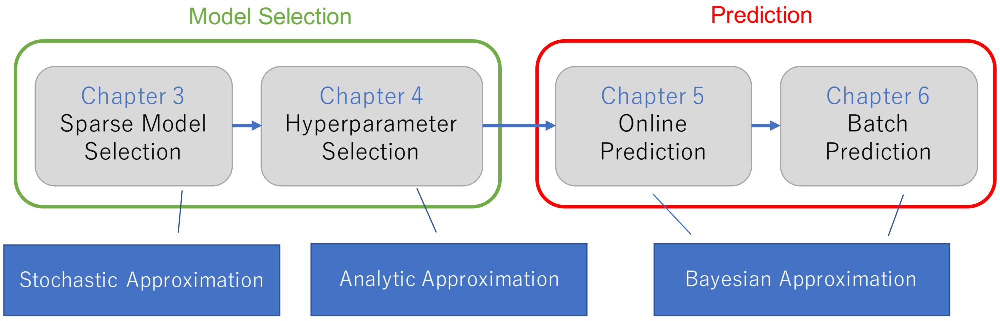


Fig. 1.2: Overview of the thesis

we will see in Chapter 2, we address these questions with techniques of relaxation.

1.4 Contributions and Outline

With regard to the above research questions, the contributions of the present study are summarized as follows:

(C3) We approach the problem of model selection by introducing continuous relaxation and stochastic approximation of the NML code length. It enables us to approximate the normalizing term without asymptotic assumption (Chapter 3).

(C4) We non-asymptotically study the NML code length of smooth models to derive a new analytic approximation. Although the smoothness of the model is assumed to apply this method, the new approximation guarantees that the approximation error is uniformly bounded. We also develop an algorithm to optimize the approximated NML with some convergence guarantees (Chapter 4).

(C5,6) We study a novel complexity measure, namely the *envelope complexity*, that provides a theoretical framework to analytically approximate the NML code length based on Bayesian approximation. We demonstrate its power by designing a tractable adaptive minimax predictor over high-dimensional ℓ_1 balls (Chapter 5). We also utilize envelope complexity to give systematic upper bounds on predictive risks. This demonstrates that envelope complexity actually is an essential complexity measure alternative to the minimax regret (Chapter 6).

Each of these offers a systematic manner of constructing minimax optimal predictors and/or evaluating minimax regrets in a high-dimensional setting. Hence, by putting them altogether, the present thesis provides theoretical foundations and tools for inductive inference based on the MDL principle with high-dimensional models. The visual summary of the overview is shown in Figure 1.2.

The rest of the thesis is organized as follows. We first provide a comprehensive review and more detailed introduction to the tools and results developed under the MDL principle in Chapter 2. Then, we present the results on high-dimensional model selection in the following consecutive two chapters (Chapter 3 and 4). In the next two chapters, we demonstrate the results on high-dimensional prediction derived from the framework of the envelope complexity (Chapter 5 and 6). Finally, we conclude the thesis and discuss future directions in Chapter 7.

Chapter 2

Preliminary

In this chapter, we introduce the existing results in the MDL literature that are relevant to our study and clarify our position among them.

To this end, we first describe the known properties and limitations of the normalized maximum likelihood (NML) code length, which is the key quantity of the MDL principle. Then, we discuss previous results and our results on the extensions and approximations of the NML code length.

2.1 Normalized Maximum Likelihood (NML) Code Length

Under the MDL principle, the NML code length is both the criterion of model selection and the optimal predictor. Thus, almost all inference tasks based on the MDL principle can be seen as the computation of NML or its approximation.

2.1.1 Definition

Let $(\mathcal{X}, \mu)$ be a measure space of data, where μ is the base measure like the Lebesgue measure or the counting measure, and denote data by $X^n = (X_1, \dots, X_n)$. Let $\mathcal{H} = \{p(X|\theta) \mid \theta \in \Omega \subset \mathbb{R}^d\}$ be a model or a set of (sub) probability densities over $\mathcal{X}$. We denotes the joint density of sequence X^n with respect to parameter θ as $p(X^n|\theta) = \prod_{i=1}^n p(X_i|\theta)$.

The (idealized) lossless code lengths L are measurable functions over $\mathcal{X}$ that satisfy the Kraft-McMillan inequality $\int \exp\{-L(X)\}\, \mu(\mathrm{d}X) \leq 1$. The NML code length, or NML for short, with respect to the model $\mathcal{H}$ is defined as

$$L_{\mathrm{NML}}(X^n) \stackrel{\mathrm{def}}{=} \inf_{\theta \in \Omega} \ln \frac{1}{p(X^n|\theta)} + \ln Z(\mathcal{H})$$

where $Z(\mathcal{H}) = \int \sup_{\theta \in \Omega} p(X^n|\theta) \mu(\mathrm{d}X^n)$ is the normalizing term to make it lossless code length. If the normalizing term is undefined, i.e., the integral is infinite, NML is not well-defined.

2.1.2 Properties

The motivation behind using the NML code length as the criterion of the MDL principle is based on the fact that NML, if defined, achieves Shtarkov's minimax regret. More precisely, it satisfies the following equality.

Theorem 1 (Shtarkov (1987)) If $Z(\lambda)$ is finite, then we have

$$\ln Z(\mathcal{H}) = \inf_{L:\text{lossless}} \sup_{X^n \in \mathcal{X}^n, \theta \in \Omega} \mathrm{REG}(L|X^n, \mathcal{H}),$$

where $\mathrm{REG}(L|X^n, \mathcal{H}) = L(X^n) - \inf_{\theta \in \Omega} \ln \frac{1}{p(X^n|\theta)}$ denotes the regret of code length L with respect to data X^n and model $\mathcal{H}$. Moreover, the minimax regret is uniquely attained with NML, i.e., the other code lengths do not achieve the minimax regret.

The proof is relegated to that of a general theorem in Section 2.2.1. It is immediately seen that $\mathrm{REG}^\star(\mathcal{H}) \overset{\text{def}}{=} \inf_{L:\text{lossless}} \sup_{X^n} \mathrm{REG}(L|X^n, \mathcal{H}) = Z(\mathcal{H})$, and hence we refer to $Z(\lambda)$ as the *Shtarkov complexity*. Note that this is minimax in the sense of an individual sequence. This must be distinguished from one in the expectation sense, which assumes that X^n is subject to some distribution, unlike ours.

Moreover, when we see the NML code length as a probability density $\bar{p}(X^n) = e^{-L_{\mathrm{NML}}(X^n)}$, the associated sequential prediction strategy $\bar{p}(X_i|X^{i-1})$ achieves minimax regret with respect to the logarithmic loss $-\ln p(X_i|X^{i-1})$. This is easily seen from the telescoping sum

$$L_{\mathrm{NML}}(X^n) = \sum_{i=1}^{n} \ln \frac{1}{\bar{p}(X_i|X^{i-1})},$$

where $\bar{p}(X_i|X^{i-1}) = \bar{p}(X^i)/\bar{p}(X^{i-1})$ denotes the conditional distribution.

2.1.3 Direct Approximation and Exact Computation of NML

Since the integration in the normalizing term is intractable in general, a number of formulae for calculating NML have been proposed so far.

Rissanen's Asymptotic Formula One of the most famous approximations is Rissanen's asymptotic expansion (Rissanen, 1996). It states that for models $\mathcal{H}$ satisfying certain regularity conditions that include some exponential families of distributions, we have

$$\ln Z(\mathcal{H}) = \frac{d}{2} \ln \frac{n}{2\pi} + \ln \int_{\Omega} \sqrt{\det I(\theta)} \mathrm{d}\theta + o(1),$$

where $o(1) \to 0$ as $n \to \infty$. Here, $I(\theta)$ denotes the Fisher information matrix $I(\theta) = -\mathbb{E}_X \nabla_\theta^2 \ln p(X|\theta)$. Note that this is equal to BIC (Schwarz et al., 1978) except with a constant, and Rissanen's expansion can be thought of as a more accurate version of BIC for the NML distribution. Later, arbitrary precise expansions for multinomial distributions are obtained through the singularity analysis (Flajolet and Odlyzko, 1990) by Kontkanen (2009).

Exact Computation for Continuous Exponential Families For some instances of continuous exponential families of distributions including normal, Gamma, exponential, and logistic distributions, non-asymptotic formulae are also developed (Hirai and Yamanishi, 2013). These are derived by transforming the integral with respect to the data space $\mathcal{X}^n$ into the parameter space Ω.

Efficient Numerical Computation The multinomial distribution is one of the most common distributions over discrete data. Accordingly, the computation of NML for the multinomial distributions has been extensively studied. The NML of multinomial models with m outcomes and n observations is reduced by recursion to that of binomial models in linear time, and the total time complexity is $O(m + n)$ (Kontkanen and Myllymäki, 2007).

Moreover, the NML of binomial models with finite precision is computed ever faster with $O(\sqrt{n})$ time (Mononen and Myllymäki, 2008).

2.1.4 Limitations

Notwithstanding the above calculation techniques, there remain some limitations. We discuss two of the largest limitations in the following.

Limitation 1) First, the NML code length, even with asymptotic approximation, is often not tractable in practical situations. In particular, most of the results on the exact computation focus on exponential families such as normal and multinomial distributions, and the asymptotic formula is only applicable to regular models, which excludes most practical models. Moreover, specifically with continuous data spaces, the normalizing term tends not to be well-defined and requires appropriate restriction of the parameter space.

Limitation 2) Secondly, although NML is the unique optimal predictor, actual predictions based on NML are almost impossible. This is because the conditional density $\bar{p}(X_i|X^{i-1})$ requires marginal density $\bar{p}(X^i)$, and the marginal density requires the integral $\int \bar{p}(X^n)\mu(\mathrm{d}X_{i+1}^n)$, which is even more intractable than the NML code length itself.

In the following, we discuss the existing approaches and our approach to these limitations.

2.2 Extension of NML Code Length

several extensions of NML have been proposed to bypass Limitation 1, i.e., to deal with practical models where the ordinary NML cannot be computed.

2.2.1 Luckiness NML

The luckiness NML (Kakade et al., 2006; Grünwald, 2007) (LNML) is an extension of NML whose complexity term takes a function over hypotheses instead of a set of hypotheses. It can also be seen as a generalization of the conditional NML (Rissanen and Roos, 2007). Let $\gamma : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a penalty function of hypotheses. The LNML with respect to γ is given by

$$L_{\mathrm{LNML}}(X^n) = \inf_{\theta \in \mathbb{R}^d} \left[\ln \frac{1}{p(X^n|\theta)} + \gamma(\theta) \right] + \ln Z(\gamma),$$

where $Z(\gamma) = \int \sup_{\theta \in \mathbb{R}^d} p(X^n|\theta) e^{-\gamma(\theta)} \mu(\mathrm{d}X^n)$ denotes the corresponding normalizer. LNML can be seen as a generalization of NML in terms of the restriction of parameter spaces from hard constraints to soft constraints. Note that LNML recovers NML when $\gamma(\theta) = 0$ for $\theta \in \Omega$, and $\gamma(\theta) = \infty$ otherwise.

LNML can be used as a remedy for the infinite complexity problem. By choosing appropriate penalty functions γ, one can bound $Z(\gamma)$ without actually hard-constraining the parameter space. Moreover, if one takes γ nicely with respect to the density $p(X^n|\theta)$, LNML is easily computed even if NML is not (e.g., see Miyaguchi (2017)). Furthermore, LNML may be utilized for deriving an upper bound on NML. If the parameter space is restricted as $\gamma(\theta) \leq B$, then we have

$$\begin{aligned}\ln Z(\mathcal{H}) &= \ln \int \sup_{\gamma(\theta)\leq B} p(X^n|\theta)\mu(\mathrm{d}X^n)\\ &\leq \ln \int \sup_{\theta\in\mathbb{R}^d} p(X^n|\theta)e^{-\gamma(\theta)+B}\mu(\mathrm{d}X^n)\\ &= B + \ln Z(\gamma).\end{aligned}$$

LNML satisfies generalized minimax regret optimality. To show this, let $\mathrm{REG}(L|X^n,\gamma) = L(X^n) - \inf_{\theta\in\mathbb{R}^d}\left[\ln\frac{1}{p(X^n|\theta)} + \gamma(\theta)\right]$ denote the generalized regret.

Theorem 2 (Minimax optimality of LNML) For all $\gamma : \mathbb{R}^d \to \overline{\mathbb{R}}_+$, we have

$$\ln Z(\gamma) = \inf_{L:\text{lossless}} \sup_{X^n\in\mathcal{X}^n} \mathrm{REG}(L|X^n,\gamma).$$

Moreover, if $Z(\gamma) < +\infty$, the minimax regret is uniquely obtained with LNML.

Proof Let L_{LNML} be the LNML code length. Suppose any code length L^0 such that there exists $E \subset \mathcal{X}^n$ and $L_{\mathrm{LNML}} \neq L^0$ for all $X \in E$ with $\mu(E) > 0$. Then, noting that $\int e^{-L^0(X^n)}\mu(\mathrm{d}X^n) \leq 1 = \int e^{-L_{\mathrm{LNML}}(X^n)}\mu(\mathrm{d}X^n)$, there exists $X^n \in \mathcal{X}^n$ such that $L^0(X^n) < L_{\mathrm{LNML}}(X^n)$, otherwise we have a contradiction. By exploiting X^n, we have

$$\begin{aligned}\sup_{Y^n\in\mathcal{X}^n} REG(L^0, Y^n, \gamma) &\geq REG(L^0, X^n, \gamma)\\ &= L^0(X^n) - \inf_{\theta\in\mathbb{R}^d}\left[\ln\frac{1}{p(X^n|\theta)} + \gamma(\theta)\right]\\ &> L_{\mathrm{LNML}}(X^n) - \inf_{\theta\in\mathbb{R}^d}\left[\ln\frac{1}{p(X^n|\theta)} + \gamma(\theta)\right]\\ &= \ln Z(\gamma)\\ &= \sup_{Y^n\in\mathcal{X}^n} REG_n(L_{\mathrm{LNML}}, Y^n, \gamma),\end{aligned}$$

and therefore L^0 does not achieve minimax regret, but L_{LNML} does. In addition, the minimax regret turns out to be $\ln Z(\gamma)$. ■

Rissanen's asymptotic expansion is also generalized for LNML (Grünwald, 2007). The generalized version is given by

$$\ln Z(\gamma) = \frac{d}{2}\ln\frac{n}{2\pi} + \ln\int_\Omega \sqrt{\det I(\theta)}e^{-\gamma(\theta)}\mathrm{d}\theta + o(1),$$

under the asymptotics of $n \to \infty$. Here, the similar regularity conditions as in the original expansion are required.

2.2.2 Latent Variable Completion

Latent variable models (LVMs) are useful generative models that have unobserved random variables $W_i \in \mathcal{W}$ whose densities are given by

$$p(X|\theta) = \int_{\mathcal{W}} p(X|\theta, W)\pi(\mathrm{d}W|\theta).$$

This class of densities includes practically important models such as mixture models, the hidden Markov models, and the restricted Boltzmann machines. In general, the NMLs of LVMs are intractable since the maximum likelihood estimation (MLE) inside the integral $Z(\mathcal{H})$ is analytically intractable.

To address this issue, the technique of latent variable completion (LVC-NML) has been developed. We estimate the latent variables $\hat{W}^n = \hat{W}^n(X^n)$ from the observables X^n and treat them as if they were the true values W^n, i.e.,

$$L_{\mathrm{LVC}}(X^n) = L_{\mathrm{NML}}(X^n, \hat{W}^n(X^n)).$$

Thus, the MLEs are often analytically solved and the integrals $Z(\mathcal{H})$ are computed by sophisticated recursions. The LVC-NMLs of the naïve Bayes models are computed utilizing the moment generating functions within $O(n^2)$ time (Mononen and Myllymäki, 2007). The LVC-NMLs of mixtures of exponential families are also reduced to that of the base exponential families based on the techniques of moment generating functions (Hirai and Yamanishi, 2013, 2017). Its computation time is $O(n^2K)$, where K is the number of mixture components. Wu et al. (2017) further extended LVC-NML to derive the decomposed NML (DNML) for more complex models such as latent Dirichlet allocation models, where the ordinary LVC-NML is intractable.

2.2.3 Our Approach

Applicability of the above extensions still appears to be remaining in a confined class of models. In other words, either with LNML or LVC-NML, one should derive their new formulae when they have encountered new complex models. Although the asymptotic expansion of LNML is exceptional from this issue, it is still not applicable to our high-dimensional setting since the expansion is only valid for the conventional large-sample limits.

In this work, to further expand the applicability to larger classes of models including high-dimensional models, we take two different approaches. For one approach, we run the stochastic gradient method on the surface of LNML (Chapter 3). By computing gradients directly and bypassing the value of LNML, the proposed method has much wider applicability to high-dimensional and complex models. On the other hand, we propose two non-asymptotic analytic approximations of LNML (Chapter 4 and 5). By abandoning the exact value, we obtain simple analytic formulae that can be systematically computed given models.

2.3 Approximation of NML as a Predictor

In this section, we review the previous work on the approximation of NML in terms of tractable prediction to address Limitation 2.

2.3.1 Sequential NML

The sequential NML (SNML) allows us to make computationally inexpensive prediction by taking the current observation as the last one for every time step (Takimoto and Warmuth, 2000; Rissanen and Roos, 2007; Roos, 2008). The equivalence of SNML and NML has been studied by Hedayati and Bartlett (2012) and they gave key insights on when NML can be simulated with inexpensive classes of predictors such as SNML and Bayes predictors.

2.3.2 Bayes Predictors

The Krichevsky–Trofimov (KT) estimator (Krichevsky and Trofimov, 1981) is one of the most traditional universal code lengths that can be seen as a Bayesian approximation of NML for multinomial models. It achieves asymptotic minimax regret for almost all ranges of data. After it was proposed, Xie and Barron (2000); Watanabe and Roos (2015) improved upon it to achieve strict asymptotic minimaxity.

Takeuchi and Barron (2013) showed that the Bayesian predictor with the Jeffreys prior asymptotically achieves minimax regret for exponential families. Barron et al. (2014) also studied numerical Bayesian simulation of NML and showed that signed discrete priors allow us to exact simulation for certain discrete variable models.

2.3.3 Our Approach

We take the Bayesian approach to approximate the NML predictors in Chapter 5. The largest difference between our method and conventional ones is that we assume the high-dimensionality of models, whereas others assume large sample size compared to the dimensionality. Moreover, in Chapter 6, we show that our analysis can naturally be extended to the batch-prediction scenario, thereby justifying the proposed approximation of NML.

Chapter 3

Graphical Model Selection via Relaxed Stochastic Complexity

Discovering a true sparse model capable of generating data is a challenging yet important problem for understanding the nature of the source of data. A major part of the challenge arises from the fact that the number of possible sparse models grows exponentially as the dimensionality of the models increases. In this study, we consider a method for estimating the true model over an exponentially large number of sparse models based on the minimum description length principle[*1]. We show that a novel criterion derived by *continuous relaxation of the stochastic complexity* induces selection of the true model by solving the ℓ_1-regularization problem for which the hyperparameters are appropriately chosen. Moreover, we provide an efficient optimization algorithm for finding the appropriate hyperparameters and select the sparse model accordingly. The experimental results we obtained for the problem of sparse graphical modeling indicate that the proposed method estimates the true model effectively in comparison to existing methods for choosing hyperparameters to solve the ℓ_1-regularization problem.

3.1 Motivation

In sparse modeling, combinatorial sets of explanatory factors are considered in order to interpret observations $X = (x_1, x_2, \cdots, x_n)^\top$. Each combinatorial set J is a subset of the universal set of explanatory factors $J \subset [d] \overset{\text{def}}{=} \{1, 2, \cdots, d\}$, where d denotes the number of available factors. Sparse modeling allows us to choose $\hat{J}$ such that the observations are well explained and $\hat{J}$ is sparse, i.e., $|\hat{J}|$ is small at the same time. Estimation of an appropriate sparsity of J is not only beneficial in terms of space and computational time, but it is also useful for constructing good predictors of the source of the data X or for discovering essential explanatory factors. Sparse modeling has had recent advances and success in applications in areas such as medicine, chemistry, and materials science (Rish and Grabarnik, 2014).

When we consider a model of probability density (or mass) $\mathcal{M} = \left\{ p(X; \theta) \;\middle|\; \theta \in \Omega \subset \mathbb{R}^d \right\}$, each dimension of the parameter θ can be seen as an explanatory factor. In this context, sets of explanatory factors J are defined as the *sparsity patterns* of parameters, i.e., $j \in J$ if and only if $\theta_j \neq 0$. In order to understand the nature of the source of the data X, we must determine the sparsity pattern of the generative parameter, which we denote by J^*. We are often motivated to solve the ℓ_1-regularization problem given by the following

[*1] The content of this chapter was published in Miyaguchi et al. (2017)

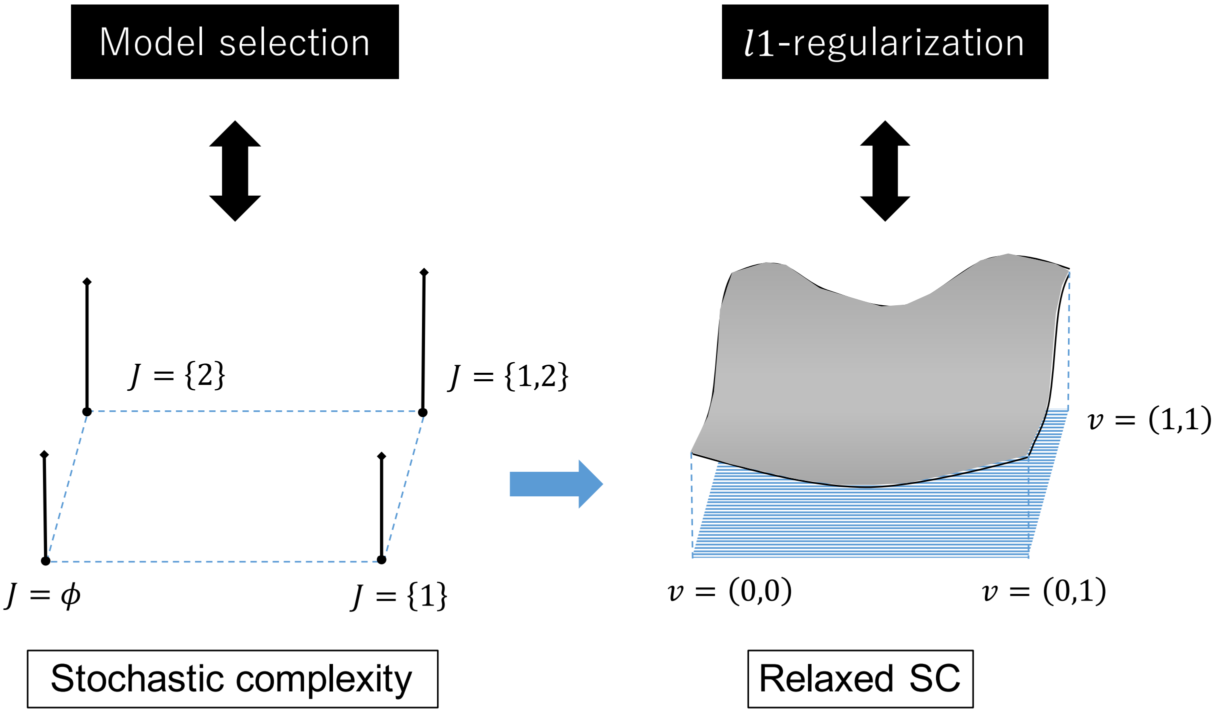


Fig. 3.1: Schematic of the proposed method for the dimensionality $d = 2$. In order to solve the minimization of the stochastic complexity over the power set of explanatory factors 2^J (bottom left), we relax it to the optimization of the *relaxed stochastic complexity* (bottom right), which can be efficiently minimized. This relaxation also reveals the underlying relationship between model selection (top left) and hyperparameter selection of ℓ_1-regularization (top right).

formulation:

$$\bar{\theta}(X, \lambda) = \operatorname*{argmin}_{\theta \in \Omega} \left\{ \ln \frac{1}{p(X;\theta)} + \sum_j \lambda_j |\theta_j| \right\} \tag{3.1}$$

given the hyperparameter $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_d) \geq \mathbf{0}$. Note that although λ is often assumed to be univariate, i.e., $\lambda = c\mathbf{1}$, it has been shown that such a simplification may give an inconsistent estimate in some cases (Zou, 2006). This is why we are concerned with the general formulation in (3.1). The sparsity patterns of $\bar{\theta}(X, \lambda)$ depend on the value of λ. In the case of $\lambda = \mathbf{0}$, we obtain the maximum likelihood (ML) estimate, which is typically dense. In contrast, $\bar{\theta}(X, \lambda)$ becomes sparse as λ increases. In general, there exists an optimal hyperparameter λ^* such that $\bar{\theta}(X, \lambda^*)$ can recover the true sparsity pattern J^*; that is, $\bar{\theta}_j(X, \lambda^*) \neq 0$ if and only if $j \in J^*$. However, λ^* is not known in real applications, and several methods for hyperparameter selection, such as cross validation and Bayesian methods, have been proposed to date.

On the other hand, as each sparsity pattern $J \subset [d]$ corresponds to a probabilistic model $\mathcal{M}_J = \{p(X;\theta) \mid \theta_j = 0, \forall j \notin J\}$, we can regard sparse modeling as a problem of choosing a probabilistic model, namely, model selection. In this study, we approach the problem of estimating the true pattern of J^* by means of model selection based on the *MDL principle* (Rissanen, 1978). According to this principle, we choose the pattern $\hat{J}$

that minimizes the NML code length, also known as *stochastic complexity*:

$$SC(X;J) = -\ln \frac{p(X;\hat{\theta}(X,J))}{\int p(Y;\hat{\theta}(Y,J))dY}, \tag{3.2}$$

where $\hat{\theta}(\cdot, J)$ denotes the ML estimator with respect to the probabilistic model $\mathcal{M}_J$. Traditionally, model selection is formulated as the minimization of criteria over finite candidates and is conducted by computing their values for all candidates. In the case of sparse modeling, the number of candidates grows exponentially as the dimensionality d increases. Therefore, it is impossible to compute all of the abovementioned criteria in a real application.

In this chapter, we consider model selection based on the MDL principle and propose an algorithm for finding the optimal sparsity pattern from the viewpoint of the MDL principle. By applying continuous relaxation for the discrete minimization of the stochastic complexity, we obtain the *relaxed stochastic complexity (RSC)* (see Figure 3.1) and derive an algorithm to minimize the relaxed criterion, in which the problem of ℓ_1-regularization in (3.1) is iteratively solved. Not only does this allow us to efficiently find the optimal model from an exponentially large number of candidates, but it also gives a criterion for choosing the hyperparameter λ in (3.1).

The remainder of the chapter is organized as follows. In Section 2, we present the conventional methods and clarify their relationship to this study. Section 3 states the notion of the MDL principle and the derivation of novel MDL-based criteria suited to sparse modeling. Section 4 introduces the problem of sparse graphical modeling as a typical issue of sparse modeling. A state-of-the-art solution to this problem, the Graphical LASSO (Friedman et al., 2008), is also discussed. In Section 5, we present an algorithm for optimizing the proposed criteria specialized for sparse graphical mode. Section 6 contains the experimental results obtained for the proposed algorithm in identifying the true sparsity pattern J^*. Section 7 provides concluding remarks and suggests future research.

3.2 Related Work

As we study the problem of sparse modeling by connecting the *model selection* with the *hyperparameter selection of the ℓ_1-regularization*, we review work related to these two domains. In general terms, the purpose of model selection and ℓ_1-regularization in sparse modeling is either generalization or interpretation. For generalization purposes, we are concerned with estimating accurate values of a parameter by selecting an appropriate model or hyperparameter in the sense that the risk of the estimates relative to some classes of the distribution of data is bounded or minimized. On the other hand, our focus is on models for interpretation purposes. We are interested in finding the sparsest model that explains the data sufficiently well, i.e., the essential model, by either directly choosing a model or by choosing a hyperparameter and examining the resulting estimate with the aim of exploiting it for knowledge discovery of the source of the data. Specifically, we consider the *consistency* of the estimated models relative to the true probability distribution.

A number of criteria for model selection have been proposed thus far, namely the Akaike information criteria (AIC) (Akaike, 1974), the Bayes information criteria (BIC) (Schwarz et al., 1978), the MDL principle (Rissanen, 1978), etc. It has been shown that the estimates provided by the BIC and MDL are consistent, whereas those of the AIC are not. Another fascinating property of the MDL is that it is well-defined even for finite samples, which is typically not the case for the other criteria. In view of continuous relaxation, definiteness is important since the resulting relaxed criteria can also be interpreted as the description length of the presented samples, and the minimization over relaxed criteria continues to make sense.

A similar observation holds for Bayesian variable selection (George and McCulloch, 1997) in that Bayesian evidence can be regarded as the description length. One of the notable differences between the stochastic complexity and Bayesian evidence is that the stochastic complexity depends only on probabilistic models, whereas the evidence depends on the parameter priors, which have infinite degrees of freedom in principle. The proposed method utilizes this simple dependence of the stochastic complexity, which is useful for applications that solely depend on models. Moreover, it differs from our intention in the sense that it manages to directly optimize a criterion (i.e., evidence) over the exponential number of discrete candidates.

Methods for minimizing the criteria of model selection, such as the BIC and MDL, present the problem of computational complexity in view of sparse modeling. The minimization is often formulated as nonconvex, essentially discrete optimization, in which there is an exponentially large number of candidates to compare. Rissanen (2000) and Roos et al. (2009) studied the use of stochastic complexity in the problem of sparse feature selection. In these studies, it was assumed that features are orthogonal to one another. This assumption is necessary, as it reduces the essential number of candidate sets of features from exponential to linear. In contrast, we remove the orthogonality assumption and consider optimizing the nonlinear interaction among features, i.e., the explanatory factors. On the other hand, within the framework of PAC-Bayesian theory, a similar problem arising from model averaging for generalization purposes can be efficiently solved with sampling methods (Alquier et al., 2011). In fact, our method also employs a paradigm of sampling when solving the continuous-relaxationized problem, and it aggregates the information of the discrete candidates in the same manner.

The problem of ℓ_1-regularization was originally introduced by Tibshirani (1996) in order to find sparse estimates by means of convex optimization, which can be efficiently solved. As for the methods of hyperparameter selection for the ℓ_1-regularization problem, techniques involving cross validation (Kohavi, 1995) and the Bayesian LASSO (Park and Casella, 2008) have been widely used. On the other hand, theoretical risk bounds of LASSO estimates have recently been investigated from the viewpoint of the MDL principle (Barron and Luo, 2008; Chatterjee and Barron, 2014; Kawakita and Takeuchi, 2016), and several methods for selecting the optimal hyperparameters that minimize the bounds have been presented in various scenarios of statistical learning. Those methods are designed to choose the appropriate λ in (3.1) to estimate accurate parameter values as $\theta = \bar{\theta}(X, \lambda)$, which rather than being typically suitable for the purposes of interpretation such as identifying the true sparsity pattern, are appropriate for the purpose of generalization. However, our method resembles the Bayesian LASSO in the technical sense that concerns optimization over a continuous space of probabilistic models.

The adaptive LASSO (Zou, 2006) was proposed considering the estimation of the true sparse model. It specifies a half line of the hyperparameter l_X, which satisfies the *oracle property* according to observed data. The oracle property ensures approximate consistency and that there exists $\lambda^+ \in l_X$ such that the sparsity pattern of the estimate $\bar{\theta}(X, \lambda^+)$ is J^* with a high probability approaching one as $n \to \infty$. However, although the theoretical rates of the hyperparameter $\lambda = \lambda_n$ are presented in the analysis of consistency, they do not provide any information on how to choose λ relative to the given data X. Moreover, conventional methods for choosing λ fail to identify the true sparse model, even if the half line with the oracle property is given.

3.3 Relaxed Dual of the MDL Principle

There are a number of studies that have applied the MDL principle to problems of model selection, e.g., the orthogonal basis selection of regression models (Rissanen, 2000; Roos et al., 2009), the cluster-number selection of Gaussian mixture models (Hirai and Yamanishi, 2013), and rank selection for nonnegative matrix factorization (Ito et al., 2016). In these works, the description length of the data X relative to the probabilistic models $\mathcal{M}_J$ is minimized with respect to the index J. The description length is typically given by the stochastic complexity in (3.2), which is the length of the optimal lossless coding of the data X in the sense of Shtarkov's minimax regret (Shtarkov, 1987). In the asymptotic setting, under standard regularity conditions for the models $\{\mathcal{M}_J\}$, the minimizer of the stochastic complexity

$$\hat{J}(X) \overset{\text{def}}{=} \operatorname*{argmin}_{J \subset [d]} SC(X; J) \tag{3.3}$$

is known to be a *consistent estimator* of the true sparsity pattern J^* (Rissanen, 2012), i.e.,

$$\lim_{n \to \infty} P\left\{\hat{J}(X) = J^*\right\} = 1.$$

That is, in the context of sparse modeling, we can theoretically estimate the true sparsity pattern given a sufficient number of observations. However, naïve minimization of the stochastic complexity in (3.3) is practically intractable since the number of candidates is 2^d, which exponentially grows as d increases. Moreover, computation of the stochastic complexity itself is often analytically intractable owing to the normalizing factor $\int p(Y; \hat{\theta}(Y, J))dY$ in (3.2).

The key idea to overcome these difficulties is to conduct *continuous relaxation*. We note that the 2^d candidates can be naturally embedded in the Euclidean space $\mathbb{R}^d$ as the vertices of the d-dimensional unit hypercube $\mathcal{C}^d$; that is, there is a natural mapping φ such that $2^{[d]} \ni J \mapsto \varphi(J) = \sum_{j \in [d]} \chi_J(j) e_j \in \mathcal{C}^d$. Here, we define $\chi_J(\cdot)$ as an indicator function of the set J. Then, we introduce the RSC, which a continuous interpolation of the stochastic complexity over the hypercube $RSC(X; v)_{v \in \mathcal{C}^d}$, such that $RSC(X; \varphi(J)) = SC(X; J)$ for all $J \subset [d]$. Once we obtain the minimizer of the RSC by utilizing continuous optimization techniques, an approximate solution of $\hat{J}(X)$ is obtained by a certain backward mapping ψ_X from the hypercube $\mathcal{C}^d$ back to the power set $2^{[d]}$. Formally, we can define a new estimator of the sparsity pattern as follows.

Definition 1 The relaxed MDL estimator of the sparsity pattern is given by

$$\bar{J}(X) \overset{\text{def}}{=} \psi_X \left(\operatorname*{argmin}_{v \in \mathcal{C}^d} RSC(X; v) \right). \tag{3.4}$$

Note that this no longer suffers from the exponentially large number of candidates.

The relaxation of the stochastic complexity starts with the minimax optimality of the stochastic complexity:

$$SC(\cdot; J) = \operatorname*{argmin}_{L:\text{lossless}} \max_{X, \theta \in \Omega_J} \left\{ L(X) - \ln \frac{1}{p(X; \theta)} \right\} \tag{3.5}$$

given a subspace $\Omega_J = \{\theta \in \Omega \mid \theta_j = 0, \forall j \notin J\}$. Let $U_J(\theta)$ be a function such that $U_J(\theta) = 0$ for all $\theta \in \Omega_J$, and $U_J(\theta) = +\infty$ otherwise. Then, we can rewrite (3.5) in a functional form:

$$SC(\cdot; J) = \operatorname*{argmin}_{L:\text{lossless}} \max_{X,\theta\in\Omega} \left\{ L(X) - \ln \frac{1}{p(X;\theta)} - U_J(\theta) \right\}. \tag{3.6}$$

The dependence of the stochastic complexity on the sparsity pattern J is reflected only in the function $U_J(\theta)$. By relaxing this function into $\bar{U}_v(\theta) \overset{\text{def}}{=} -\sum_j |\theta_j| \ln v_j$, we have continuous relaxation of the stochastic complexity (RSC):

$$RSC(\cdot; v) \overset{\text{def}}{=} \operatorname*{argmin}_{L:\text{lossless}} \max_{X,\theta\in\Omega} \left\{ L(X) - \ln \frac{1}{p(X;\theta)} - \bar{U}_v(\theta) \right\}. \tag{3.7}$$

By the definition above, we can see that $RSC(X; \varphi(J)) = SC(X; J)$ for all $J \subset [d]$ since $\bar{U}_{\varphi(J)}(\theta) \equiv U_J(\theta)$ with the conventional definition of $\ln 0 = -\infty$. The RSC can be also written in the explicit form

$$RSC(X; v) = -\ln \frac{\max_{\theta\in\Omega} p(X;\theta) \prod_j v_j^{|\theta_j|}}{\int \max_{\theta'\in\Omega} p(Y;\theta') \prod_j v_j^{|\theta'_j|} dY}, \tag{3.8}$$

which is also known as a special case of the LNML code length (Grünwald, 2007). Here, the integral is taken over all possible values of the data Y. Observing the right-hand side of (3.8), it is straightforward to see that the RSC is continuous with respect to $v \in \mathcal{C}^d$ under standard conditions on $\mathcal{M}$.

In order to construct the backward mapping $\psi_X : \mathcal{C}^d \to 2^{[d]}$, we notice the ML estimator $\hat{\theta}(X, J)$ as an indicator of the sparsity pattern of the model J. Since the ML estimate can be characterized as the maximizer of the regret on the right-hand side of (3.6), we consider the maximizer of the relaxed regret on the right-hand side of (3.7) as the ancillary relaxation of the ML estimate. In fact, the relaxed ML estimate is equivalent to the solution $\bar{\theta}(X, \lambda)$ of (3.1) under a transformation of variables; that is, we can see that

$$\bar{\theta}(X, \lambda(v)) = \operatorname*{argmax}_{\theta\in\Omega} \left\{ -\ln \frac{1}{p(X;\theta)} - \bar{U}_v(\theta) \right\}, \tag{3.9}$$

where $\lambda_j(v) = -\ln v_j$. Therefore, a natural backward operator $\psi_X(\cdot)$ can be defined as the sparsity pattern of $\bar{\theta}(X, \lambda(\cdot))$ as

$$\psi_X(v) = \left\{ j \in [d] \;\middle|\; \bar{\theta}_j(X, \lambda(v)) \neq 0 \right\}. \tag{3.10}$$

Hereafter, we refer to $\lambda(v)$ as simply λ if the dependency is clear from the context.

Up to this point, we have given a feasible approximation of consistent model selection in (3.3) by the technique of continuous relaxation in (3.4) in view of the MDL principle. The relaxed objective function in (3.8) can be efficiently minimized by the family of gradient descent methods, as will be described in Section 5, and the pulling-back operation of the minimizer $\bar{v}$ back to the original domain is essentially equivalent to solving the problem of ℓ_1-regularization in (3.1).

3.4 Graphical LASSO

Undirected graphical models are used to express statistical dependence such as the correlation or noncorrelation among the variables of interest; that is, there is an edge between node i and node j if and only if the i-th and j-th variables are dependent, where the other variables are conditioned. Assuming that we have n independent observations of the m variables denoted as $X \in \mathbb{R}^{n\times m}$, we are motivated to infer the underlying graphical model in order to understand the relationship between the variables of high-dimensional observations in real life, e.g., see Menéndez et al. (2010).

It is known that the adjacency matrix of a graphical model is equivalent to the sparsity pattern of the precision matrix $\Theta = \Sigma^{-1} \in \mathbb{R}^{m\times m}$, given that the variables are drawn from an m-dimensional Gaussian distribution with zero mean $\mathcal{N}_m[\mathbf{0}, \Sigma]$. Using the notation of the empirical covariance $S = X^\top X/n$, the sparsity pattern of Θ can be inferred by solving the following graphical LASSO problem (Friedman et al., 2008):

$$\bar{\Theta}(S, \Lambda) = \operatorname*{argmin}_{\Theta\succ 0} \left\{ \operatorname{tr}[S\Theta] - \ln\det\Theta + \sum_{i,j} \Lambda_{ij} |\Theta_{ij}| \right\} \tag{3.11}$$

for a given symmetric matrix $\Lambda \in [0,1]^{m\times m}$. This minimization is well-defined as long as S is positive definite. Here, $\Lambda = \mathbf{1}$ gives the sparsest solution, i.e., no edges in the graph; therefore, we restrict the range of the hyperparameters to $0 \le \Lambda_{ij} \le 1$. The data X are assumed to be normalized, i.e., $\operatorname{diag} S = \mathbf{1}$, for the sake of scale invariance.

Note that this minimization is a special case of the formulation in (3.1). Hence, according to the discussion in Section 3.3, the estimator in (3.11) plays an important role in mapping the relaxed continuous domain to the original discrete domain, i.e., (3.10). This is important with regard to the backward operation and also a key quantity in the minimization of the RSC. For the sake of later use in the minimization, we introduce notation for the loss of graphical LASSO, denoted by

$$h(S, \Lambda) \overset{\text{def}}{=} \min_{\Theta\succ 0} \left\{ \operatorname{tr}[S\Theta] - \ln\det\Theta + \sum_{i,j} \Lambda_{ij} |\Theta_{ij}| \right\}. \tag{3.12}$$

The graphical LASSO is a convex optimization problem, and a number of efficient algorithms are known for it with a complexity of at most $O(m^3)$ per iteration. We employ one of them (Duchi et al., 2012) in the following section, which is suitable for the scenario in which each coefficient Λ_{ij} is an independent free hyperparameter.

3.5 Persistent Contrastive LASSO Algorithm for Sparse Graphical Modeling

In this section, we formulate the problem of sparse graphical modeling via the paradigm of model selection described in Section 3.3. Under the assumption of a Gaussian distribution, we have the universal probabilistic model $p(X;\Theta) = \left(\frac{\det\Theta}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\operatorname{tr}[S\Theta]\right)$. By the definition in (3.8), the RSC for sparse graphical modeling is given by

$$RSC(X; v) = \frac{n}{2} h(S, \Lambda) + \ln \int e^{-\frac{n}{2}h(S',\Lambda)} dX', \tag{3.13}$$

given $\Lambda_{ij} = -\ln v_{ij}$ for $1 \le i, j \le m$.

3.5.1 Computing Gradient of RSC

Our objective is to minimize the RSC with respect to v, or equivalently, with respect to Λ. The gradient of the RSC with respect to Λ is given by

$$\frac{\partial}{\partial\Lambda_{ij}} RSC(X; v) = n\left\{ \left|\bar{\Theta}_{ij}(S, \Lambda)\right| - \mathbb{E}_q \left|\bar{\Theta}_{ij}(S', \Lambda)\right| \right\}, \tag{3.14}$$

where the expectation is taken with respect to the probability density $q(X') = e^{-RSC(X';v)}$, and we define $S' = X'^{\top}X'/n$. The derivation of this formula is provided in Section A.1.1. Note that $q(X)$ is a proper density function because the RSC is the length of the lossless coding that attains the continuous version of Kraft's upper bound, i.e., $\int e^{-RSC(X;v)}dX = 1$. Since the above expectation is analytically intractable, regular gradient descent (SGD) algorithms cannot be applied. Therefore, we approximate the expectation by sampling and then utilizing the SGD algorithm. The update formula of Λ_{ij} at the t-th iteration is formulated as

$$\begin{aligned} \Lambda_{ij}^{(t)} &\leftarrow \Pi_{[0,1]}\left[\Lambda_{ij}^{(t-1)} - \eta_{ij}^{(t)}\Delta_{ij}\right], \\ \Delta_{ij} &= \left|\bar{\Theta}_{ij}(S, \Lambda^{(t-1)})\right| - \left|\bar{\Theta}_{ij}(S^{(t)}, \Lambda^{(t-1)})\right|, \end{aligned} \tag{3.15}$$

given the step size $\eta_{ij}^{(t)}$ and the empirical covariance of the t-th sample $S^{(t)} = X^{(t)\top}X^{(t)}/n$. Here, $\Pi_{[0,1]}$ denotes the projection operator onto the interval $[0,1]$. We employ the Adagrad (Duchi et al., 2011) algorithm to determine the step size $\eta_{ij}^{(t)}$, which has only one hyperparameter to choose.

3.5.2 Sampling from RSC

Now, we consider a sampling method for the density $q(X) = e^{-RSC(X;v)}$. There are two major problems that make the sampling difficult. One is the high dimensionality of the variable $X \in \mathbb{R}^{n\times m}$, which increases linearly with the number of samples n. In order to reduce the dependency on n, we project the target distribution onto the space of the sample covariance $S \in \mathbb{R}^{m\times m}$. Since the integrand $\left|\bar{\Theta}_{ij}(S, \Lambda)\right|$ only depends on X through S, we can change the variable as

$$\begin{aligned} \mathbb{E}_q\left|\bar{\Theta}_{ij}(S', \Lambda)\right| &= \int \left|\bar{\Theta}_{ij}(S', \Lambda)\right| q(X')dX' \\ &= \int \left|\bar{\Theta}_{ij}(S, \Lambda)\right| q(S)dS, \end{aligned}$$

where the projected density is given by

$$q(S) \propto (\det S)^{\frac{n-m-1}{2}} \exp\left\{-\frac{n}{2}h(S, \Lambda)\right\}.$$

The second problem is that we cannot directly sample from the density q owing to the normalization factor. In order to avoid computing the density directly, we employ the Metropolis–Hastings algorithm (MHA), which allows us to produce effective samples (i.e., uncorrelated samples) in the long run by making use of the sample of the previous iteration, namely

$$\begin{aligned} S^{(t)} &\leftarrow z\tilde{S} + (1-z)S^{(t-1)}, \\ \tilde{S} &= S^{(t-1)} + \sigma N^{(t)}, \\ z &\sim \text{Bernoulli}\left(\min\left[1, e^{-\beta_q}\right]\right), \end{aligned} \tag{3.16}$$

where each $N_{ij}^{(t)} = N_{ji}^{(t)}$ is independently subject to the normal distribution, and $N_{ii}^{(t)} = 0$

for all $1 \leq i \leq m$. Here, we define β_q as the rejection factor

$$\begin{aligned}\beta_q &\overset{\text{def}}{=} -\ln \frac{q(\tilde{S})}{q(S^{(t-1)})}\\ &= \frac{n}{2}\left\{h(\tilde{S},\Lambda) - h(S^{(t-1)},\Lambda)\right\} - \frac{n-m-1}{2}\ln\frac{\det\tilde{S}}{\det S^{(t-1)}}.\end{aligned}$$

Further, we can adopt the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Rosenthal, 1998), which scales better than the vanilla MHA. It scales down the mixing period of the Markov chain from $O(m^2)$ to $O(m^{2/3})$ by utilizing first-order information as follows:

$$\tilde{S} = S^{(t-1)} + \sigma^{(t)}N^{(t)} + \frac{\sigma^{(t)2}}{2}G^{(t)},$$

where $\sigma^{(t)} = \sigma\det S^{(t-1)}$ is a scale factor considering the geometry of the set of symmetric positive definite matrices, and $G^{(t)} = \left\{G_{ij}(S^{(t-1)})\right\}_{1\leq i,j\leq m}$ denotes the gradient of the logarithmic density, where

$$\begin{aligned}G_{ii}(S) &\overset{\text{def}}{=} 0,\\ G_{ij}(S) &\overset{\text{def}}{=} \frac{\partial}{\partial S_{ij}}\ln q(S)\\ &= (n-m-1)S_{ij}^{-1} - n\bar{\Theta}_{ij}(S,\Lambda) \quad (i \neq j). \qquad (3.17)\end{aligned}$$

The derivation of (3.17) is given in Section A.1.2. The rejection probability is also modified accordingly as

$$z \sim \text{Bernoulli}\left(\min\left[1, e^{-\beta_q-\beta_\pi}\right]\right),$$

where β_π denotes the additional rejection factor due to the asymmetric transition probability $\pi(\tilde{S}|S^{(t-1)})$,

$$\begin{aligned}\beta_\pi &\overset{\text{def}}{=} -\ln\frac{\pi(S^{(t-1)}|\tilde{S})}{\pi(\tilde{S}|S^{(t-1)})}\\ &= \frac{1}{2}\left\|\rho N^{(t)} + \frac{\sigma}{2}\left\{\rho^2 G(S^{(t-1)}) + G(\tilde{S})\right\}\right\|_{\text{F}}^2 -\\ &\quad \frac{1}{2}\left\|N^{(t)}\right\|_{\text{F}}^2 - \frac{m(m-1)}{2}\ln\rho,\end{aligned}$$

given $\rho = \det S^{(t-1)}/\det\tilde{S}$.

In summary, our algorithm consists of two interleaving update chains: the update of the SGD on $\Lambda^{(t)}$ in (3.15) and the update of the MHA on $S^{(t)}$ in (3.16). Note that because the probability density q implicitly depends on $\Lambda^{(t)}$ and $\Lambda^{(t)}$ is not constant, the density q is not constant during iteration. This means that the output of the MHA is not exactly subject to the target density unless we iterate the update of the MHA from line 7 to line 23 in Algorithm 1 an infinite number of times before the update of the SGD. However, it is empirically known that just one update of $S^{(t)}$ is sufficient for practical use if the step size η_t is appropriately decreased. This type of algorithm is known as the *persistent contrastive divergence* algorithm (Tieleman, 2008) for training restricted Boltzmann machines. Since two LASSO estimates are iteratively compared in terms of their magnitudes in our algorithm, we call it the *persistent contrastive LASSO* algorithm (PCLA).

The detailed procedure of the PCLA is presented in Algorithm 1. It takes the arguments $X \in \mathbb{R}^{m\times n}$ as the data, $\Lambda^{(0)} \in \mathbb{R}_+^{m\times m}$ as the initial guess of the hyperparameter, $T > 0$ as the number of iterations, and $\sigma > 0$ and $\eta > 0$ as the scale parameters of the MHA and SGD steps, respectively. Note that there are three lines of computation for the graphical LASSO estimator per iteration. Since the estimator $\bar{\Theta}(S, \Lambda)$ is computed with an iterative algorithm, we can apply warm starting to the computation of the t-th iteration by utilizing the result of the $(t-1)$-th iteration.

Algorithm 1 Sparse graphical modeling via the PCLA

```
Input: $X \in \mathbb{R}^{n\times m}$, $\Lambda^{(0)} \in \mathbb{R}_+^{m\times m}$, $T, \sigma, \eta > 0$
# Initialize
$S \leftarrow X^\top X/n$
$S^{(0)} \leftarrow I_m$
$\bar{\Theta}^{(0)} \leftarrow \bar{\Theta}(S^{(0)}, \Lambda^{(0)})$
$V \leftarrow \mathbf{0}_{m\times m}$
for $t = 1, 2, \dots, T$ do
    # MHA step via the MALA
    # Propose new sample
    $N \leftarrow \texttt{symmetric\_offdiagonal\_normal}(m, m)$
    $\hat{G} \leftarrow G(S^{(t-1)}, \Lambda^{(t-1)})$
    $\tilde{S} \leftarrow S^{(t-1)} + \sigma^{(t)} N + \frac{\sigma^{(t)2}}{2}\hat{G}$
    $\tilde{\Theta} \leftarrow \bar{\Theta}(\tilde{S}, \Lambda^{(t-1)})$
    # Accept/reject new sample
    if $\tilde{S}$ is symmetric positive definite then
        $\hat{h} \leftarrow h(S^{(t-1)}, \Lambda^{(t-1)}), \tilde{h} \leftarrow h(\tilde{S}, \Lambda^{(t-1)})$
        $\tilde{G} \leftarrow G(\tilde{S}, \Lambda^{(t-1)})$
        $\beta_q \leftarrow \frac{n}{2}(\tilde{h} - \hat{h}) + \frac{n-m-1}{2} \ln \rho.$
        $\beta_\pi \leftarrow \frac{\left\|\rho N + \frac{\sigma}{2}\left\{\rho^2 \hat{G} + \tilde{G}\right\}\right\|_{\mathrm{F}}^2}{2} - \frac{\|N\|_{\mathrm{F}}^2}{2} - \frac{m(m-1)}{2} \ln \rho$
        $z \leftarrow \texttt{Bernoulli}\left(\min\left\{1, \exp(-\beta_q - \beta_\pi)\right\}\right)$
        $S^{(t)} \leftarrow z\tilde{S} + (1-z)S^{(t-1)}$
        $\bar{\Theta}^{(t)} \leftarrow z\tilde{\Theta} + (1-z)\bar{\Theta}^{(t-1)}$
    end if
    # SGD step via AdaGrad
    $\Delta \leftarrow \left|\bar{\Theta}(S, \Lambda^{(t-1)})\right| - \left|\bar{\Theta}^{(t)}\right|$
    $V_{ij} \leftarrow V_{ij} + \Delta_{ij}^2$ for all $i$ and $j$
    $\Lambda_{ij}^{(t)} \leftarrow \Pi_{[0,1]}\left[\Lambda_{ij}^{(t-1)} - \frac{\eta}{\sqrt{V_{ij}}}\Delta_{ij}\right]$ for all $i$ and $j$
    $\bar{\Theta}^{(t)} \leftarrow \bar{\Theta}(S^{(t)}, \Lambda^{(t)})$
end for
return $\texttt{sparsity\_pattern}(\bar{\Theta}(S, \Lambda^{(T)}))$
```

3.6 Experimental Results

In order to validate the proposed algorithm, we evaluate PCLA quantitatively answering the following two questions. Firstl, (i) *does PCLA successfully minimize the relaxed stochastic complexity?* Although PCLA is designed to minimize RSC, it is a stochastic iterative algorithm and we do not have any meaningful stopping conditions. Hence, when it is stopped with finite iterations, its performance should be validated empirically. Secondly, (ii) *is the sparsity pattern estimation with PCLA consistent as expected?* Our idea for minimizing RSC to find sparse models is based on the fact that the minimizer of the stochastic complexity is consistent, i.e., in the large sample limit, the estimated sparsity pattern coincides with the true one in probability. However, as we have applied continuous relaxation on the stochastic complexity to obtain RSC, we should also check whether the estimation with PCLA is actually consistent.

Settings We have conducted three experiments, and in each experiment, we have a distinct generative graphical model. The first model is a chain-shaped graph with a size of five, in which the nodes are connected in a line with each partial correlation of ±0.3. The second one is a star-shaped graph with a size of five, in which there is a central node connected to all other nodes with the partial correlation of ±0.2. The third one is a random graph with a size of 10 whose precision matrix Θ^* is generated with the following procedure according to Mazumder and Hastie (2012). Let $A \in \mathbb{R}^{10\times 10}$ be a random matrix with each entry drawn from an i.i.d. normal distribution $\mathcal{N}[0, 1]$, and let $B \in \mathbb{R}^{10\times 10}$ be a symmetric matrix whose entries are identical to $(A + A^\top)/2$, except with 70 randomly chosen entries in symmetric positions shrunk to zero. Then, we have $\Theta^* = B+(1-\kappa_B)I_m$, where κ_B denotes the smallest eigenvalue of B. By its construction described above, Θ^* is symmetric positive definite and hence is a valid precision parameter for Gaussian distributions. The corresponding graphical model consists of 10 vertices and 10 random edges. The generated graphical model is shown in Figure 3.2. Note that the number of possible models amounts to $2^{_{10}C_2} \approx 35$trillion.

Next, we generated an i.i.d. sequence $X = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^{n\times m}$ from the multivariate Gaussian distribution with zero mean, $\mathcal{N}[\mathbf{0}, \Theta^{*-1}]$, and then estimated the precision matrix Θ^* with the PCLA. We generated five independent sequences and executed the PCLA for each experimental configuration to obtain 25 distinct minimum RSC estimates $\bar{J}(X)$. Here, we set $T = 10^6$ and $\eta = 10^{-3}$ and chose σ so that the acceptance rate is near 0.574, as suggested in Roberts and Rosenthal (1998). We chose the initial guess of the hyperparameter as $\Lambda^{(0)} = \mathbf{0}_{m\times m}$ to avoid the trivial local minimizer $\Lambda = \mathbf{1}_{m\times m}$.

We compared the PCLA with the method of five-fold cross validation with an equally spaced grid with a size of 256 for $\Lambda = c\mathbf{1}_{m\times m}$ $(0 \le c \le 1)$, which we call CV here, as a conventional method of hyperparameter selection for the graphical LASSO. We have also included the PCLA with the initial value $\Lambda^{(0)}$ given by CV, called CV+PCLA.

Result: Code Length Since RSC itself is difficult to compute exactly, we evaluate an upper bound of RSC. We substitute the normalizing term of RSC with the upper bound computed with the Bayesian minimax regret, which is introduced in Chapter 5. It is given as

$$RSC(X;\Lambda) \le h(S,\Lambda) + \sum_{i<j} \ln\left[1 + \frac{2R^2}{\Lambda_{ij}^2}\exp\left(-\frac{\Lambda_{ij}^2}{2R^2}\right)\right],$$

where R^{-1} is a lower bound of the smallest eigenvalue of the precision matrix Θ. In our setting, the eigenvalue can be arbitrarily small and we cannot fix $R < +\infty$ before training. Thus, we adopt a two-stage coding approach such that

$$R\tilde{S}C(X;\Lambda) \overset{\mathrm{def}}{=} \min_{k\in\mathbb{N}} h(S,\Lambda) + \sum_{i<j} \ln\left[1 + \frac{2R_k^2}{\Lambda_{ij}^2 e}\exp\left(-\frac{\Lambda_{ij}^2}{2R_k^2}\right)\right] + L(R_k),$$

where $R_k = 2^{-k}$, and $L(R_k)$ denotes the code length for the lower bound itself. We adopt the universal code length for integers suggested by Rissanen (1983).

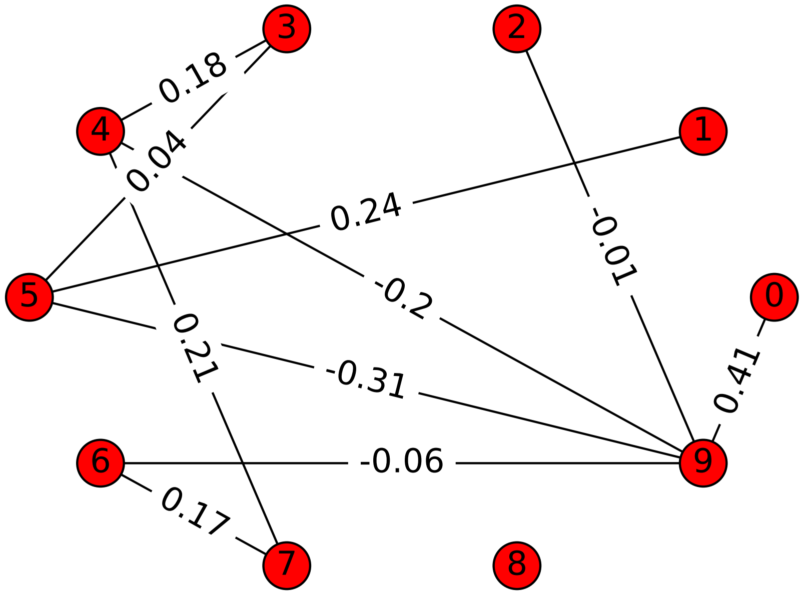


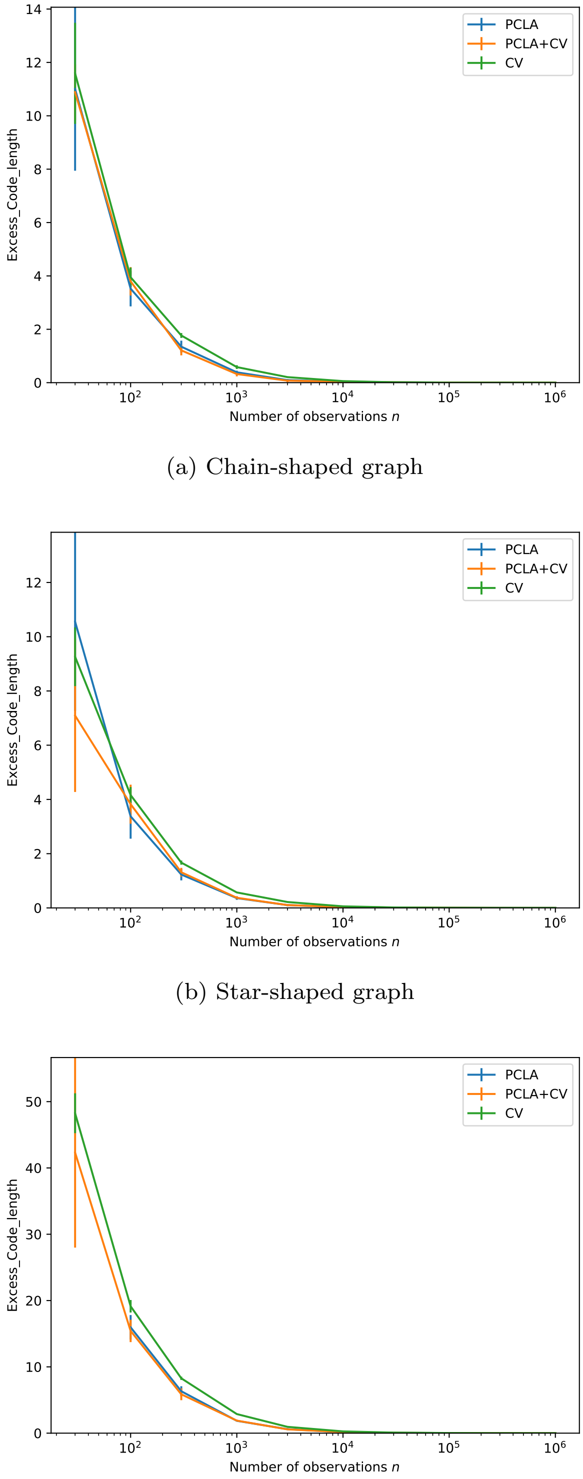
Fig. 3.2: Synthetic random undirected graph with a size of 10. There are 10 randomly generated edges in the graph. The partial correlation coefficients according to the true precision matrix, $-\Theta^*_{ij}/\sqrt{\Theta^*_{ii}\Theta^*_{jj}}$, are presented on the corresponding edges.

Figure 3.3 shows the upper bound of RSC corresponding to the selected Λ. In comparison to CV, the code lengths of the estimates produced with PCLA are no larger. Specifically, when the size of model m is large, they are significantly smaller than CV (note the difference in the scale of vertical axes among figure parts (a)–(c)). Therefore, it is indirectly implied that PCLA successfully minimizes RSC as we intended.

Result: Consistency Figure 3.4 shows the Hamming distance of the estimated sparsity pattern $\bar{J}(X)$ relative to the true pattern J^* when varying the number of observations n. As shown in the figure, the estimate of the PCLA converges to the true one for the first two experiments as n increases. For the last experiment, owing to the lightweight edges in the graph, e.g., $e = (2,9),(3,5)$, the Hamming distance does not converge to zero. However, it still gives a much better approximation than CV for large n. Further, the warm-starting PCLA with CV improves the performance, where n is relatively small and gives almost identical performance with large n.

3.7 Concluding Remarks

We presented a paradigm that relaxes the problem of sparse model selection to the problem of hyperparameter selection of ℓ_1-regularization in view of the MDL principle. The relax-

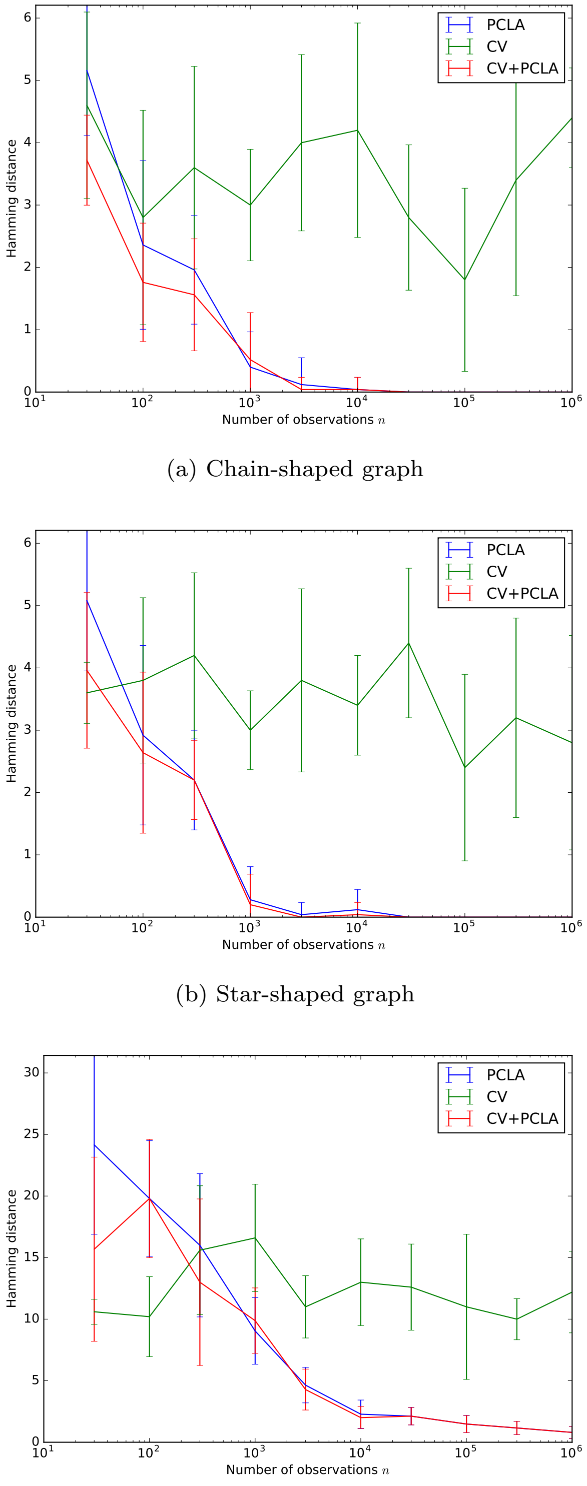


(a) Chain-shaped graph

(b) Star-shaped graph

(c) Random graph

Fig. 3.3: Upper bounds of the excess of the relaxed stochastic complexity relative to the expected optimal code length. Each figure shows the results of the respective experiment, and the vertical whiskers in them show the ranges of ±1 standard deviation. The horizontal axes represent the number of observations n. It is shown that the PCLA successfully lowers RSC compared to CV, especially with the larger model (c), and the initialization with CV takes some effect with smaller sample size.



(a) Chain-shaped graph

(b) Star-shaped graph

(c) Random graph

Fig. 3.4: Hamming distance of the estimates of the sparsity pattern relative to the true one. Each figure shows the results of the respective experiment, and the vertical whiskers in them show the ranges of ±1 standard deviation. The horizontal axes represent the number of observations n. The distance of the PCLA converges to zero at $n = 1000$ for the first two experiments but not in the third experiment. On the other hand, the distance of CV is volatile throughout all experiments. Although CV+PCLA performs almost identically to the PCLA for large n, it tends to improve the performance of the PCLA where n is small.

ation immediately induces criteria for the hyperparameter selection of ℓ_1-regularization as a feasible approximation of the MDL criteria for model selection. We have also derived an iterative algorithm for its optimization, the PCLA, by which we can efficiently choose the optimal undirected graphical model out of an exponentially large number of candidates. The experiments show that the resulting estimate successfully identifies the true structure of graphical models as the number of observations increases.

To conclude, let us present a number of possible applications and extensions of the proposed scheme. In the proposed algorithm, the variable v obtained by the continuous relaxation can be regarded as the storage of messages passed by the data X about the preference for the sparsity patterns $J \subset [d]$. Therefore, in the context of a time series analysis for example, we suggest that the variable v by itself—not $\psi_X(v)$—can be utilized to detect changes in the true sparsity pattern. Moreover, the iterative nature of the PCLA allows efficient online learning of v for streaming data. Other interesting future work includes the extension of the PCLA towards the problem of *regression* with ℓ_1-regularization. We also require theoretical analyses of the validity of the relaxation, such as a convex analysis of the relaxed objective function $RSC(X; v)$ and the consistency of its minimizer $\bar{J}(X)$.

Chapter 4

High-dimensional Penalty Selection via Analytic Approximation of Minimax Regret

We tackle the problem of penalty selection of regularization on the basis of the MDL principle. In particular, we consider that the design space of the penalty function is high-dimensional. In this situation, the LNML-minimization approach is favorable, because LNML quantifies the goodness of regularized models with any forms of penalty functions in view of the MDL principle and guides us to a good penalty function through the high-dimensional space. However, the minimization of LNML entails two major challenges: 1) the computation of the normalizing factor of LNML and 2) its minimization in high-dimensional spaces. In this chapter, we present a novel regularization selection method (MDL-RS), in which a tight upper bound of LNML (uLNML) is minimized with local convergence guarantee[*1]. Our main contribution is the derivation of uLNML, which is a uniform-gap upper bound of LNML in an analytic expression. This solves the above challenges in an approximate manner because it allows us to accurately approximate LNML and then efficiently minimize it. The experimental results show that MDL-RS improves the generalization performance of regularized estimates specifically when the model has redundant parameters.

4.1 Motivation

We are concerned with the problem of learning with *redundant* models (or hypothesis classes). This setting is not uncommon in real-world machine learning and data mining problems. This is because the amount of available data is sometimes limited owing to the cost of data collection (e.g., in biomedical data analyses), while researchers can come up with an unbounded number of models for explaining the data that may contain a number of irrelevant features. For example, in sparse regression, one may consider a number of features that is much larger than the number in the data, assuming that useful features are actually scarce (Rish and Grabarnik, 2014). Another example is statistical conditional-dependency estimation, in which the number of the parameters to estimate is quadratic compared to the number of random variables, while the number of nonzero coefficients is often expected to be sub-quadratic.

In the context of such a redundant model, there is a danger of overfitting, which is where the model fits the present data excessively well but does not generalize well. To

[*1] The content of this chapter was published in Miyaguchi and Yamanishi (2018b)

address this, we introduce regularization and reduce the complexity of the models by taking the regularized empirical risk minimization (RERM) approach (Shalev-Shwartz and Ben-David, 2014). In RERM, we minimize the sum of the loss and penalty functions to estimate parameters. However, the choice of penalty function should be made cautiously as it controls the bias-variance trade-off of the estimates, and hence has a considerable effect on the generalization capability.

In conventional methods for selecting such hyperparameters, a two-step approach is usually followed. First, a candidate set of penalty functions is configured (possibly randomly). Then, a penalty selection criterion is computed for each candidate and the best one is chosen. Note that this method can be applied to any penalty selection criteria. Sophisticated approaches like Bayesian optimization (Mockus et al., 2013) and gradient-based methods (Larsen et al., 1996) also tend to leave the criterion as a black-box. Although leaving it as a black-box is advantageous in that it works for a wide range of penalty selection criteria, a drawback is that the full information of each specific criterion cannot be utilized. Hence, the computational costs can be unnecessarily large if the design space of the penalty function is high-dimensional.

In this chapter, we propose a novel penalty selection method that utilizes information about the objective criterion efficiently on the basis of the MDL principle (Rissanen, 1978). We especially focus on the LNML code length (Grünwald, 2007) because the LNML code length measures the complexity of regularized models without making any assumptions on the form of the penalty functions. Moreover, it places a tight bound on the generalization error (Grünwald and Mehta, 2017). However, the actual use of LNML on large models is limited so far. This is owing to the following two issues.

I1) LNML contains a normalizing constant that is difficult to compute, especially for large models. This tends to make the evaluation of the code length intractable.

I2) Since the normalizing term is defined as a non-closed form of the penalty function, efficient optimization of LNML is non-trivial.

Next, solutions are described for the above issues. First, we derive an upper bound of the LNML code length, namely uLNML. The key idea is that the normalizing constant of LNML, which is not analytic in general, is characterized by the smoothness of loss functions, which can often be upper-bounded by an analytic quantity. As such, uLNML exploits the smoothness information of the loss and penalty functions to approximate LNML with much smaller computational costs, which solves I1. Moreover, within the framework of the concave-convex procedure (CCCP) (Yuille and Rangarajan, 2003), we propose an efficient algorithm for finding a local minimima of uLNML, i.e., finding a good penalty function in terms of LNML. This algorithm only adds an extra analytic step to the iteration of the original algorithm for the RERM problem, regardless of the dimensionality of the penalty design. Thus, I2 is addressed. We combine these two methods and propose a novel method of penalty selection named *MDL regularization selection (MDL-RS)*.

We also validate the proposed method from both a theoretical and empirical perspective. Specifically, as our method relies on approximation of uLNML and the CCCP algorithm relies on uLNML, the following questions arise.

Q1) How well does uLNML approximate LNML?

Q2) Does the CCCP algorithm on uLNML perform well with respect to generalization compared to the other methods for penalty selection?

To answer Q1, we show that the gap between uLNML and LNML is uniformly bounded under smoothness and convexity conditions. As for Q2, from our experiments on example models involving both synthetic and benchmark datasets, we found that MDL-RS is at least comparable to the other methods and even outperforms them when models are highly redundant, as we expected. Therefore, the answer is affirmative.

The remainder of this chapter is organized as follows. In Section 2, we introduce a novel penalty selection criteria called uLNML with uniform gap guarantees. Section 3 demonstrates some examples of the calculation of uLNML. Section 4 provides the minimization algorithm of uLNML and discusses its convergence property. Conventional methods for penalty selection are reviewed in Section 5. Experimental results are shown in Section 6. Finally, Section 7 concludes the chapter and discusses future work.

4.2 Method: Analytic Upper Bound of LNMLs

In this section, we first briefly review the definition of RERM and the notion of penalty selection. Then, we introduce the LNML code length. Finally, as our main result, we show an upper bound of LNML (uLNML) with approximation error analyses and several examples.

4.2.1 Preliminary: Regularized Empirical Risk Minimization (RERM)

Let $f_X : \mathbb{R}^p \to \overline{\mathbb{R}}(= \mathbb{R} \cup \{\infty\})$ be an extended-value loss function of parameter $\theta \in \mathbb{R}^p$ with respect to data $X = (x_1, \dots, x_n) \in \mathcal{X}^n$. We assume $f_X(\theta)$ is a log-loss (but not limited to i.i.d. loss), i.e., it is normalized with respect to some base measure ν over $\mathcal{X}^n$, where $\int_{\mathcal{X}^n} \exp\{-f_X(\theta)\}\, d\nu(X) = 1$ for all θ in some closed subset $\Omega_0 \subset \mathbb{R}^p$. Here, x_i can be a pair of datum and label (x_i, y_i) in the case of supervised learning. For simplicity, e drop the subscript of X and write $f(\theta) = f_X(\theta)$ if there is no confusion.

The regularized empirical risk minimization (RERM) with convex domain $\Omega \subset \Omega_0$ is defined as the following minimization:

$$\mathrm{RERM}(\lambda): \qquad \text{minimize } f_X(\theta) + g(\theta, \lambda) \quad \text{s.t.} \quad \theta \in \Omega, \tag{4.1}$$

where $g : \mathbb{R}^p \times \mathbb{A} \to \overline{\mathbb{R}}$ denotes the penalty function, and $\lambda \in \mathbb{A} \subset \mathbb{R}^d$ is the hyperparameter that parametrizes the shape of penalty on θ. We assume that at least one minimizer always exists in Ω and denote it as $\hat{\theta}(X, \lambda)$. Here, we focus on a special case of RERM in which the penalty is linear in terms of λ:

$$g(\theta, \lambda) = \sum_{j=1}^{d} \lambda_j g_j(\theta), \quad \lambda_j \geq 0 \quad (j = 1, \dots, d), \tag{4.2}$$

and $\mathbb{A} \subset \mathbb{R}_+^d$ is a convex set of positive vectors. Finally, let us define $D(\lambda) \overset{\text{def}}{=} \{X \in \mathcal{X}^n \mid \hat{\theta}(X, \lambda) \in \Omega^{\circ}\}$, where Ω° is the interior of the set Ω. Then, we assume that the following regularity condition holds:

Assumption 1 (Regular penalty) $D(\lambda)$ is monotonically increasing, i.e.,

$$\lambda \leq \lambda' \Rightarrow D(\lambda) \subset D(\lambda'),$$

or equivalently,

$$\lambda \leq \lambda',\ \hat{\theta}(X, \lambda) \in \Omega^{\circ} \Rightarrow \hat{\theta}(X, \lambda') \in \Omega^{\circ}.$$

Regularization is beneficial from two perspectives. It improves the condition number of the optimization problem, and hence it enhances the numerical stability of the estimates. It also prevents the estimate from overfitting to the training data X, which hence reduces generalization error.

However, these benefits come with an appropriate penalization. If the penalty is too large, the estimate will be biased. If the penalty is too small, the regularization no longer takes effect and the estimate is likely to overfit. Therefore, we are motivated to select good λ as a function of data X.

4.2.2 LNML

In order to select an appropriate hyperparameter λ, we introduce the LNML code length as a criterion for the penalty selection. The LNML code length associated with RERM(λ) is given by

$$\mathcal{L}(X \mid \lambda) \overset{\text{def}}{=} \min_{\theta \in \Omega} f_X(\theta) + g(\theta, \lambda) + \ln Z(\lambda), \tag{4.3}$$

where $Z(\lambda) \overset{\text{def}}{=} \int \max_{\theta \in \Omega} \exp\{-f_X(\theta) - g(\theta, \lambda)\} \, d\nu(X)$ is the normalizing factor of LNML.

Note that LNML is originaly derived by generalization of the Shtarkov's minimax coding strategy (Shtarkov, 1987; Grünwald, 2007). The normalizing factor $Z(\lambda)$ can be seen as a penalization of the complexity of RERM(λ). It quantifies how much RERM(λ) will overfit to random data. If the penalty g is small such that the minimum in (4.1) always takes a low value for all $X \in \mathcal{X}^n$, $Z(\lambda)$ becomes large. Specifically, any constant shift on the RERM objective that does not change the RERM estimator $\hat{\theta}$ does not change LNML since $Z(\lambda)$ cancels it out. Moreover, recent advances in the analysis of LNML show that it bounds the generalization error of $\hat{\theta}(X, \lambda)$ (Grünwald and Mehta, 2017). Thus, our primary goal is to minimize the LNML code length (4.3).

4.2.3 Upper Bound of LNML (uLNML)

The direct computation of the normalizing factor $Z(\lambda)$ is often impossible because it requires integration of the RERM objective (4.1) over all possible data. To avoid computational difficulty, we introduce an upper bound of $Z(\lambda)$ that is analytic with respect to λ. Then, adding the upper bound to the RERM objective, we have an upper bound of the LNML code length itself.

To derive the bound, let us define the following $\overline{H}$-upper smoothness condition.

Definition 2 ($\overline{H}$-upper smoothness) Let $\overline{H} \in \mathbb{S}^p_{++} \subset \mathbb{R}^{p \times p}$ be a symmetric positive definite matrix. A function $f : \mathbb{R}^p \to \overline{\mathbb{R}}$ is called $\overline{H}$-upper smooth, or $(\overline{H}, c, r)$-upper smooth to avoid any ambiguity, if there exists a constant $c \geq 0$, a vector-valued function $\xi : \mathbb{R}^p \to \mathbb{R}^p$, and a monotonically increasing function $r : \mathbb{R} \to \overline{\mathbb{R}}_+$ such that

$$f(\psi) - f(\theta) \leq c + \langle \xi(\theta), \psi - \theta \rangle + \frac{1}{2} \|\psi - \theta\|^2_{\overline{H}} + r(\|\theta - \psi\|^2), \quad \forall \psi \in \mathbb{R}^p, \forall \theta \in \Omega,$$

where $\|\psi - \theta\|^2_{\overline{H}} \overset{\text{def}}{=} (\psi - \theta)^\top \overline{H}(\psi - \theta)$ and $r(t^2) = o(t^2)$.

There are a few remarks to make about this definition. The $\overline{H}$-upper smoothness is a condition that is weaker than that of standard smoothness. In particular, ρ-smoothness implies ρI_p-upper smoothness, and all the bounded functions are upper smooth with their respective triples $(\overline{H}, c, r)$. Moreover, the function r only describes the behavior of the function outside of Ω. Thus, if the penalty $g_j(\theta)$ is upper smooth, we can take the corresponding r as zero without changing the solution of RERM.

Now, assume that f_X and g_j $(j = 1, \ldots, d)$ are upper-smooth:

Assumption 2 (Upper-smooth objective) Both of the following equivalent conditions are satisfied:

i) f_X is $(\overline{H}_0, c_0, r)$-upper smooth for all $X \in \mathcal{X}^n$ and g_j is $(\overline{H}_j, c_j, 0)$-upper smooth $(j = 1, \dots, d)$.
ii) $f_X(\cdot) + g(\cdot, \lambda)$ is $(\overline{H}(\lambda), c(\lambda), r)$-upper smooth for all $X \in \mathcal{X}^n$ and all $\lambda \geq \mathbf{0}$, where $\overline{H}(\lambda) \overset{\mathrm{def}}{=} \overline{H}_0 + \sum_{j=1}^d \lambda_j \overline{H}_j$ and $c(\lambda) \overset{\mathrm{def}}{=} c_0 + \sum_{j=1}^d \lambda_j c_j$.

Then, the following theorem states that the upper bound depends on f_X and g only through their smoothness.

Theorem 3 (Upper bound of $Z(\lambda)$) Suppose that Assumption 2 holds. Let $R(H; U) = \mathbb{E}_{z \sim \mathcal{N}_p[\mathbf{0}, H^{-1}]} \left[\mathbb{1}U(z) \exp\left\{ -r\left(\|z\|^2 \right) \right\} \right]$. Then, for all the symmetric neighbors of the origin $U \subset \mathbb{R}^p$ satisfying $\Omega + U \subset \Omega_0$, we have

$$Z(\lambda) \leq \bar{Z}(\lambda) \overset{\mathrm{def}}{=} \frac{1}{R(\overline{H}_0; U)} \frac{e^{c(\lambda)} \det \overline{H}(\lambda)^{\frac{1}{2}}}{\sqrt{2\pi}^p} \int_{\Omega + U} e^{-g(\theta, \lambda)} d\theta. \tag{4.4}$$

Proof Let $q_\lambda(X) \overset{\mathrm{def}}{=} \int_{\Omega+U} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\theta$. First, by Hölder's inequality, we have

$$\begin{aligned} Z(\lambda) &= \int_{\mathcal{X}^n} \max_{\theta \in \Omega} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\nu(X) \\ &\leq \left\| \frac{\max_{\theta \in \Omega} \exp\{-f_{\cdot}(\theta) - g(\theta, \lambda)\}}{q_\lambda(\cdot)} \right\|_\infty \|q_\lambda(\cdot)\|_{L^1(\nu)} \\ &= \underbrace{\sup_{X \in \mathcal{X}^n} \max_{\theta \in \Omega} \frac{\exp\{-f_X(\theta) - g(\theta, \lambda)\}}{q_\lambda(X)}}_{A} \underbrace{\int_{\mathcal{X}^n} q_\lambda(X) d\nu(X)}_{B}, \end{aligned}$$

where $\|\cdot\|_\infty$ denotes the uniform norm, and $\|\cdot\|_{L^1(\nu)}$ denotes the L^1-norm with respect to measure ν. Then, we will bound A and B in the right-hand side, respectively. Since we assume that $f_X(\theta)$ is a logarithmic loss if $\theta \in \Omega_0$, the second factor is simply evaluated using Fubini's theorem,

$$\begin{aligned} B &= \iint_{(\Omega+U) \times \mathcal{X}^n} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\theta d\nu(X) \\ &= \int_{\Omega+U} e^{-g(\theta, \lambda)} d\theta. \end{aligned}$$

On the other hand, by the $\overline{H}(\lambda)$-upper smoothness of $f(\theta) + g(\theta, \lambda)$, we have

$$\begin{aligned}
A^{-1} &= q_\lambda(X) \exp\left\{f_X(\theta) + g(\theta)\right\} \\
&= \int_{\Omega+U} \exp\left\{f_X(\theta) + g(\theta,\lambda) - f_X(\psi) - g(\psi,\lambda)\right\} d\psi \\
&\geq \int_{\Omega+U} \exp\left\{-c(\lambda) - \langle \xi(\theta), \psi - \theta\rangle - \frac{1}{2}\left\|\psi - \theta\right\|^2_{\overline{H}(\lambda)} - r(\|\psi - \theta\|^2)\right\} d\psi \\
&\geq e^{-c(\lambda)} \int_U \exp\left\{- \langle \xi(\theta), z\rangle - \frac{1}{2}\left\|z\right\|^2_{\overline{H}(\lambda)} - r(\|z\|^2)\right\} dz \\
&\geq e^{-c(\lambda)} \int_U \exp\left\{- \frac{1}{2}\left\|z\right\|^2_{\overline{H}(\lambda)} - r(\|z\|^2)\right\} dz \\
&= e^{-c(\lambda)} \frac{\sqrt{2\pi}^p}{\det \overline{H}(\lambda)^{\frac{1}{2}}} R(\overline{H}(\lambda); U) \\
&\geq e^{-c(\lambda)} \frac{\sqrt{2\pi}^p}{\det \overline{H}(\lambda)^{\frac{1}{2}}} R(\overline{H}_0; U).
\end{aligned}$$

This concludes the proof. ■

The upper bound in Theorem 3 can be easily computed by ignoring the constant factor $R(\overline{H}_0, U)^{-1}$ given the upper smoothness of f_X and $g(\cdot, \lambda)$. In particular, the integral $\int_{\Omega+U} e^{-g(\theta,\lambda)} d\theta$ can be evaluated in a closed form if one chooses a suitable class of penalty functions with a suitable neighbor U (e.g., quadratic penalty functions with $U = \mathbb{R}^p$). Therefore, we adopt this upper bound as an alternative of the LNML code length, namely *uLNML*:

$$\bar{\mathcal{L}}(X|\lambda) \overset{\text{def}}{=} \min_{\theta\in\Omega} f_X(\theta) + g(\theta,\lambda) + \ln \bar{Z}(\lambda), \tag{4.5}$$

where $\bar{Z}(\lambda)$ is defined in Theorem 3. Note that the symmetric set U should be fixed beforehand. In practice, we recommend simply taking $U = \mathbb{R}^p$ because uLNML with $U = \mathbb{R}^p$ bounds uLNMLs with $U \neq \mathbb{R}^p$, and then we have

$$\bar{\mathcal{L}}(X|\lambda) = \min_{\theta\in\Omega} f_X(\theta) + g(\theta,\lambda) + c(\lambda) + \frac{1}{2}\ln\det \overline{H}(\lambda) + \ln \int_{\mathbb{R}^p} e^{-g(\psi,\lambda)} d\psi + \text{const.}$$

However, for the sake of the later analysis, we leave U to be arbitrary.

We present two useful specializations of uLNML with respect to the penalty function $g(\theta, \lambda)$. One is the Tikhonov regularization, known as the ℓ_2-regularization.

Corollary 1 (Bound for Tikhonov regularization) Suppose that Assumption 2 holds with f_X. Suppose that $g(\theta, \lambda) = \frac{1}{2}\sum_{j=1}^p \lambda_j \theta_j^2$, where $\lambda_j > 0$ for all $1 \leq j \leq p$. Then, we have

$$Z(\lambda) \leq \frac{e^{c_0}}{R(\overline{H}_0; \mathbb{R}^p)} \sqrt{\frac{\det(\overline{H}_0 + \operatorname{diag}\lambda)}{\det \operatorname{diag}\lambda}}.$$

Proof This claim follows from the setting of $U = \mathbb{R}^p$ in Theorem 3 and the fact that $g(\cdot, \lambda)$ is $(\operatorname{diag}\lambda, 0, 0)$-upper smooth. ■

The other specialization is that of LASSO (Tibshirani, 1996), known as ℓ_1-regularization. This is useful if one requires sparse estimates of $\hat{\theta}(X, \lambda)$.

Corollary 2 (Bound for LASSO) Suppose that Assumption 2 holds with f_X. Suppose that $g(\theta,\lambda)=\sum_{j=1}^{p}\lambda_j\,|\theta_j|$, where $\lambda_j>0$ for all $1\le j\le p$. Then, we have

$$Z(\lambda)\le\frac{e^{c_0}}{R(\overline{H}_0;\mathbb{R}^p)}\sqrt{\frac{e}{2\pi}}^{\,p}\sqrt{\frac{\det(\overline{H}_0+(\operatorname{diag}\lambda)^2)}{\det(\operatorname{diag}\lambda)^2}}.$$

Proof As in the proof of Corollary 1, it follows from Theorem 3 and the fact that $g(\cdot,\lambda)$ is $((\operatorname{diag}\lambda)^2,1/2,0)$-upper smooth. ■

Finally, we present a useful extension for RERMs with Tikhonov regularization, which contains the inverse temperature parameter $\beta\in[a,b]$ $(0<a\le b)$ as a part of the parameter:

$$f_X(\beta,\theta)=\beta\tilde{f}_X(\theta)+\ln C(\beta), \tag{4.6}$$

$$g(\beta,\theta,\lambda)=\beta\tilde{g}(\theta,\lambda)=\beta\sum_{j=1}^{d}\frac{\lambda_j}{2}\theta_j^2, \tag{4.7}$$

where $C(\beta)\overset{\text{def}}{=}\int e^{-\beta\tilde{f}_X(\theta)}d\nu(X)<\infty$ is the normalizing constant of the loss function. Here, we assume that $C(\beta)$ is independent of the non-temperature parameter θ. Interestingly, the uLNML of variable temperature models (4.6) (4.7) coincides with that of the fixed temperature models given in Corollary 1 except with a constant.

Corollary 3 (Bound for variable temperature model) Let $(\beta,\theta)\in[a,b]\times\Omega$ be the parameter of the model (4.6). Suppose that $\tilde{f}_X(\theta)$ is $(\overline{H}_0,c_0,r)$-upper smooth for all $X\in\mathcal{X}^n$. Then, there exists a constant $C_{[a,b]}$ such that

$$Z(\lambda)\le\frac{C_{a,b}\;e^{(b+a/2)c_0}}{R(\frac{a}{2}\overline{H}_0;\mathbb{R}^p)}\sqrt{\frac{\det\left(\overline{H}_0+\operatorname{diag}\lambda\right)}{\det\operatorname{diag}\lambda}}.$$

Proof Let $\tilde{F}_X(\lambda)=\min_{\theta\in\Omega}\tilde{f}_X(\theta)+\tilde{g}(\theta,\lambda)$. Let $W=[a/2,b+a/2]$ and $\tilde{q}_\lambda(X)=\int_W\exp\left\{-\beta\tilde{F}_X(\lambda)-\ln C(\beta)\right\}d\beta$. Note that $\ln C(\beta)$ is continuous and hence bounded over W, which implies that it is upper smooth. Let $(h_\beta,c_\beta,r_\beta)$ be the upper smoothness of $\ln C(\beta)$ over W. Then, using the same techniques as in Theorem 3, we have

$$\begin{aligned}
Z(\lambda)&=\int\max_{\beta\in[a,b],\ \theta\in\Omega}\exp\left\{-\beta\left[\tilde{f}_X(\theta)-\tilde{g}(\theta,\lambda)\right]-\ln C(\beta)\right\}d\nu(X)\\
&\le\max_{\beta\in[a,b]}\sup_{X\in\mathcal{X}^n}\exp\left\{-\beta\tilde{F}_X(\lambda)-\ln C(\beta)-\ln\tilde{q}_\lambda(X)\right\}\int\tilde{q}_\lambda(X)d\nu(X)\\
&\le C_{a,b}\int_W d\beta\int\max_{\theta\in\Omega}\exp\left\{-\beta\tilde{f}_X(\theta)-\beta\tilde{g}(\theta,\lambda)-\ln C(\beta)\right\}d\nu(X)\\
&\le C_{a,b}\int_W d\beta\frac{e^{\beta c_0}}{R(\beta\overline{H}_0;\mathbb{R}^p)}\sqrt{\frac{\det\beta\left(\overline{H}_0+\operatorname{diag}\lambda\right)}{\det\beta\operatorname{diag}\lambda}}\\
&=\frac{C_{a,b}\;e^{(b+a/2)c_0}}{R(\frac{a}{2}\overline{H}_0;\mathbb{R}^p)}\sqrt{\frac{\det\left(\overline{H}_0+\operatorname{diag}\lambda\right)}{\det\operatorname{diag}\lambda}},
\end{aligned}$$

where $C_{a,b}\overset{\text{def}}{=}\frac{e^{c_\beta}}{R_\beta(h_\beta;[-a/2,a/2])}\sqrt{\frac{h_\beta}{2\pi}}$. ■

4.2.4 Gap between LNML and uLNML

In this section, we evaluate the tightness of uLNML. To this end, we now bound LNML from below. The lower bound is characterized by the strong convexity of f_X and $g(\cdot, \lambda)$.

Definition 3 ($\underline{H}$-strong convexity) Let $\underline{H} \in \mathbb{S}^p_{++} \subset \mathbb{R}^{p \times p}$ be a symmetric positive definite matrix. A function $f(\theta)$ is $\underline{H}$-strongly convex if there exists a vector-valued function $\xi : \mathbb{R}^p \to \mathbb{R}^p$ such that

$$f(\psi) - f(\theta) \geq \langle \xi(\theta), \psi - \theta \rangle + \frac{1}{2} \|\psi - \theta\|^2_{\underline{H}}, \quad \forall \psi \in \mathbb{R}^p, \forall \theta \in \mathbb{R}^p.$$

Note that $\underline{H}$-strong convexity can be seen as the matrix-valued version of the standard strong convexity. Now, assume the strong convexity of f_X and g_j:

Assumption 3 (Strongly convex objective) Both of the following equivalent conditions are satisfied:

i) f_X is $\underline{H}_0$-strongly convex for all $X \in \mathcal{X}^n$, and g_j is $\underline{H}_j$-strongly convex ($j = 1, \ldots, d$).

ii) $f_X(\cdot) + g(\cdot, \lambda)$ is $\underline{H}(\lambda)$-strongly convex for all $X \in \mathcal{X}^n$ and all $\lambda \geq \mathbf{0}$, where $\underline{H}(\lambda) \overset{\text{def}}{=} \underline{H}_0 + \sum_{j=1}^d \lambda_j \underline{H}_j$.

Then, we have the following lower bound on $Z(\lambda)$.

Theorem 4 (Lower bound on $Z(\lambda)$) Suppose that Assumption 1 and 3 hold. Let $T(V) \overset{\text{def}}{=} \inf_{\psi \in V} \int_{D(\mathbf{0})} \exp\{-f_X(\psi)\} d\nu(X)$. Then, for all $V \subset \Omega_0$, we have

$$Z(\lambda) \geq T(V) \frac{\det \underline{H}(\lambda)^{\frac{1}{2}}}{\sqrt{2\pi}^p} \int_V e^{-g(\theta,\lambda)} d\theta. \tag{4.8}$$

Proof Let $q_\lambda(X) \overset{\text{def}}{=} \int_V \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\theta$. First, from the positivity of q_λ, we have

$$\begin{aligned} Z(\lambda) &= \int_{\mathcal{X}^n} \max_{\theta \in \Omega} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\nu(X) \\ &\geq \int_{D(\lambda)} \max_{\theta \in \Omega^\circ} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\nu(X) \\ &\geq \inf_{X \in D(\lambda)} \underbrace{\max_{\theta \in \Omega^\circ} \frac{\exp\{-f_X(\theta) - g(\theta, \lambda)\}}{q_\lambda(X)}}_{A} \underbrace{\int_{D(\lambda)} q_\lambda(X) d\nu(X)}_{B}. \end{aligned}$$

Then, we bound from below A and B in the right-hand side, respectively. Since we assumed that $f_X(\theta)$ is a logarithmic loss, the second factor is simply evaluated using Fubini's theorem:

$$\begin{aligned} B &\geq \iint_{V \times D(\mathbf{0})} \exp\{-f_X(\theta) - g(\theta, \lambda)\} d\theta d\nu(X) \\ &\geq T(V) \int_V e^{-g(\theta,\lambda)} d\theta, \end{aligned}$$

where the first inequality follows from Assumption 1.

On the other hand, by Lemma 18 in Section A.2, we have

$$\begin{aligned}
A^{-1} &= q_\lambda(X) \min_{\theta \in \Omega^\circ} \exp\left\{ f_X(\theta) + g(\theta) \right\} \\
&= \int_\Omega \exp\left\{ \min_{\theta \in \Omega^\circ} f_X(\theta) + g(\theta, \lambda) - f_X(\psi) - g(\psi, \lambda) \right\} d\psi \\
&\leq \int_\Omega \exp\left\{ -\frac{1}{2} \|z\|^2_{\underline{H}(\lambda)} \right\} dz \\
&\leq \frac{\sqrt{2\pi}^p}{\det \underline{H}(\lambda)^{\frac{1}{2}}}.
\end{aligned}$$

This concludes the proof. ■

The lower bound in Theorem 4 has a similar form to the upper bound $\bar{Z}(\lambda)$. Therefore, combining Theorem 4 with Theorem 3, we have a uniform gap bound of uLNML.

Theorem 5 (Uniform gap bound of uLNML) Suppose that the assumptions made in Theorem 3 and 4 is satisfied. Suppose that the penalty function is quadratic, i.e., $\overline{H}_j = \underline{H}_j$ and $c_j = 0$ for all $j = 1, \ldots, d$. Then, the gap between LNML and uLNML is uniformly bounded for all $X \in \mathcal{X}^n$ and $\lambda \in \mathbb{A}$ as

$$\bar{\mathcal{L}}(X|\lambda) - \mathcal{L}(X|\lambda) \leq c_0 + \frac{1}{2} \ln \frac{\det \overline{H}_0}{\det \underline{H}_0} - \ln R(\overline{H}_0; U) - \ln T(\Omega + U), \tag{4.9}$$

where $R(\overline{H}_0; U)$ and $T(V)$ are defined as in the preceding theorems.

Proof From Theorem 3 and Theorem 4, we have

$$\begin{aligned}
\bar{\mathcal{L}}(X|\lambda) - \mathcal{L}(X|\lambda) &\leq \ln \bar{Z}(\lambda) - \ln Z(\lambda) \\
&\leq c(\lambda) + \frac{1}{2} \ln \frac{\det \overline{H}(\lambda)}{\det \underline{H}(\lambda)} - \ln R(\overline{H}_0; U) - \ln T(V) \\
&\quad + \ln \frac{\int_{\Omega + U} e^{-g(\theta, \lambda)} d\theta}{\int_V e^{-g(\theta, \lambda)} d\theta},
\end{aligned}$$

where $c(\lambda) = c_0$ from the assumption. Taking $V = \Omega + U$ to cancel out the last term, we have

$$\bar{\mathcal{L}}(X|\lambda) - \mathcal{L}(X|\lambda) \leq c_0 + \frac{1}{2} \ln \frac{\det \overline{H}(\lambda)}{\det \underline{H}(\lambda)} - \ln R(\overline{H}_0; U) - \ln T(\Omega + U). \tag{4.10}$$

Let $\kappa(Q) \overset{\text{def}}{=} \ln \frac{\det(\overline{H}_0 + Q)}{\det(\underline{H}_0 + Q)}$ for $Q \in \mathbb{S}^p_+$, and let $\underline{Q} = \overline{H}_0^{-\frac{1}{2}} Q \overline{H}_0^{-\frac{1}{2}}$ and $\overline{Q} = \underline{H}_0^{-\frac{1}{2}} Q \underline{H}_0^{-\frac{1}{2}}$. Then, we have

$$\begin{aligned}
\frac{\partial}{\partial t} \kappa(tQ) &= \operatorname{tr}\left((\overline{H}_0 + tQ)^{-1} Q - (\overline{H}_0 + tQ)^{-1} Q \right) \\
&= \operatorname{tr}\left((I + t\underline{Q})^{-1} \underline{Q} - (I + t\overline{Q})^{-1} \overline{Q} \right) \leq 0,
\end{aligned}$$

where the last inequality follows from $\underline{Q} \preceq \overline{Q}$. This implies that

$$\ln \frac{\det \overline{H}(\lambda)}{\det \underline{H}(\lambda)} = \kappa\left(\sum_{j=1}^d \overline{H}_j \right) \leq \kappa(O) = \ln \frac{\det \overline{H}_0}{\det \underline{H}_0},$$

which, combined with (4.10), completes the proof. ■

The theorem implies that uLNML is a constant-gap upper bound of the LNML code length if f_X is strongly convex. Moreover, the gap bound (4.9) can be utilized for choosing a good neighbor U. Suppose that there is no effective boundary in the parameter space $\Omega = \Omega^{\mathrm{o}}$. Then, we can simplify the gap bound, and the optimal neighbor U is explicitly given.

Corollary 4 (Uniform gap bound for no-boundary case) Suppose that the assumptions made in Theorem 5 are satisfied. Then, if $\Omega = \Omega^{\mathrm{o}}$, we have a uniform gap bound

$$\bar{\mathcal{L}}(X|\lambda) - \mathcal{L}(X|\lambda) \le c_0 + \frac{1}{2}\ln\frac{\det \overline{H}_0}{\det \underline{H}_0} - \ln R(\overline{H}_0; U) \tag{4.11}$$

for all $X \in \mathcal{X}^n$ and all $\lambda \in \mathbb{A}$. This bound is minimized with the largest U, i.e., $U = \bigcap_{\theta\in\Omega}[\Omega_0 - \{\theta\}]$.

Proof According to Theorem 5, it suffices to show that $T(V) \equiv 1$ for all $V \subset \Omega_0$. From the existence of the RERM estimate in Ω, we have $\hat{\theta}(X, \lambda) \in \Omega = \Omega^{\mathrm{o}}$ for all $X \in \mathcal{X}^n$ and all $\lambda \in \mathbb{A}$. Therefore, we have $D(\lambda) = \mathcal{X}^n = D_\star$, and hence $\int_{D_\star} e^{-f_X(\psi)} d\nu(X) \equiv 1$ for all $\psi \in \Omega_0$, which is followed by $T(V) \equiv 1$. The second argument follows from the monotonicity of $R(H; \cdot)$. ■

As a remark, if we assume in addition that f_X is a smooth i.i.d. loss, i.e., $f_X = \sum_{i=1}^n f_{x_i}$ and $c_0 = 0$, the gap bound is also uniformly bounded with respect to the sample size n. This is derived from the fact that the right-hand side of (4.11) turns out to be

$$\ln\frac{\det n\overline{H}}{\det n\underline{H}} - \ln \mathbb{E}_{z\sim\mathcal{N}_m[\mathbf{0}, \frac{1}{n}\overline{H}_0^{-1}]}\left[\mathbb{1}U(z)e^{-r(\|z\|^2)}\right] \overset{n\to\infty}{\longrightarrow} \ln\frac{\det \overline{H}}{\det \underline{H}} < \infty.$$

4.2.5 Discussion

In previous sections, we derived an upper bound of the normalizing constant $Z(\lambda)$ and defined an easy-to-compute alternative for the LNML code length called uLNML. We also presented uniform gap bounds of uLNML for quadratic penalty functions. Note that uLNML characterizes $Z(\lambda)$ with upper smoothness of the loss and penalty functions. This is both advantageous and disadvantageous. The upper smoothness can often be easily computed even for complex models like deep neural networks. This means that uLNML is applicable to a wide range of loss functions. On the other hand, if the Hessian of the loss function drastically varies across Ω, the gap can be considerably large. In this case, one may tighten the gap by reparametrizing Ω to make the Hessian as uniform as possible.

The derivation of uLNML relies on the upper smoothness of the loss and penalty functions. In particular, our current analysis on the uniform gap guarantee given by Theorem 5 holds only if the penalty function is smooth. This is violated if one employs the ℓ_1-penalties.

We note that there exists an approximation of LNML called Rissanen's asymptotic expansion (RAE), which was originally given by Rissanen (1996) for a special case and then generalized by Grünwald (2007). RAE approximates LNML except for the $o(1)$ term with respect to n:

$$\mathcal{L}(X|\lambda) = \min_{\theta \in \Omega} f_X(\theta) + g(\theta, \lambda) + \frac{p}{2} \ln \frac{n}{2\pi} + \ln \int_{\Omega} \sqrt{\det I(\psi)} e^{-g(\psi,\lambda)} d\psi + o(1),$$

where $I(\psi) \stackrel{\text{def}}{=} \int \left[\nabla f_X(\theta) \nabla f_X(\theta)^\top\right] e^{-f_X(\theta)} d\nu(X)$ denotes the Fisher information matrix. Differences between RAE and uLNML are seen from two perspectives: their approximation errors and tractability.

As for the approximation errors, one of the largest differences is in their boundedness. RAE's $o(1)$ term is not necessarily uniformly bounded with respect to λ, and actually it can be unboundedly large for every fixed n as $\|\lambda\| \to \infty$ in the case of, for example, the Tikhonov regularization. This is in contrast to uLNML, in that the approximation gap is uniformly bounded with respect to λ according to Corollary 5, but it does not necessarily tend to zero as $n \to \infty$. This difference can be significant, especially in the scenario of penalty selection, where one compares different λ while n is fixed.

In terms of tractability, uLNML is usually easier to compute than RAE. One of the major obstacles when one computes RAE is that the integrand $\sqrt{\det I(\psi)} e^{-g(\psi,\lambda)}$ depends on both f_X and λ. Unless the analytic value of the integral is known, which is unlikely especially for complex models, one may employ the Monte Carlo approximation to evaluate it, which is typically computationally demanding for high-dimensional models. On the other hand, in uLNML, the unwieldy integral is replaced with the upper-smoothness term and the integral of the penalty. The upper smoothness can be computed with differentiation, which is often easier than integration, and the penalty integral does not depend on f_X anymore. Therefore, uLNML is often applicable to a wider class of models than RAE is. See Section 4.3.2 for an example.

4.3 Examples of uLNML

In the previous section, we have shown that the normalizing factor of LNML is bounded if the upper smoothness of $f_X(\theta)$ is bounded. The upper smoothness can be easily characterized for a wide range of loss functions. Since we cannot cover all of it here, we present a few examples that will be used in the experiments below.

4.3.1 Linear Regression

Let $X \in \mathbb{R}^{n \times m}$ be a fixed design matrix and let $y \in \mathbb{R}^n = \mathcal{X}^n$ represent the corresponding response variables. Then, we want to find $\beta \in \mathbb{R}^m$ such that $y \approx X\beta$. We assume that the 'useful' predictors may be sparse, and hence most of the coefficients of the best β for generalization may be close to zero. As such, we are motivated to solve the following ridge regression problem:

$$\min_{\sigma^2 \in [a,b],\ \beta \in \mathbb{R}^d} -\ln p(y|X, \beta, \sigma^2) + \frac{1}{2\sigma^2} \sum_{j=1}^{p} \lambda_j \beta_j^2, \tag{4.12}$$

where $-\ln p(y|X, \beta, \sigma^2) = \frac{1}{2\sigma^2} \|y - X\beta\|^2 + \frac{n}{2} \ln 2\pi\sigma^2$. According to Corollary 3, the uLNML of the ridge regression is given by

$$\begin{aligned}\bar{\mathcal{L}}(X|\lambda) = &\min_{\sigma^2 \in [a,b],\ \beta \in \mathbb{R}^d} -\ln p(y|X, \beta, \sigma^2) + \frac{1}{2\sigma^2} \sum_{j=1}^{p} \lambda_j \beta_j^2 \\ &+ \frac{1}{2} \ln \frac{\det(C + \operatorname{diag} \lambda)}{\det \operatorname{diag} \lambda} + \text{const.},\end{aligned}$$

where $C \overset{\text{def}}{=} X^\top X$. Note that the gap of the uLNML here is uniformly bounded, because the LNML of the variable temperature model (4.12) is bounded from below with that of fixed-variance models, which coincides with the above uLNML except with a constant.

4.3.2 Conditional Dependence Estimation

Let $X = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^{n\times m} = \mathcal{X}^n$ be a sequence of n observations independently drawn from the m-dimensional Gaussian distribution $\mathcal{N}_m[0, \Sigma]$. We assume that the conditional dependence among the m variables in X is scarce, which means that most of the coefficients of precision $\Theta = \Sigma^{-1} \in \mathbb{R}^{m\times m}$ are (close to) zero. Thus, to estimate the precision matrix Θ, we penalize the nonzero coefficients and consider the following RERM

$$\min_{\Theta\in\Omega} - \ln p(X|\Theta) + \frac{1}{2}\sum_{i\neq j} \lambda_{ij}\Theta_{ij}^2, \tag{4.13}$$

where $-\ln p(X|\Theta) = \frac{1}{2}\left\{\operatorname{tr} X^\top X\Theta - n\ln\det 2\pi\Theta\right\}$ denotes the probability density function of the Gaussian distribution. Here, we take $\Omega = \left\{\Theta \in \mathbb{S}^m_{++} \;\middle|\; \Theta \succeq R^{-1}I_m\right\}$ such that the Hessian is appropriately bounded: $\nabla^2_\Theta f_X = \Theta^{-1}\otimes\Theta^{-1} \preceq \overline{H}_0 = \frac{R^2}{2} I_{m\times m}$. As for the choice of R, we can use any upper-bound estimates of the largest eigenvalue of Θ^{-1}. Specifically, we employed $R = \left\|X^\top X/n\right\|_\infty$ in the experiments. One may include the code length of the hyperparameter itself, $\mathcal{L}(R)$, to make the code length complete. However, when we use any universal code length, including Rissanen (1983), the effect of the additional code length is at most $O(\ln R)$. Moreover, we can utilize the renormalization technique (Rissanen, 2000) to further reduce the dependency to $O(\ln\ln R)$. Hence we omit it for simplicity.

As it is an instance of the Tikhonov regularization, from Corollary 1 with $\overline{H}_0 = \frac{n}{2}R^2 I_{m^2}$, the uLNML for the graphical model is given by

$$\bar{\mathcal{L}}(X|\lambda) = \min_{\Theta\in\Omega} - \ln p(X|\Theta) + \frac{1}{2}\sum_{i\neq j}\left[\lambda_{ij}\Theta_{ij}^2 + \ln\left(1 + \frac{nR^2}{2\lambda_{ij}}\right)\right].$$

4.4 Minimization of uLNML

Given data $X \in \mathcal{X}^n$, we want to minimize uLNML (4.5) with respect to $\lambda \in \mathbb{A}$ as it bounds the LNML code length, which is a measure of the goodness of the penalty with respect to the MDL principle (Rissanen, 1978; Grünwald, 2007). Furthermore, it bounds the risk of the RERM estimate $\mathbb{E}_Y f_Y(\hat{\theta}(X,\lambda))$ Grünwald and Mehta (2017). The problem is that grid-search-like algorithms are inefficient since the dimensionality of the domain $\mathbb{A} \subset \mathbb{R}^d$ is high.

In order to solve this problem, we derive a CCCP for uLNML minimization. The algorithm is justified with the convergence properties that result from the CCCP framework. Then, we also give concrete examples of the computation required in the CCCP for typical RERMs.

4.4.1 CCCP for uLNML Minimization

In the forthcoming discussion, we assume that $\mathbb{A}$ is closed, bounded, and convex for computational convenience. We also assume that the upper bound of the normalizing factor $\ln\bar{Z}(\lambda)$ is convex with respect to λ. This is not a restrictive assumption as the true normalizing term $\ln Z(\lambda) = \ln\int\exp\left\{\max_{\theta\in\Omega} -f_X(\theta) - g(\theta,\lambda)\right\} d\nu(X)$ is always convex if the penalty is linear, as given in (4.2). In particular, it is actually convex for the Tikhonov regularization and LASSO, as in Corollary 1 and Corollary 2, respectively.

Recall that the objective function,uLNML, is written as

$$\bar{\mathcal{L}}(X|\lambda) = \min_{\theta\in\Omega} f_X(\theta) + g(\theta,\lambda) + \ln \bar{Z}(\lambda).$$

Therefore, the goal is to find a $\lambda^\star \in \mathbb{A}$ that attains

$$\bar{\mathcal{L}}(X|\lambda^\star) = \min_{\theta\in\Omega,\lambda\in\mathbb{A}} h_X(\theta,\lambda),$$

where $h_X(\theta,\lambda) \overset{\mathrm{def}}{=} f_X(\theta) + g(\theta,\lambda) + \ln \bar{Z}(\lambda)$. Note that the existence of $\lambda^\star$ follows from the continuity of the objective function $\bar{\mathcal{L}}(X|\lambda)$ and the closed nature of the domain $\mathbb{A}$.

The minimization problem can be solved by alternate minimization of h_X with respect to θ and λ as in Algorithm 2, which we call MDL regularization selection (MDL-RS). In general, minimization with respect to θ is the original RERM (4.1) itself. Thus, it can often be solved with existing software or libraries associated with the RERM problem. On the other hand, for minimization with respect to λ, we can employ standard convex optimization techniques since $h_X(\theta,\cdot)$ is convex as both $g(\theta,\cdot)$ and $\ln \bar{Z}(\cdot)$ are convex. Specifically, for some types of penalty functions, we can derive closed-form formulae of the update of λ. If one employs the Tikhonov regularization and $\overline{H}_0$ is diagonal, then

$$\frac{\partial}{\partial\lambda_j}\left[g(\theta_t,\lambda) + \bar{Z}(\lambda)\right] = 0 \Leftrightarrow \lambda = \frac{\overline{H}_{0,jj}}{2}\left[\sqrt{1+\frac{4}{\theta_{t,j}^2\overline{H}_{0,jj}}} - 1\right] \quad \left(= \tilde{\lambda}_{t,j}\right).$$

Therefore, if $\mathbb{A} = [a_1,b_1]\times\cdots\times[a_d,b_d]$, the convex part is completed by $\lambda_{t,j} = \Pi_{[a_j,b_j]}\tilde{\lambda}_{t,j}$, where $\Pi_{[a_j,b_j]}$ is the projection of the j-th coordinate. Similarly, if one employs LASSO,

$$\tilde{\lambda}_{t,j} = \sqrt[3]{\alpha + \sqrt{\alpha^2 + \left(\frac{\overline{H}_{0,jj}}{3}\right)^3}} + \sqrt[3]{\alpha - \sqrt{\alpha^2 + \left(\frac{\overline{H}_{0,jj}}{3}\right)^3}},$$

where $\alpha = \overline{H}_{0,jj}/\left|\theta_{t,j}\right|$. The projection procedure is the same as that for Tikhonov regularization.

Algorithm 2 MDL regularization selection (MDL-RS)

```
Input: $X \in \mathcal{X}^n$, $\lambda_0 \in \mathbb{A}$
$t \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    $\theta_t \leftarrow \mathrm{argmin}_{\theta\in\Omega} f_X(\theta) + g(\theta,\lambda_{t-1})$
    $\lambda_t \leftarrow \mathrm{argmin}_{\lambda\in\mathbb{A}} g(\theta_t,\lambda) + \ln \bar{Z}(\lambda)$
until stopping condition is met
return $\theta_t, \lambda_t$
```

The MDL-RS algorithm can be regarded as a special case of CCCP (Yuille and Rangarajan, 2003). First, the RERM objective is concave as it is the minimum of linear functions $F_X(\lambda) = \min_{\theta\in\Omega} f_X(\theta) + g(\theta,\lambda)$. Hence, uLNML is decomposed into the sum of concave and convex functions

$$\bar{\mathcal{L}}(X|\lambda) = F_X(\lambda) + \ln \bar{Z}(\lambda).$$

Secondly, $\tilde{F}_X^{(t)}(\lambda) \overset{\text{def}}{=} f_X(\theta_t) + g(\theta_t, \lambda)$ is a linear majorization function of $F_X(\lambda)$ at $\lambda = \lambda_{t-1}$, i.e., $\tilde{F}_X^{(t)}(\lambda) \geq F_X(\lambda)$ for all $\lambda \in \mathbb{A}$ and $\tilde{F}_X^{(t)}(\lambda_{t-1}) = F_X(\lambda_{t-1})$. Therefore, as we can write $\lambda_t = \operatorname{argmin}_{\lambda \in \mathbb{A}} \tilde{F}_X^{(t)}(\lambda) + \ln \bar{Z}(\lambda)$, MDL-RS is a concave-convex procedure for minimizing uLNML. The CCCP interpretation of MDL-RS immediately implies the following convergence arguments. Please refer to Yuille and Rangarajan (2003) for the proofs.

Theorem 6 (Monotonicity of MDL-RS) Let $\{\lambda_t\}_{t=0}^{\infty}$ be the sequence of solutions produced by Algorithm 2. Then, we have $\bar{\mathcal{L}}(X|\lambda_{t+1}) \leq \bar{\mathcal{L}}(X|\lambda_t)$ for all $t \geq 0$.

Theorem 7 (Local convergence of MDL-RS) Algorithm 2 converges to one of the stationary points of uLNML $\bar{\mathcal{L}}(X|\lambda)$.

Even if the concave part, i.e., minimization with respect to θ, cannot be solved exactly, MDL-RS still monotonically decreases uLNML as long as the concave part monotonically decreases the objective value $f_X(\theta_t) + g(\theta_t, \lambda_{t-1}) \leq f_X(\theta_{t-1}) + g(\theta_{t-1}, \lambda_{t-1})$ for all $t \geq 1$. This can be confirmed by seeing that $\bar{\mathcal{L}}(X|\lambda_t) = h_X(\theta_{t+1}, \lambda_t) \leq h_X(\theta_t, \lambda_t) \leq h_X(\theta_t, \lambda_{t-1}) = \bar{\mathcal{L}}(X|\lambda_{t-1})$, where $h_X(\theta, \lambda) = f_X(\theta) + g(\theta, \lambda)$. Moreover, if the subroutine of the concave part is iterative, early stopping may even be beneficial in terms of the computational cost.

4.4.2 Discussion

We previously introduced the CCCP algorithm for minimizing uLNML, namely MDL-RS. The monotonicity and local convergence property follow from the CCCP framework. One of the most prominent features of the MDL-RS algorithm is that the concave part is left completely black-boxed. Thus, it can be easily applied to the existing RERM.

There exists another approach for minimization of LNMLs in which a stochastic minimization algorithm is proposed (Miyaguchi et al., 2017). Instead of approximating the value of LNML, this directly approximates the gradient of LNML with respect to λ in a stochastic manner. However, since the algorithm relies on the stochastic gradient, there is no trivial way of judging if it converges or not. On the other hand, MDL-RS can exploit the information of the exact gradient of uLNML to stop the iteration.

Approximating LNML with uLNML benefits us more because can combine MDL-RS with grid search. Since MDL-RS could be trapped at fake minima, i.e., local minima and saddle points, starting from multiple initial points may be helpful to avoid poor fake minima and help it achieve lower uLNML.

4.5 Related Work

In the literature of model selection based on the MDL principle, there exist a number of methods that are concerned with discrete sets of candidate models (example.g., see Roos et al. (2009) and Hirai and Yamanishi (2011)). Note that in the problem of regularization selection, the candidate models are infinite in general and hence typical methods of the MDL model selection cannot be straightforwardly applied. Nevertheless, some of the RERM problems are addressed utilizing such methods. For example, the ℓ_0-norm RERM can be cast into the discrete model selection over all the subsets of features (e.g., see Dhillon et al. (2011) and Miyaguchi et al. (2017)). On the other hand, our method is applicable to arbitrary penalty functions as long as they are reasonably upper-smooth, although this is not the case with the ℓ_0-penalty. Therefore, our method can be regarded as a complement of the conventional discrete model selection.

As compared to existing methods of regularization selection, MDL-RS is distinguished by its efficiency in searching for penalties and its ease of systematic computation. Conventional penalty selection methods for large-dimensional models are roughly classified into three categories. Below, we briefly describe each one emphasizing its relationship and difference to the MDL-RS algorithm.

4.5.1 Grid Search with Discrete Model Selection Criteria

The first category is grid search with a discrete model selection criterion such as the cross validation score, Akaike's information criterion (AIC) (Akaike, 1974), or Bayesian information criterion (BIC) (Schwarz et al., 1978; Chen and Chen, 2008). In this method, we choose a model selection criterion and a candidate set of the hyperparameter $\{\lambda^{(k)}\}_{k=1}^{K} \in \mathbb{A} \subset \mathbb{R}^d$ in advance. Then, we calculate the RERM estimates for each candidate $\theta^{(k)} = \hat{\theta}(X, \lambda^{(k)})$. Finally, we select the best estimate according to the pre-determined criterion. This method is simple and universally applicable for any model selection criteria. However, the time complexity grows linearly as the number of candidates increases, and an appropriate configuration of the candidate set can vary corresponding to the data. This is specifically problematic for high dimensional design spaces, i.e., $d \gg 1$, where the combinatorial number of possible configurations is much larger than the feasible number of candidates.

On the other hand, the computational complexity of MDL-RS often scales better. Though it depends on the time complexity of the subroutine for the original RERM problem, the MDL-RS algorithm is not explicitly affected by the curse of dimensionality. However, it can be used for model selection in combination with the grid search. Although MDL-RS provides a more efficient approach to seeking a good λ in a (possibly) high-dimensional space as compared to simple grid search, it is useful to combine the two. Since uLNML is nonconvex in general, MDL-RS may converge to a fake minimum, such as local minima and saddle points, depending on the initial point λ_0. In this case, starting MDL-RS with multiple initial points $\lambda_0 = \lambda^{(k)}$ may improve the objective value.

4.5.2 Evidence Maximization

The second category is evidence maximization. In this methodology, one interprets the RERM as a Bayesian learning problem. The approach involves converting loss functions and penalty functions into conditional probability density functions $p(X|\theta) = e^{-f_X(\theta)}$ and prior density functions $p(\theta; \lambda) = e^{-g(\theta,\lambda)}(\int e^{-g(\psi,\lambda)} d\psi)^{-1}$, respectively. Then, the evidence is defined as $p(X; \lambda) = \int p(X|\theta)p(\theta; \lambda)d\theta$ and it is maximized with respect to λ. A successful example of evidence maximization is the relevance vector machine (RVM) proposed by Tipping (2001). It is a Bayesian interpretation of ridge regression with different penalty weights λ_j on different coefficients, as described in Corollary 1. This results in so-called automatic relevance determination (ARD) and makes the approach applicable to redundant models.

The maximization of the evidence can also be thought of as an instance of the MDL principle, as it is equivalent to minimizing $-\ln p(X; \lambda)$ with respect to λ, which is a codelength function of X. Moreover, LNML and the evidence each have an intractable integral in them. A notable difference between the two is the computational cost to optimize them. Though LNML contains an intractable integral in its normalizing term $\ln Z(\lambda)$, it can be systematically approximated by uLNML, and uLNML is efficiently minimized via CCCP. On the other hand, in the case of evidence, we do not know of any approximation that is as easy to optimize and as systematic as uLNML. Although a number of approximations have been developed for evidence, such as the Laplace approximation, variational Bayesian inference (VB), and Markov chain Monte Carlo sampling (MCMC), these tend to be combined with grid search (e.g., Yuan and Lin (2005)), except for some special cases such as the RVM and Gaussian processes (Rasmussen and Williams, 2006).

4.5.3 Error Bound Minimization

The last category is error bound minimization. The generalization capability of RERM has been extensively studied in bounding generalization errors, specifically on the basis of the PAC learning theory (Valiant, 1984) and PAC-Bayes theory (Shawe-Taylor and Williamson, 1997; McAllester, 1999). There also exist a considerable number of studies that relate error bounds with the MDL principle, including (but not limited to) Barron and Cover (1991), Yamanishi (1992), and Chatterjee and Barron (2014). One might determine the hyperparamter λ by minimizing the error bound. However, being used to prove the learnability of new models, such error bounds are often not used in practice more than the cross validation score. MDL-RS can be regarded as an instance of the minimization of an error bound. Actually, uLNML bounds the LNML code length, which was recently shown to be bounding the generalization error of the RERM estimate under some conditions including boundedness of the loss function and fidelity of hypothesis classes (Grünwald and Mehta, 2017).

4.6 Experiments

In this section, we empirically investigate the performance of the MDL-RS algorithm.[*2] We compare MDL-RS with conventional methods, applying the two models introduced in Section 4.3 on both synthetic and benchmark datasets.

We employ two models, namely linear regression with Gaussian noise and conditional dependency estimation. For each model, a comparison is conducted from two perspectives: First, as a preliminary experiment, we check if the MDL-RS algorithm can actually minimize uLNML. Secondly, we evaluate the generalization performance of the MDL-RS algorithm. In particular, because we expect that the proposed method performs better than the other methods if the model is high-dimensional, we focus on the dependency of the performance on the dimensionality.

4.6.1 Linear Regression

Setting. For the linear regression, we compared MDL-RS with ARD regression with RVM Tipping (2001) and deterministic and random grid search methods for the cross validation score. As for the deterministic cross validation, we employ ridge regression and LASSO with penalty weights of 20 points spread logarithmically evenly over $\lambda_j = \lambda \in [10^{-4}, 10^0]$ $(j = 1, \dots, p)$. As for the random cross validation, 100 random points are drawn from the log-uniform distribution over $[10^{-4}, 10^0]^p$. The performance metric is test logarithmic loss (log-loss) $-\ln p(y|X, \beta, \sigma^2)$ (see Section 4.3.1) on 5-fold cross validation. Figure 4.2 shows the results of the comparison with six datasets, namely three synthetic datasets and three real-world dataset. In the synthetic datasets, the number of design variables m varies from 5 to 100, and the true coefficients β are randomly chosen with some set to zero. The real-world examples are taken from the Diabetes dataset[*3],

[*2] The source codes and datasets of the following experiments are available at `https://github.com/koheimiya/pymdlrs`.

[*3] http://www4.stat.ncsu.edu/ boos/var.select/diabetes.html

ResidentialBuilding dataset Rafiei and Adeli (2015)[*4], and YearPredictionMSD dataset[*5]. We note that there is huge variety in the dimensionality of parameter spaces: $m = 14$ for Diabetes, $m = 90$ for YearPredictionMSD, and $m = 103$ for ResidentialBuilding dataset.

Result. Figure 4.1 shows the results of the preliminary experiments with linear regression. It is shown there that the MDL-RS algorithm successfully minimizes uLNML compared to the other non-MDL algorithms.

As for the generalization errors, from the overall results, we can see that MDL-RS and RVM are comparable to one another and outperform the other three. Figure 4.2a, Figure 4.2b, and Figure 4.2c suggest that the proposed method performs well in all the synthetic experiments. Figure 4.2d, Figure 4.2e, and Figure 4.2f show the results of the real-world experiments. One can observe the same tendency as in the synthetic ones; Both MDL-RS and RVM outperform the rest in terms of generalization error (log-loss) and the difference is bigger when the sample size is smaller. However, note that in the YearPredictionMSD dataset, RVM converges to a poor local minima and hence fails to lower the log-loss well, even with large training samples. It is also noteworthy that the performance of the random grid-search method becomes poor and unstable for the high-dimensional cases, which is due to the curse of dimensionality of the design space. These results emphasize the efficiency of MDL-RS in optimizing uLNML with high-dimensional models.

4.6.2 Conditional Dependence Estimation

Setting. For the estimation of conditional dependencies, we compared MDL-RS with the grid search of LASSO Friedman et al. (2008) with AIC, (extended) BIC, and the cross validation score. We generated data $X \in \mathbb{R}^{n\times m}$ from m-dimensional double-ring Gaussian graphical models ($m = 10, 20, 50, 100$) in which each variable $j \in [1, m]$ is conditionally dependent to its 2-neighbors $j-2, j-1, j+1$, and $j+2 \pmod m$ with a coefficient of 0.25. Note that MDL-RS can be applied to the graphical model just by computing the upper smoothness, while RVM cannot be applied straightforwardly.

Result. Figure 4.3 shows the results of the preliminary experiments for conditional dependency estimation. It is shown that the MDL-RS algorithm is the best for minimizing uLNML as we expected, especially when the dimensionality d is relatively large compared to the sample size n. We note that these results are a natural consequence of our experimental design, as the other methods are not designed for minimizing uLNML. We have simply confirmed that the MDL-RS algorithm works well in accordance with our intention.

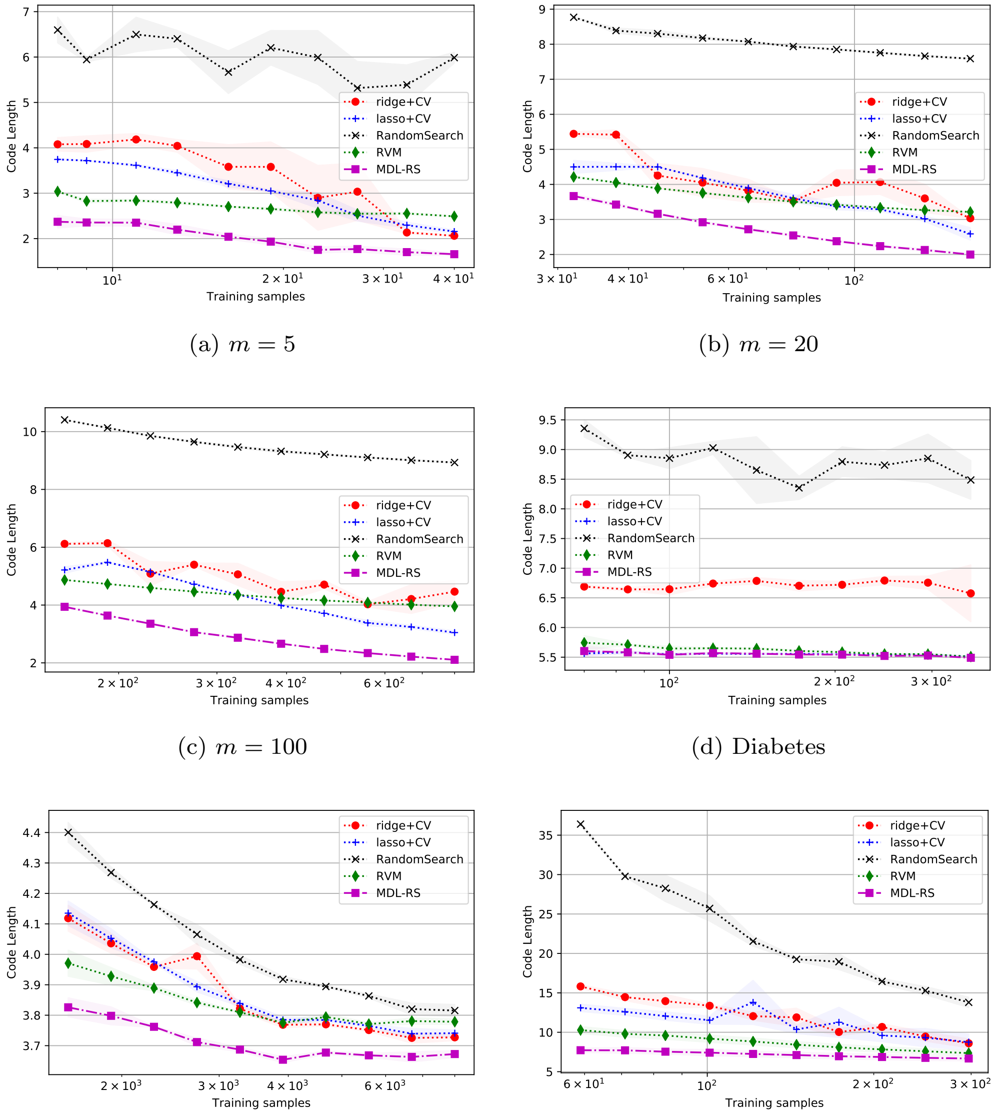
Figure 4.4 shows the results of the experiment on the generalization errors. It is seen that all the estimators converge at the same rate of $O(n^{-1})$, whereas MDL-RS gives the least Kullback–Leibler divergence by far, specifically with large m. In particular, when $m = 100$, the proposed estimator outperforms the others by more than a factor of five. This supports our claim that penalty selection in high-dimensional design spaces has a considerable effect on generalization capability when the model is redundant.

4.6.3 Discussion

Both results indicate that MDL-RS performs well, specifically when the model is high-dimensional, as expected. Note that the generalization error LNML and uLNML bound is

[*4] https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set

[*5] https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd



(a) $m = 5$

(b) $m = 20$

(c) $m = 100$

(d) Diabetes

(e) YearPredictionMSD

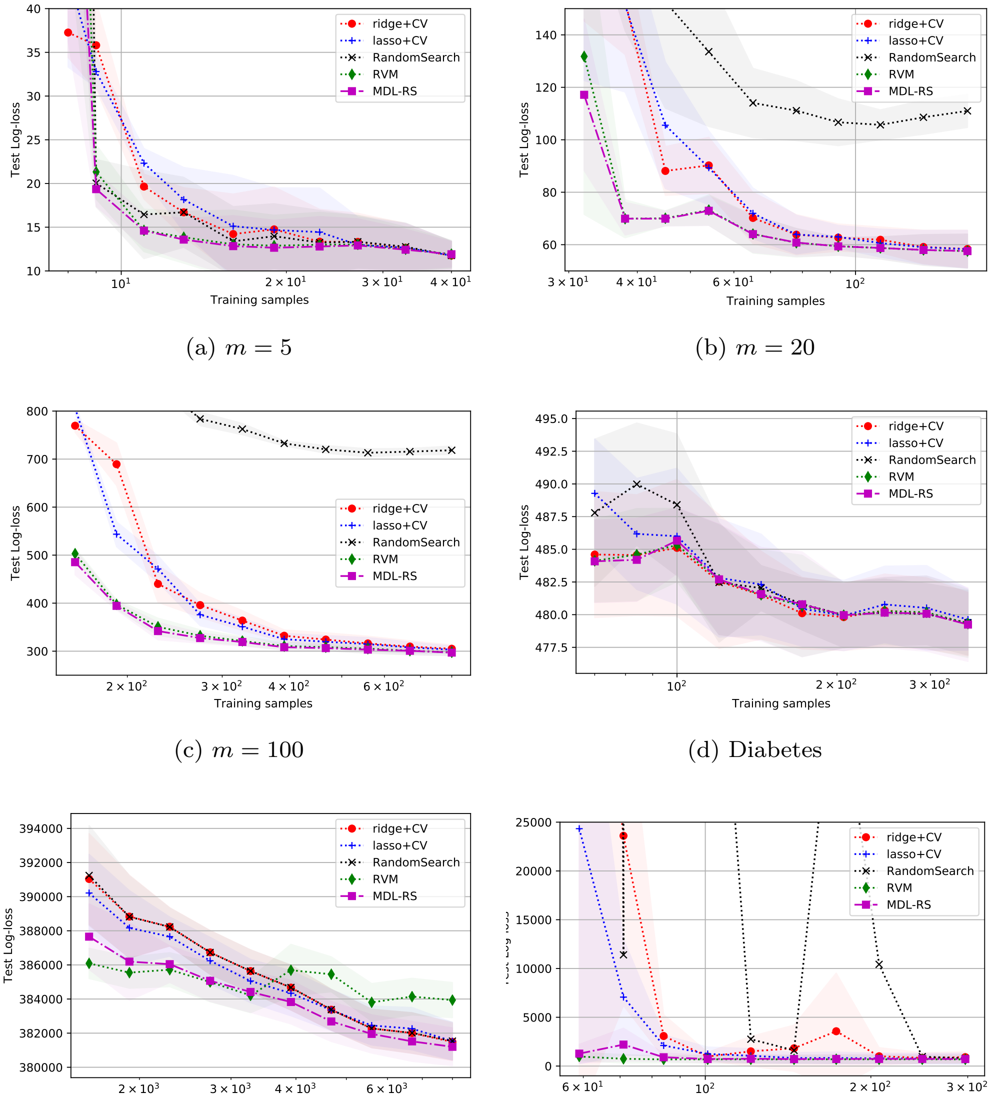
(f) ResidentialBuilding

Fig. 4.1: **Convergence of normalized uLNML in linear regression**
The horizontal axes show the number of training samples in logarithmic scale, while the vertical axes show uLNML per sample. Each shaded area shows ±one-standard deviation.

the expected logarithmic loss $\mathbb{E}_{X,Y} f_Y(\hat{\theta}(X,\lambda))$, and the performance metric we employed in the experiments is (an unbiased estimator of) the logarithmic loss itself. Hence, if the metric is changed, the result could be different.

4.7 Concluding Remarks

In this chapter, we proposed a new method for penalty selection on the basis of the MDL principle. Our main contribution was the introduction of uLNML, which is a tight upper bound of LNML for smooth RERM problems. This can be analytically computed, except for a constant, given the (upper) smoothness of the loss and penalty functions. We also



(a) $m = 5$

(b) $m = 20$

(c) $m = 100$

(d) Diabetes

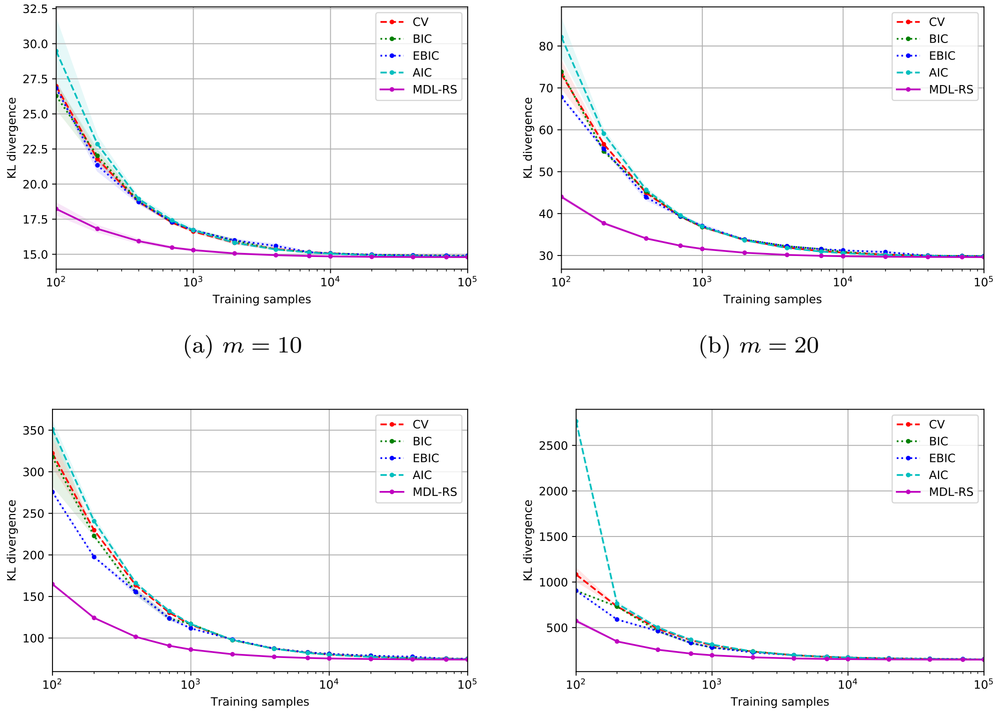
(e) YearPredictionMSD

(f) ResidentialBuilding

Fig. 4.2: **Convergence of log-loss in linear regression**
The horizontal axes show the number of training samples in logarithmic scale, while the vertical axes show the test log-loss. Each shaded area shows ±one-standard deviation.

presented the MDL-RS algorithm, which is a minimization algorithm of uLNML with convergence guarantees. Experimental results indicated that MDL-RS's generalization capability was comparable to that of conventional methods. In the high-dimensional setting we are interested in, it even outperformed conventional methods.

In related future work, further applications to various models such as latent variable models and deep learning models must be analyzed. As the above models are not (strongly) convex, the extension of the lower bound of LNML to the non-convex case would also be an interesting topic of future study. While we bounded LNML with the language of Euclidean spaces, the only essential requirement of our analysis is upper smoothness of loss functions defined over parameter spaces. Therefore, we believe that it is possible to generalize uLNML to Hilbert spaces to deal with infinite-dimensional models.



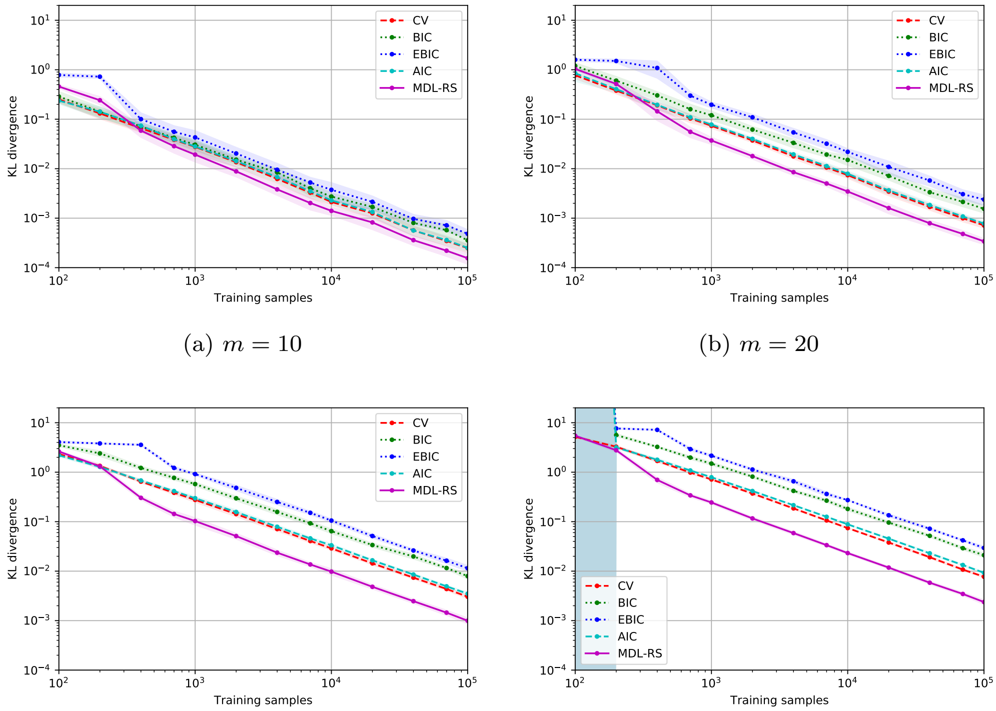
(a) $m = 10$

(b) $m = 20$

(c) $m = 50$

(d) $m = 100$

Fig. 4.3: **Convergence of normalized uLNML for graphical models**
The horizontal axes show the number of training samples in logarithmic scale, while the vertical axes show the uLNML per sample associated with penalty weight λ. Each shaded area shows ±one-standard deviation.



(a) $m = 10$

(b) $m = 20$

(c) $m = 50$

(d) $m = 100$

Fig. 4.4: **Convergence of Kullback–Leibler divergence for graphical models** The horizontal axes show the number of training samples in logarithmic scale, while the vertical axes show the divergence of estimates relative to true distributions. Each shaded area shows ±one-standard deviation.

Chapter 5

Minimax Regret for Smooth Logarithmic Losses over High-Dimensional ℓ_1-Balls

We develop a new theoretical framework called *envelope complexity* to analyze the minimax regret with logarithmic loss functions and derive a Bayesian predictor that adaptively achieves 2-approximate minimax regret over high-dimensional ℓ_1-balls. The prior is newly derived for achieving the minimax regret and is called the *spike-and-tails (ST) prior*, named after its look. The resulting regret bound is so simple that it is completely determined with the smoothness of the loss function and the radius of the balls except with logarithmic factors, and it has a generalized form of existing regret/risk bounds. In the preliminary experiment, we confirm that the ST prior outperforms the conventional minimax-regret prior under non-high-dimensional asymptotics[*1].

5.1 Motivation

As a notion of complexity of predictive models (sets of predictors), *minimax regret* has been considered in the literature of online learning (Cesa-Bianchi and Lugosi, 2006) and the MDL principle (Rissanen, 1978; Grünwald, 2007). The minimax regret of a model $\mathcal{H}$ is given by

$$\mathrm{REG}^\star(\mathcal{H}) = \inf_{\hat{h}\in\hat{\mathcal{H}}} \sup_{X\in\mathcal{X}} \left\{ f_X(\hat{h}) - \inf_{h\in\mathcal{H}} f_X(h) \right\}, \tag{5.1}$$

where $f_X(h)$ denotes the loss of the prediction over data X made by h, $\hat{\mathcal{H}}$ denotes the feasible predictions, and $\mathcal{X}$ is the space of data. Here, the data may consist of a sequence of datum $X = X^n = (X_1, \dots, X_n)$, and the loss maybe additive ($f_X(h) = \sum_{i=1}^n f_{X_i}(h)$), but we keep them implicit for generality. The minimax regret is a general complexity measure in the sense that it is defined without any assumptions on the generation process of X. For instance, one can bound statistical risks with $\mathrm{REG}^\star(\mathcal{H})$ regardless of the distribution of data (Littlestone, 1989; Cesa-Bianchi et al., 2004; Cesa-Bianchi and Gentile, 2008). Therefore, bounding the minimax regret and constructing the corresponding predictor $\hat{h}$ is important to make a good and robust prediction.

We consider that $\mathcal{H}$ is parametrized by a real-valued vector $\theta \in \mathbb{R}^d$: $\mathcal{H} = \{h_\theta \mid \gamma(\theta) \le B,\ \theta \in \mathbb{R}^d\}$, where $\gamma(\theta)$ denotes a radius function such as norms of θ.

[*1] The contents of this chapter was published in Miyaguchi and Yamanishi (2018a).

Thus, we may consider the luckiness minimax regret (Grünwald, 2007)

$$\mathrm{LREG}^{\star}(\gamma) = \inf_{\hat{h}\in\hat{\mathcal{H}}} \sup_{X\in\mathcal{X}} \left\{ f_X(\hat{h}) - \inf_{\theta\in\mathbb{R}^d} \left[f_X(\theta) + \gamma(\theta) \right] \right\} \tag{5.2}$$

instead of the original minimax regret. Through abuse of notation, we say $f_X(\theta) = f_X(h_\theta)$. There are at least three reasons for adopting this formulation. First, as we do not assume the underlying distribution of X, it may be plausible to pose a soft restriction as in (5.2) rather than the hard restriction in (5.1). Secondly, it is straightforwardly shown that the luckiness minimax regret bounds above the minimax regret. Thus, it is often sufficient to bound $\mathrm{LREG}^{\star}(\gamma)$ for bounding $\mathrm{REG}^{\star}(\mathcal{H})$. Finally, the luckiness minimax regret is including the original minimax regret as a special case such that $\gamma(\theta) = 0$ if $\theta \in \mathcal{H}$, and $\gamma(\theta) = \infty$ otherwise. Therefore, we may avoid possible computational difficulties of the minimax regret by choosing the penalty γ carefully.

That being said, the closed-form expression of the exact (luckiness) minimax regret is intractable except with few special cases (e.g., Shtarkov (1987); Koolen et al. (2014)).

However, if we focus on information-theoretic settings, i.e., the model $\mathcal{H}$ is a set of probabilistic distributions, everything becomes explicit. Now, let predictors be sub-probability distributions $P(\cdot \mid \theta)$, and adopt the logarithmic loss function $f_X(\theta) = -\ln \frac{\mathrm{d}P}{\mathrm{d}\nu}(X|\theta)$ with respect to an appropriate base measure ν, such as counting or Lebesgue measures. Note that a number of important practical problems, such as logistic regression and data compression, can be handled with this framework. With the logarithmic loss, the closed form of the luckiness minimax regret is given by Shtarkov (1987); Grünwald (2007) as

$$\mathrm{LREG}^{\star}(\gamma) = \ln \int e^{-m(f_X+\gamma)} \nu(\mathrm{d}X) \overset{\mathrm{def}}{=} S(\gamma), \tag{5.3}$$

where m denotes the minimum operator given by $m(f) = \inf_{\theta\in\mathbb{R}^d} f(\theta)$. We refer to the left-hand-side value as the *Shtarkov complexity*. Moreover, when all the distributions in $\mathcal{H}$ are i.i.d. regular distributions of n-sequences $X = (X_1, \ldots, X_n)$, under some regularity conditions, the celebrated asymptotic formula (Rissanen, 1996; Grünwald, 2007) is given by

$$S(\gamma) = \frac{d}{2} \ln \frac{n}{2\pi} + \int \sqrt{\det I(\theta)} e^{-\gamma(\theta)} \mathrm{d}\theta + o(1), \tag{5.4}$$

where $I(\theta)$ is the Fisher information matrix, and $o(1) \to 0$ as $n \to \infty$. More importantly, although the exact minimax-regret predictor achieving $S(\gamma)$ is still intractable, the asymptotic formula implies that it is asymptotically achieved with the Bayesian predictor associated with the *tilted Jeffreys prior* $\pi(\mathrm{d}\theta) \propto \sqrt{\det I(\theta)} e^{-\gamma(\theta)} \mathrm{d}\theta$.

Here, our research questions are as follows: First, **(Q1)** *How can we evaluate $S(\gamma)$ in modern high-dimensional contexts?* In particular, the asymptotic formula (5.4) does not withstand high-dimensional learning problems where d increases as $n \to \infty$. The exact evaluation of the Shtarkov complexity (5.3) on the other hand is often intractable due to the minimum operator inside the integral. Secondly, **(Q2)** *How can we achieve the minimax regret with computationally feasible predictors?* It is important to provide the counterpart of the tilted Jeffreys prior in order to make actual predictions.

Regarding the above questions, our contributions are summarized as follows:

- We introduce *envelope complexity*, which is a non-asymptotic approximation of the Shtarkov complexity $S(\gamma)$ that allows us systematic computation of its upper bounds and predictors achieving these bounds. In particular, we show that the regret of the predictor is characterized with the smoothness.

- We demonstrate its usefulness by giving a Bayesian predictor that adaptively achieves the minimax regret within a factor of two over any high-dimensional smooth models under ℓ_1-constraints $\|\theta\|_1 \leq B$.

The remainder of this chapter is organized as follows: In Section 5.2, we introduce the notion of Bayesian minimax regret as an approximation of the minimax regret within the 'feasible' set of predictors. We then develop a complexity measure called *envelope complexity* in Section 5.3 as a mathematical abstraction of the Bayesian minimax regret. We also present a collection of techniques for bounding the envelope complexity to the Shtarkov complexity. In Section 5.4, we utilize the envelope complexity to construct a near-minimax Bayesian predictor under ℓ_1-penalization, namely the ST prior. We also show that it achieves the minimax rate over $\mathcal{H} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq B\}$ under high-dimensional asymptotics. In Section 5.5, we demonstrate numerical experiments to visualize our theoretical results. Discussion on these results in comparison to existing results is given in Section 5.6. Finally, we conclude the chapter in Section 5.7.

5.2 Bayesian Minimax Regret

The minimax regret with logarithmic loss is given by the Shtarkov complexity $S(\gamma)$. The computation of the Shtarkov complexity $S(\gamma)$ is often intractable if we consider practical models such as deep neural networks. This is because the landscapes of loss functions $f \in \mathcal{F}$ are as complex as the models are, so their minimums $m(f)$ and the complexity, which is an integral over the function of $m(f)$, are not tractable. Moreover, computations of the optimal predictor $h^\star$ are still often intractable, even if $S(\gamma)$ are given. For instance, the minimax-regret prediction for Bernoulli models over n results in a time cost of $O(n2^n)$. Clearly there exist some special cases for which closed forms of $\hat{h}$ are given. However, so far, they are limited to exponential families.

One cause of this issue is that we seek for the best predictor $\hat{h}$ among all the possible predictors $\hat{\mathcal{H}}$, i.e., all probability distributions. This is too general, so it is potentially impossible to compute $\hat{h}$ and $\mathrm{REG}^\star(\gamma)$. To avoid this difficulty, we narrow the set of feasible predictors $\hat{\mathcal{H}}$ to the Bayesian predictors. Let $w \in \mathcal{M}_+(\mathbb{R}^d)$ be a positive measure over $\mathbb{R}^d$, which we may refer to as *pre-prior*, and let $\overline{h}_w$ be the Bayesian predictor associated with the prior $\pi(\mathrm{d}\theta) \propto e^{-\gamma(\theta)} w(\mathrm{d}\theta)$. Then, we have

$$f_X(\overline{h}_w) = \ln \frac{w\left[e^{-\gamma}\right]}{w\left[e^{-f_X-\gamma}\right]} \overset{\text{def}}{=} f_X(w), \tag{5.5}$$

where $w\left[\cdot\right]$ denotes the integral operation with respect to $w(\mathrm{d}\theta)$. Now, we consider the Bayesian (luckiness) minimax regret given by

$$\begin{aligned} \mathrm{LREG}^{\mathrm{Bayes}}(\gamma) &\overset{\text{def}}{=} \inf_{w \in \mathcal{M}_+(\mathbb{R}^d)} \mathrm{LREG}(w|\gamma), \\ \mathrm{LREG}(w|\gamma) &\overset{\text{def}}{=} \sup_{X \in \mathcal{X}} \left\{ f_X(w) - m\left(f_X + \gamma\right) \right\}. \end{aligned}$$

One advantage of considering the Bayesian minimax regret is that given a measure w, one can compute $\overline{h}_w$ analytically or numerically utilizing techniques developed in the literature of Bayesian inference. In particular, a number of sophisticated variants of Monte Carlo Markov chain (MCMC) methods, such as stochastic gradient Langevin Dynamics (Welling and Teh, 2011), have been developed for sampling θ from complex posteriors.

Note that there does exist a case where the Bayesian minimax regret strictly differs from the minimax regret. The following example is taken from Barron et al. (2014).

Example 1 (Existence of gap) Consider a single observation of a ternary variable $X \in \{1,2,3\}$ under a model consisting of three predictors with no penalty, h_1, h_2, h_3, defined as follows.

$$P(\cdot|h_1) = \left(\frac{1}{2}, \frac{1}{2}, 0\right), \; P(\cdot|h_2) = \left(0, \frac{1}{2}, \frac{1}{2}\right), \; P(\cdot|h_3) = \left(\frac{2}{7}, \frac{3}{7}, \frac{2}{7}\right).$$

Let $\gamma(j) = 0$ for all $j = 1,2,3$ and $\gamma(\theta) = \infty$ otherwise. The maximum likelihood values are given by $1/2$ for all X since the maximum of $1/2$ is achieved by either h_1 or h_2, the minimax distribution is the uniform distribution $P(X|\hat{h}) = 1/3$ $(X = 1,2,3)$.

Now, assume that the minimax distribution is a Bayesian predictor with pre-prior w: $\hat{h} = h_{\gamma,w}$. Then, we have $P(X|h_{\gamma,w}) = \sum_{j=1}^{3} w\left[P(X|h_j)\right] = 1/3$ for all X. It is only satisfied if $w \propto (-1,-1,7/2)$, which cannot be a proper pre-prior. This yields a contradiction, and therefore by the continuity of the regret with respect to w, it implies that $\mathrm{REG}^\star(\gamma) < \mathrm{LREG}^{\mathrm{Bayes}}(\gamma)$.

It implies that narrowing the range of predictors to Bayesian may worsen the achievable worst-case regret. However, as we will show shortly, the gap between these minimax regrets can be controlled with model γ.

5.3 Envelope Complexity

We have introduced the Bayesian minimax regret $\mathrm{LREG}^{\mathrm{Bayes}}(\gamma)$. In this section, we present a set representation of Bayesian minimax regret, namely the *envelope complexity* $C(\gamma, \mathcal{F})$. Then, we show that the Shtarkov complexity is bounded by the envelope complexity, which can be easily bounded even if the models are complex.

5.3.1 Set Representation of Bayesian Minimax Regret

The envelope complexity is a simple mathematical abstraction of Bayesian minimax regret and gives a fundamental basis for systematic computation of upper bounds on the (Bayesian) minimax regret. Let $\mathcal{F}$ be a set of continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ that is not necessarily logarithmic. Define the Bayesian envelope of $\mathcal{F}$ as

$$\mathcal{E}(\mathcal{F}) \overset{\mathrm{def}}{=} \left\{ w \in \mathcal{M}_+(\mathbb{R}^d) \;\middle|\; \forall f \in \mathcal{F}, w\left[e^{-f+m(f)}\right] \geq 1 \right\},$$

and define the envelope complexity as

$$C(\gamma, \mathcal{F}) \overset{\mathrm{def}}{=} \inf_{w \in \mathcal{E}(\mathcal{F})} \ln w\left[e^{-\gamma}\right].$$

Then, the envelope complexity characterizes Bayesian minimax regret.

Theorem 8 (Set representation) Let $\mathcal{F} = \{f_X + \gamma \mid X \in \mathcal{X}\}$. Then, all measures in the envelope $w \in \mathcal{E}(\mathcal{F})$ satisfy that

$$\mathrm{LREG}(w|\gamma) \leq \ln w\left[e^{-\gamma}\right].$$

Moreover, we have

$$\mathrm{LREG}^{\mathrm{Bayes}}(\gamma) = C\left(\gamma, \mathcal{F}\right).$$

Proof Let $c(w) = \inf_{f \in \mathcal{F}} w[e^{-f+m(f)}]$. Observe that

$$\begin{aligned}
\ln \frac{w\left[e^{-\gamma}\right]}{c(w)} &= \sup_{f \in \mathcal{F}} \left\{ \ln \frac{w\left[e^{-\gamma}\right]}{w\left[e^{-f}\right]} - m(f) \right\} \\
&= \sup_{X \in \mathcal{X}} \left\{ \ln \frac{w\left[e^{-\gamma}\right]}{w\left[e^{-f_X-\gamma}\right]} - m(f_X+\gamma) \right\} \\
&\quad (f = f_X + \gamma) \\
&= \mathrm{LREG}(w|\gamma). \\
&\quad (\because\ (5.5))
\end{aligned}$$

Then, since $c(w) \geq 1$ for all $w \in \mathcal{E}(\mathcal{F})$, we have the inequality.

Note that $\bar{w} = w/c(w) \in \mathcal{E}(\mathcal{F})$ for any $w \in \mathcal{M}_+(\mathbb{R}^d)$, and $\bar{w}\left[e^{-\gamma}\right] \leq w\left[e^{-\gamma}\right]$ whenever $w \in \mathcal{E}(\mathcal{F})$. Then, we have

$$\begin{aligned}
C(\gamma, \mathcal{F}) &= \inf_{w \in \mathcal{M}_+(\mathbb{R}^d)} \ln \frac{w\left[e^{-\gamma}\right]}{c(w)} \\
&= \inf_{w \in \mathcal{M}_+(\mathbb{R}^d)} \mathrm{LREG}(w|\gamma) && \text{(the above equality)} \\
&= \mathrm{LREG}^{\mathrm{Bayes}}(\gamma),
\end{aligned}$$

yielding the equality. This completes the proof. ■

We have seen that the envelope complexity is equivalent to the Bayesian minimax regret. Below, we present upper bounds of the Shtarkov complexity that the remainder of this chapter is based.

Theorem 9 (Bounds on Shtarkov complexity) Let $\mathcal{F} = \{f_X + \gamma \mid X \in \mathcal{X}\}$, where f_X is logarithmic. Then, for all $w \in \mathcal{E}(\mathcal{F})$, we have

$$S(\gamma) \leq C(\gamma, \mathcal{F}) \leq \ln w\left[e^{-\gamma}\right].$$

Proof The first inequality follows from the fact that the envelope minimax regret is no less than the minimax regret, as the range of infimum is shrunk from $\hat{\mathcal{H}}$ to the Bayes class $\left\{\overline{h}_w\right\}$. The second inequality follows from the definition of the envelope complexity. This completes the proof. ■

5.3.2 Useful Lemmas for Evaluating Envelope Complexity

Next, we show several lemmas that highlight the computational advantage of the envelope complexity. We start to show that the envelope complexity is easily evaluated with the surrogate relation. We say a function g is *surrogate* of another function f if and only if $f - m(f) \leq g - m(g)$, which is denoted by $f \preceq g$. Moreover, if there is one-to-one correspondence between $g \in \mathcal{G}$ and $f \in \mathcal{F}$ such that $f \preceq g$, then we may write $\mathcal{F} \preceq \mathcal{G}$.

Lemma 10 (Monotonicity) Let $\mathcal{F} \preceq \mathcal{G}' \subset \mathcal{G}$. Then, we have

$$.\mathcal{E}(\mathcal{F}) \supset \mathcal{E}(\mathcal{G}),$$

and therefore

$$C(\gamma, \mathcal{F}) \leq C(\gamma, \mathcal{G}).$$

Proof Note that $e^{-f+m(f)} \geq e^{-g+m(g)}$ if $f \preceq g$, which means $\mathcal{E}(\mathcal{F}) \supset \mathcal{E}(\mathcal{G}')$. Also, as increasing the argument from $\mathcal{G}'$ to $\mathcal{G}$ only strengthens the predicate of the envelope, we have $\mathcal{E}(\mathcal{G}') \supset \mathcal{E}(\mathcal{G})$. Therefore, we have

$$\begin{aligned} C(\gamma, \mathcal{F}) &= \inf_{w \in \mathcal{E}(\mathcal{F})} \ln w \left[e^{-\gamma}\right] \\ &\leq \inf_{w \in \mathcal{E}(\mathcal{G}')} \ln w \left[e^{-\gamma}\right] && \mathcal{E}(\mathcal{F}) \supset \mathcal{E}(\mathcal{G}') \\ &\leq \inf_{w \in \mathcal{E}(\mathcal{G})} \ln w \left[e^{-\gamma}\right] && \mathcal{E}(\mathcal{G}') \supset \mathcal{E}(\mathcal{G}) \\ &= C(\gamma, \mathcal{G}). \end{aligned}$$

■

This is especially useful when the loss functions $\mathcal{F}$ are complex but there exist simple surrogates $\mathcal{G}$. Consider any models such that the landscapes of the associated loss functions $f \in \mathcal{F}$ are not fully understood and the evaluation of $m(f)$ is expensive. It is impossible to check if w is in the envelope $w \in \mathcal{E}(\mathcal{F})$, and therefore Theorem 9 cannot be used directly. However, even in such cases, one can possibly find a surrogate class $\mathcal{G}$ of $\mathcal{F}$. If the surrogate $\mathcal{G}$ is simple enough for checking if $w \in \mathcal{E}(\mathcal{G})$, it is possible to bound the envelope complexity utilizing Lemma 10 and Theorem 9.

In the following, we consider the specific instance of the surrogate relation based on the smoothness. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *L-upper smooth* if and only if, for all $\theta, \theta_0 \in \mathbb{R}^d$, there exists $g \in \mathbb{R}^d$ such that

$$f(\theta) \leq f(\theta_0) + g^\top(\theta - \theta_0) + \frac{L}{2} \|\theta - \theta_0\|_2^2 . \tag{5.6}$$

Note that the upper smoothness is weaker than (Lipschitz) smoothness. Thus, if f is L-upper smooth and has at least one minima $\theta_0 \in \arg m(f)$, we can construct a simple quadratic surrogate of f: $\theta \mapsto \frac{L}{2} \|\theta - \theta_0\|_2^2$ $(\succeq f)$.

Motivated by the smoothness assumption, below we present more specific bounds for quadratic functions. Let $\mathcal{Q}$ be the set of all quadratic functions with curvature one, defined as $\mathcal{Q} = \{\theta \mapsto \frac{1}{2} \|\theta - u\|^2 \mid u \in \mathbb{R}^d\}$. Moreover, for all sets of loss functions $\mathcal{F}$ and penalty functions $\gamma : \mathbb{R} \to \overline{\mathbb{R}}$, we write $\mathcal{F}_\gamma = \mathcal{F} + \gamma = \{f + \gamma \mid f \in \mathcal{F}\}$. Then, the envelope complexity of $\mathcal{F}_\gamma$ is evaluated with that of $\mathcal{Q}_\gamma$.

Lemma 11 (Bounds of smoothness) Suppose that all $f \in \mathcal{F}$ are L-upper smooth. Let $\varphi(\theta) = \sqrt{L}^{-1}\theta$ be the scaling function. Then, we have

$$\mathcal{E}(\mathcal{Q}_{\gamma \circ \varphi}) \circ \varphi^{-1} \subset \mathcal{E}(\mathcal{F}_\gamma),$$

and moreover,

$$C(\gamma, \mathcal{F}_\gamma) \leq C(\gamma \circ \varphi, \mathcal{Q}_{\gamma \circ \varphi}).$$

Proof Note that $\mathcal{F}_\gamma \preceq (L\mathcal{Q})_\gamma = (\mathcal{Q} \circ \varphi^{-1})_\gamma$ since $\mathcal{F}$ is a set of L-upper smooth functions. Observe that, for all $\mathcal{F}$,

$$\begin{aligned} \mathcal{E}(\mathcal{F} \circ \varphi) &= \left\{ w \;\middle|\; w \left[e^{-f \circ \varphi - m(f \circ \varphi)}\right] \geq 1,\ \forall f \in \mathcal{F} \right\} \\ &= \left\{ w \;\middle|\; w \circ \varphi^{-1} \left[e^{-f - m(f)}\right] \geq 1,\ \forall f \in \mathcal{F} \right\} \\ &= \left\{ \tilde{w} \circ \varphi \;\middle|\; \tilde{w} \left[e^{-f - m(f)}\right] \geq 1,\ \forall f \in \mathcal{F} \right\} \\ &= \mathcal{E}(\mathcal{F}) \circ \varphi, \end{aligned}$$

where w and $\tilde{w}$ range over $\mathcal{M}_+(\mathbb{R}^d)$. Thus, by Lemma 10, we have $\mathcal{E}(\mathcal{F}_\gamma) \supset \mathcal{E}((\mathcal{Q} \circ \varphi^{-1})_\gamma) = \mathcal{E}(\mathcal{Q}_{\gamma\circ\varphi}) \circ \varphi^{-1}$. This proves the inclusion. Now, we also have

$$\begin{aligned} C(\gamma, \mathcal{F}_\gamma) &= \inf_{w \in \mathcal{E}(\mathcal{F}_\gamma)} \ln w \left[e^{-\gamma}\right] \\ &\leq \inf_{w \in \mathcal{E}(\mathcal{Q}_{\gamma\circ\varphi})\circ\varphi^{-1}} \ln w \left[e^{-\gamma}\right] \\ &= \inf_{w \in \mathcal{E}(\mathcal{Q}_{\gamma\circ\varphi})} \ln w \circ \varphi^{-1} \left[e^{-\gamma}\right] \\ &= \inf_{w \in \mathcal{E}(\mathcal{Q}_{\gamma\circ\varphi})} \ln w \left[e^{-\gamma\circ\varphi}\right] \\ &= C(\gamma \circ \varphi, \mathcal{Q}_{\gamma\circ\varphi}), \end{aligned}$$

which yields the inequality. ■

This lemma shows that as long as we consider the envelope complexity of of upper smooth functions $\mathcal{F}$, it suffices to bound above them to evaluate the envelope complexity of penalized quadratic functions $\mathcal{Q}_\gamma$.

Further, according to the lemma below, we can restrict ourselves to one-dimensional parametric models without loss of generality if the penalty functions γ are separable. Here, γ is said to be separable if and only if it can be written in the form of $\gamma(\theta) = \sum_{j=1}^d \gamma_j(\theta_j)$.

Lemma 12 (Separability) Suppose that γ is separable. Then, the envelope complexity of $\mathcal{Q}_\gamma$ is bounded by a separable function, i.e.,

$$C(\gamma, \mathcal{Q}_\gamma) \leq \sum_{j=1}^d C(\gamma_j, \mathcal{Q}^1_{\gamma_j}),$$

where $\mathcal{Q}^1$ is the set of normalized one-dimensional quadratic functions with curvature one: $\mathcal{Q}^1 = \{x(\in \mathbb{R}) \mapsto \frac{1}{2}(x-u)^2 \mid u \in \mathbb{R}\}$.

Proof Note that all $f \in \mathcal{Q}_\gamma$ are separable, i.e., $f(\theta) = \sum_{j=1}^d f_j(\theta_j)$, where $f_j \in \mathcal{Q}^1_{\gamma_j}$ and $\gamma(\theta) = \sum_{j=1}^d \gamma_j(\theta_j)$. Let $\mathcal{E}^d = \mathcal{E}(\mathcal{Q}^1_{\gamma_1}) \otimes \cdots \otimes \mathcal{E}(\mathcal{Q}^1_{\gamma_d})$. Then, we have

$$\begin{aligned} C(\gamma, \mathcal{Q}_\gamma) &= \inf_{w \in \mathcal{E}(\mathcal{Q}_\gamma)} \ln w[e^{-\gamma}] \\ &\leq \inf_{w \in \mathcal{E}^d} \ln w[e^{-\gamma}] && \mathcal{E}^d \subset \mathcal{E}(\mathcal{Q}_\gamma) \\ &= \sum_{j=1}^d \inf_{w_j \in \mathcal{E}(\mathcal{Q}^1_{\gamma_j})} \ln w_j[e^{-\gamma_j}] \\ &= \sum_{j=1}^d C(\gamma_j, \mathcal{Q}^1_{\gamma_j}). \end{aligned}$$

■

Summary We have defined the Bayesian envelope and envelope complexity. The envelope complexity $C(\gamma, \mathcal{F})$ is equal to the Bayesian minimax regret if $\mathcal{F}$ is the set of penalized logarithmic loss functions. Any measures w in the Bayesian envelope $\mathcal{E}(\mathcal{F})$ can be utilized for bounding the Shtarkov complexity through the envelope complexity. Most importantly, the envelope complexity satisfies some useful properties such as monotonicity, parametrization invariance, and separability. Specifically, the monotonicity differentiates the envelope complexity from the Shtarkov complexity.

5.4 The ST Prior for High-Dimensional Prediction

We leverage the envelope complexity to give a Bayesian predictor closely achieving $\mathrm{LREG}^{\star}(\gamma)$, where $\gamma(\theta) = \lambda \|\theta\|_1$, namely the ST prior. Moreover, the predictor is shown to be also approximately minimax without luckiness, where $e^n \geq d/\sqrt{n} \to \infty$.

5.4.1 Envelope Complexity for ℓ_1-Penalties

Let γ be the weighted ℓ_1-norm given by

$$\gamma(\theta) = \lambda \|\theta\|_1 , \tag{5.7}$$

where $\lambda > 0$. Let π_λ be the ST prior over $\mathbb{R}^d$ given by

$$\pi_\lambda^{\mathrm{ST}}(\mathrm{d}\theta) \propto e^{-\lambda\|\theta\|_1} \prod_{j=1}^{d} w_\lambda^{\mathrm{ST}}(\mathrm{d}\theta_j), \tag{5.8}$$

$$w_\lambda^{\mathrm{ST}}(\mathrm{d}x) = \delta_0(\mathrm{d}x) + \frac{e^{\lambda^2/2}}{\lambda^2 e} \mathbb{1}\{|x| \geq \lambda\}\,\mathrm{d}x, \tag{5.9}$$

where δ_t denotes Kronecker's delta measure at t. We call it the ST prior because it consists of a delta measure (spike) and two exponential distributions (tails), as shown in Figure 5.1.

Then, envelope complexities for quadratic loss functions can be bounded as follows.

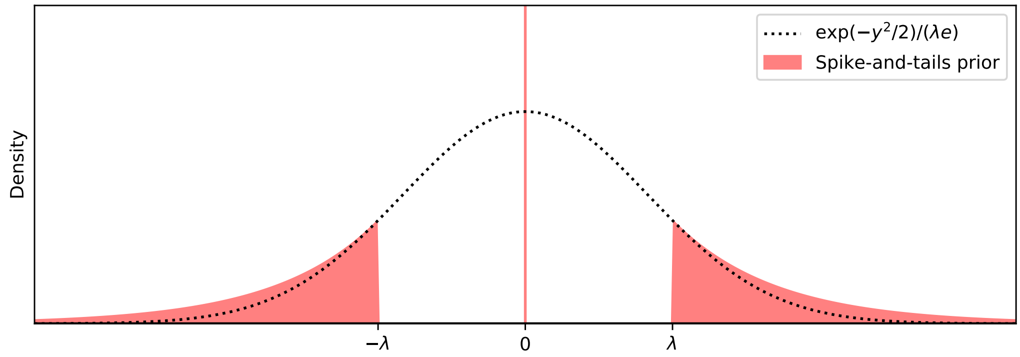


Fig. 5.1: Density of the spike-and-tails prior

Lemma 13 (Sharp bound on envelope complexity) Take γ as given by (5.7). Then, we have $w_\lambda^{\mathrm{ST}} \in \mathcal{E}(\mathcal{Q}_\gamma)$ and

$$d \ln\left(1 + \frac{e^{-\lambda^2/2}}{\lambda^3(c + o(1))}\right) \leq C(\gamma, \mathcal{Q}_\gamma) \leq \ln w_\lambda^{\mathrm{ST}}\left[e^{-\gamma}\right]$$
$$= d \ln\left(1 + \frac{2e^{-\lambda^2/2}}{\lambda^2 e}\right)$$

for some constant c, where $o(1) \to 0$ as $\lambda \to \infty$.

Proof Consider the logarithmic loss functions of the d-dimensional standard normal location model given by $f_X(\theta) = \frac{1}{2}\|X-\theta\|_2^2 + \frac{d}{2}\ln 2\pi$, $X \in \mathcal{X} = \mathbb{R}^d$, and let $\mathcal{F} = \{f_X \mid X \in \mathbb{R}^d\}$. Note that $\mathcal{F} \preceq \mathcal{Q}$. Then, the lower bound follows from Lemma 19 in Section A.3.1 with $S(\gamma) \le C(\gamma, \mathcal{F}_\gamma) \le C(\gamma, \mathcal{Q}_\gamma)$.

Note that γ is separable, and by Lemma 11, we restrict ourselves to the case of $d = 1$. Let c and t be positive real numbers. Let $w = \delta + cU$ be a measure over the real line, where δ denotes the delta measure and U denotes the Lebesgue measures restricted to $[-\lambda,\lambda]^c = \mathbb{R}\setminus[-t,t]$. That is, we have $w(E) = \mathbb{1}_{0\in E} + c\,|E\setminus[-t,t]|$ for measurable sets $E \subset \mathbb{R}$. Then, we have

$$\ln w\left[e^{-\gamma}\right] = \ln\left(1 + \frac{2c}{\lambda}e^{-t\lambda}\right). \tag{5.10}$$

We want to minimize (5.10) with respect to $w \in \mathcal{E}(\mathcal{Q}_\gamma)$. Let $f_u(\theta) = \frac{1}{2}(\theta-u)^2 + \lambda|\theta|$. Then, we have $m(f_u) = \frac{1}{2}u^2$ if $|u| \le \lambda$, and $m(f_u) = \lambda|u| - \frac{1}{2}\lambda^2$ otherwise. It suffices for c and t to have $w\left[e^{-f_u}\right] \ge e^{-m(f_u)}$ for all $u \in \mathbb{R}$. Here, we only care about the case of $u \ge \lambda$ since it is symmetric with respect to u and trivially we have $w\left[e^{-f_u}\right] \ge \delta\left[e^{-f_u}\right] \ge e^{-m(f_u)}$ for all $u \in [-\lambda,\lambda]$. Now, for $x = u - \lambda \ge 0$, we have

$$\begin{aligned} w\left[e^{-f_u}\right] &= e^{-\frac{1}{2}u^2} + ce^{-t\lambda}\left(\int_{-\infty}^{-t} + \int_t^\infty\right) e^{-\frac{1}{2}(\theta-u)^2}d\theta \\ &\ge e^{-\frac{1}{2}u^2} + ce^{-t\lambda}\int_t^\infty e^{-\frac{1}{2}(\theta-u)^2}d\theta \\ &= e^{-m(f_u)}\left(e^{-\frac{1}{2}x^2} + c\int_{t-x}^\infty e^{-\frac{1}{2}y^2}dy\right). \end{aligned}$$

Let $A(x) = e^{-\frac{1}{2}x^2} + c\int_{t-x}^\infty e^{-\frac{1}{2}y^2}dy$. Thus a sufficient condition for $w \in \mathcal{E}(\mathcal{Q}_\gamma)$ is that $A'(x) = ce^{-\frac{1}{2}(t-x)^2} - xe^{-\frac{1}{2}x^2} \ge 0$, which is satisfied with $c = \frac{1}{t}\exp\left(\frac{1}{2}t^2 - 1\right)$. Finally, evaluating (5.10) at $t = \lambda$ yields the ST pre-prior $w = w_\lambda^{\mathrm{ST}}$. Therefore, we have $w_\lambda^{\mathrm{ST}} \in \mathcal{E}(\mathcal{Q}_\gamma)$, and the upper bound is shown. The equality is a result of straightforward calculation of $\ln w\left[e^{-\gamma}\right]$. ■

According to Lemma 13, the ST prior bounds the envelope complexity in a quadratic rate as $\lambda \to \infty$. The exponent $-\frac{1}{2}\lambda^2/2$ is optimally sharp since the lower bound $C(\gamma, \mathcal{Q}_\gamma) = \Omega(d\exp\left[-\frac{1}{2}\lambda^2\right]/\lambda^3)$ has the same exponent.

This gives an upper bound on the envelope complexity for general smooth loss functions. Let $\pi_{\lambda,L}^{\mathrm{ST}}$ and $w_{\lambda,L}^{\mathrm{ST}}$ be the scale-corrected ST preprior and prior given respectively by

$$\pi_{\lambda,L}^{\mathrm{ST}}(\mathrm{d}\theta) = \pi_{\lambda/\sqrt{L}}^{\mathrm{ST}}(\sqrt{L}\mathrm{d}\theta), \qquad w_{\lambda,L}^{\mathrm{ST}}(\mathrm{d}\theta) = w_{\lambda/\sqrt{L}}^{\mathrm{ST}}(\sqrt{L}\mathrm{d}\theta).$$

The following is a direct corollary of Lemma 11, Lemma 12, Lemma 13, and Lemma 10.

Corollary 5 If all $f \in \mathcal{F}$ are L-upper smooth with respect to θ, and if γ is given by (5.7), then $w_{\lambda,L}^{\mathrm{ST}} \in \mathcal{E}(\mathcal{F}_\gamma)$, and therefore

$$C(\gamma, \mathcal{F}_\gamma) \le \ln w_{\lambda,L}^{\mathrm{ST}}\left[e^{-\gamma}\right] = d\ln\left(1 + \frac{2L}{e\lambda^2}e^{-\frac{1}{2L}\lambda^2}\right).$$

5.4.2 Regret Bound with the ST Prior

Now, we utilize Corollary 5 for bounding actual prediction performance of the ST prior. Here, we consider the scenario of online-learning under ℓ_1-constraint.

Setup Let $X^n = (X_1, \dots, X_n) \in \mathcal{X}^n$ be a sequence of outcomes. Let f_X be a logarithmic loss function such that $\int e^{-f_X(\theta)} d\nu(X) \leq 1$. Then, the conditional Bayesian pre-posterior with respect to $w \in \mathcal{M}_+(\mathbb{R}^d)$ given X^t $(0 \leq t \leq n)$ is given by

$$w(\mathrm{d}\theta|X^t) = w(\mathrm{d}\theta) \prod_{i=1}^{t} \exp\left\{-f_{X_i}(\theta)\right\}.$$

The online regret of the predictor is defined as

$$\mathrm{REG}_n(w|\mathcal{H}) \overset{\text{def}}{=} \sup_{X^n \in \mathcal{X}^n, \theta^* \in \mathcal{H}} \sum_{t=1}^{n} \left\{f_{X_t}(w(\cdot|X^{t-1})) - f_{X_t}(\theta^*)\right\}. \tag{5.11}$$

Now, we can bound the online regret of the ST prior as follows.

Theorem 14 (Adaptive minimaxity over ℓ_1-balls) Suppose that f_{X_i} are L-upper smooth and logarithmic. Let $\mathcal{H}_B = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq B\}$. Take $\lambda = \sqrt{2Ln \ln(d/\sqrt{Ln})}$. Then, with $\omega(1) = \ln(d/\sqrt{n}) = o(n)$, we have

$$\mathrm{REG}_n(w^{\mathrm{ST}}_{\lambda,Ln}|\mathcal{H}_B) \leq B\sqrt{2Ln \ln \frac{d}{\sqrt{Ln}}(1 + o(1))}$$

for all $B > 0$. Moreover, this is adaptive minimax rate and not improvable by more than a factor of two, even if B is fixed and non-Bayesian predictors are involved.

Proof Let f_{X^n} be the cumulative loss $f_{X^n} = \sum_{i=1}^n f_{X_i}$, and observe that f_{X^n} is Ln-upper smooth and logarithmic. Let $\mathcal{F} = \{f_{X^n} \mid X^n \in \mathcal{X}^n\}$ and $\gamma(\theta) = \lambda \|\theta\|_1$. Also, let γ_0 be the indicator penalty of the set $\mathcal{H}_B$ such that $\gamma_0(\theta) = 0$ if and only if $\theta \in \mathcal{H}_B$, and $\gamma_0(\theta) = \infty$ otherwise. Then, we have $\mathrm{REG}_n(w|\mathcal{H}_B) = \mathrm{LREG}(w|\gamma_0)$, where LREG is taken with respect to f_{X^n}. Now, observe that

$$\begin{aligned}
\mathrm{LREG}(w^{\mathrm{ST}}_{\lambda,Ln}|\gamma_0) &\leq \mathrm{LREG}(w^{\mathrm{ST}}_{\lambda,Ln}|\gamma - \lambda B) \\
&\quad (\because \gamma_0 \geq \gamma - \lambda B) \\
&\leq \ln w^{\mathrm{ST}}_{\lambda,Ln}\left[e^{-\gamma+\lambda B}\right], \\
&\quad (\because \text{Theorem 8}) \\
&= \lambda B + \ln w^{\mathrm{ST}}_{\lambda,Ln}\left[e^{-\gamma}\right],
\end{aligned}$$

which, combined with Corollary 5 where $\lambda = \sqrt{2Ln \ln(d/\sqrt{Ln})}$, yields the asymptotic equality. The proof of the minimaxity is adopted from the existing analysis on the minimax *risk* (see Section A.3.2 for the rigorous proof and Section 5.6.5 for detailed discussions). ■

5.5 Visual Comparison of the ST Prior and the Tilted Jeffreys Prior

Now, we verify the results on the ℓ_1-regularization obtained above. In particular, we compare the worst-case regrets achievable with Bayesian predictors to the minimax regret, i.e., the Shtarkov complexity.

Setting We adopted the one-dimensional quadratic loss functions with curvature one, $q \in \mathcal{Q}^1$, and the ℓ_1-penalty function $\gamma(\theta) = \lambda\,|\theta|$. We varied the penalty weight λ from 10^{-1} to 10^1 and observed how the worst-case regret of each Bayesian predictor changes. Specifically, we employed the ST prior (5.9) and the tilted Jeffreys prior for the predictors. Note that in this case the tilted Jeffreys prior is nothing more than the double exponential prior given by $\pi_\lambda^{\mathrm{Jeff}'}(\mathrm{d}\theta) = \frac{\lambda}{2}e^{-\lambda|\theta|}\mathrm{d}\theta$.

Results In Figure 5.2, the worst-case regrets of the ST prior and the Jeffreys prior are shown along with the minimax regret (Optimal). While the regret of the tilted Jeffreys prior is almost same as the optimal regret where λ is small, it performs poorly where λ is large. On the other hand, the ST prior performs robustly well in the entire range of λ. Specifically, it converges to zero quadratically where λ is large. Therefore, since one must take λ sufficiently large if d is large, it is implied that the ST prior is a better choice than the tilted Jeffreys prior.

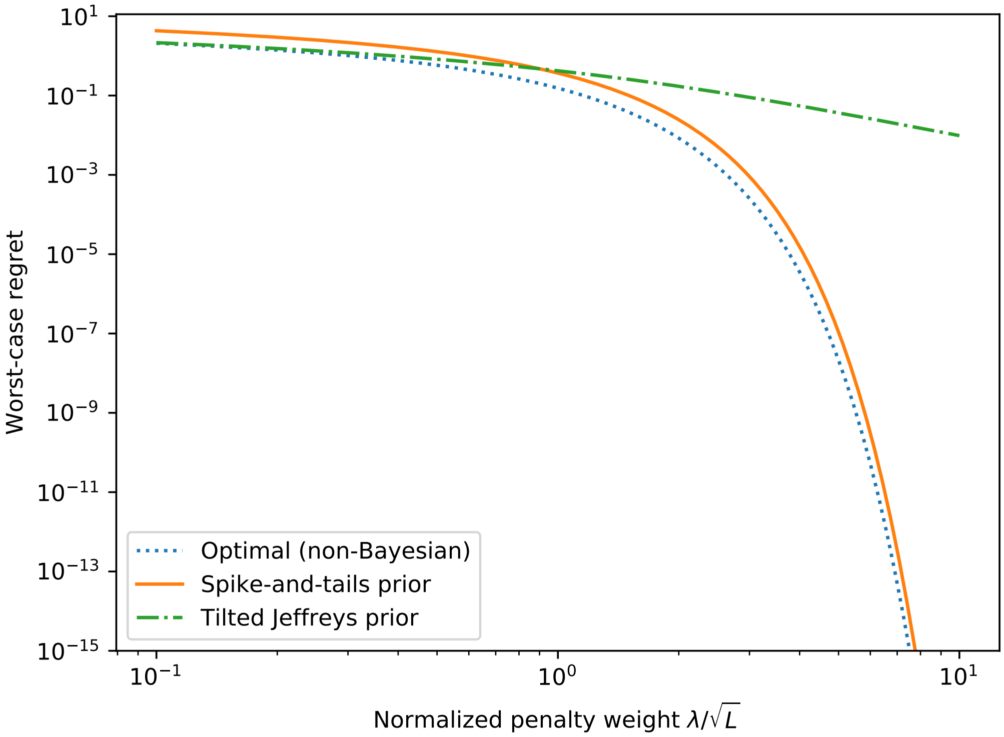


Fig. 5.2: Worst-case regrets of the spike-and-tails prior and the tilted Jeffreys prior

5.6 Implications and Discussions

In this section, we discuss interpretations of the results and present solutions to some technical difficulties.

5.6.1 Gap between $\mathrm{LREG}^\star$ and $\mathrm{LREG}^{\mathrm{Bayes}}$

One may wonder if there exists a prior that achieves the lower bound $\mathrm{LREG}^\star(\gamma)$, where $\gamma(\theta) = \lambda\,\|\theta\|_1$, $\lambda > 0$. Unfortunately, the answer is no. With a similar technique of higher-order differentiations used by Hedayati and Bartlett (2012), we can show that if γ is convex and not differentiable like the ℓ_1-norm, then the gap is nonzero, i.e., $\mathrm{LREG}^\star(\gamma) < \mathrm{LREG}^{\mathrm{Bayes}}(\gamma)$. The detailed statement and proof of this is in Section A.3.3.

5.6.2 Infinite-dimensional Models

If the dimensionality d of the parameter space is countably infinite, the minimax regret $\mathrm{REG}^\star(\mathcal{H}_B)$ with any nonzero radius B diverges. In this case, one may apply different penalty weights to different dimensions. For instance, taking different penalty weights for different dimensions, e.g., $\gamma(\theta) = \sum_{j=1} \lambda_j |\theta_j|$ for $\lambda_j = \sqrt{2L\,\mathrm{Ln}\{j\,\mathrm{Ln}\,j\}}$ and $\mathrm{Ln}\,x = \ln\max\{e, x\}$, the separability of the envelope complexity guarantees that $C(\gamma, \mathcal{F}_\gamma) \leq \sum_{j=1}^{\infty} \left(j\,\mathrm{Ln}^2\,j\right)^{-1} < +\infty$. Then, the corresponding countably-infinite tensor product of the one-dimensional ST prior $\pi^{\mathrm{ST}}_{\{\lambda_j\}}(\mathrm{d}\theta) = \prod_{j=1}^{\infty} \pi^{\mathrm{ST}}_{\lambda_j}(\mathrm{d}\theta_j)$ gives a finite regret with respect the infinite-dimensional models $\mathcal{H} = \{\theta \in \mathbb{R}^{\mathbb{N}} \mid \gamma(\theta) \leq B\}$.

5.6.3 Comparison to the Titled Jeffreys Priors and Others

There have been previous studies on the minimax regret with Bayesian predictors (Takeuchi and Barron, 1998, 2013; Watanabe and Roos, 2015; Xie and Barron, 2000). In these studies, the Bayesian predictor based on the Jeffreys prior (namely Jeffreys predictor) is proved to attain minimax-regret asymptotically under some regularity conditions. The tilted Jeffreys prior, which takes the effect of penalization γ into consideration, is given by Grünwald (2007) as $\pi_{\mathrm{Jeff}'}(\mathrm{d}\theta) \propto \mathrm{d}\theta\sqrt{\det I(\theta)}e^{-\gamma(\theta)}$, where $I(\theta)$ denotes the Fisher information matrix. In the case of quadratic loss functions $\mathcal{Q}$, as the Fisher information is equal to identity, we have $\pi_{\mathrm{Jeff}'}(\mathrm{d}\theta) \propto e^{-\gamma}\mathrm{d}\theta$. Therefore, it is implied that that taking the uniform pre-prior $w(\mathrm{d}\theta) \propto \mathrm{d}\theta$ is good for smooth models under the conventional large-sample limit. This is in very strong contrast with our result, where completely nonuniform preprior w^{ST}_λ performs better with high-dimensional models.

5.6.4 Comparison to Online Convex Optimization

So far, we have considered the luckiness minimax regret, which leads to the adaptive minimax regret. Perhaps surprizingly, our minimax regret bound coincides with the results given in the literature of online convex optimization, where different assumptions on the loss functions and predictors are made. Specifically, with $\lambda = \sqrt{2L\ln d}$, the regret bound is reduced to $\sqrt{2L\ln d} + 1/e$. This coincides with the standard no-regret rates of online learning such as Hedge algorithm (Freund and Schapire, 1997) and high-dimensional online regression (Gerchinovitz and Yu, 2014), where L is referred to as the number of trials T and d is referred to as the number of experts or dimensions n. Moreover, with $\lambda = 1$, the regret bound is reduced to $O(d\ln L)$. This is equal to the minimax-regret rate achieved under large-sample asymptotics such as in Hazan et al. (2007); Cover (2011).

Note that the conditions assumed in those two regimes are somewhat different. In our setting, loss functions are assumed to be upper smooth and satisfy some normalizing condition to be logarithmic losses, while the boundedness and convexity of loss functions is often assumed in online learning. Moreover, we have employed Bayesian predictors, whereas more simple online predictors are typically used in the context of online learning.

5.6.5 Comparison to Minimax Risk over ℓ_1-balls

In the literature of high-dimensional statistics, the minimax rate of *statistical risk* is also achieved with ℓ_1-regularization (Donoho and Johnstone, 1994), where the true parameter θ is in the unit ℓ_1-ball. Although both risk and regret are performance measures of prediction, there are two notable differences. One is that risks are calculated under some assumptions on a true statistical distribution, whereas regrets are defined without any assumptions on data. The other is that risks are typically considered with an in-model predictor, i.e., predictors are restricted to a given model, whereas regrets are often considered with out-model predictors such as Bayesian predictors and online predictors. Therefore, the minimax regret can be regarded as a more agnostic complexity measure than the minimax risk.

If we assume Gaussian noise models and adopt the logarithmic loss functions, the minimax rate of the risk is given as $\sqrt{2L \ln d/\sqrt{L}}$ according to Donoho and Johnstone (1994). Interestingly, this is same with the rate of the regret bound given by Theorem 14, where $L = Ln$. Moreover, the minimax-risk optimal penalty weights λ are also minimax-regret optimal in this case. Therefore, if the dimensionality d is large enough compared to L (n in case of online-learning), making no distributional assumption on data costs nothing in terms of the minimax rate.

5.7 Conclusion

In this study, we presented a novel characterization of the minimax regret for logarithmic loss functions called the envelope complexity, with ℓ_1-regularization problems. The virtue of envelope complexity is that it is much easier to evaluate than the minimax regret itself and is able to produce upper bounds systematically. Then, using the envelope complexity, we proposed the ST prior, which almost achieves the luckiness minimax regret against smooth loss functions under ℓ_1-penalization. We also show that the ST prior actually adaptively achieves the 2-approximate minimax regret under high-dimensional asymptotics $\omega(1) = \ln d/\sqrt{n} = o(n)$. Experimentally, we have confirmed our theoretical results: the ST prior outperforms the tilted Jeffreys prior where the dimensionality d is high, whereas the tilted Jeffreys prior is optimal if $n \gg d$.

Limitations and future work The present work is relying on the assumption of smoothness and logarithmic property on the loss functions. The smoothness assumption may be removed by considering the smoothing effect of stochastic algorithms like SGD, as in Kleinberg et al. (2018). As for the logarithmic assumption, it will be generalized to evaluate complexities with non-logarithmic loss functions with the help of tools that have been developed in the literature of information theory, such as in Yamanishi (1998). Finally, since our regret bound with the ST prior is quite simple (there are only the smoothness L and the radius B except with the logarithmic term), applying these results to concrete models such as deep learning models would be interesting future work, as would comparison to the existing generalization error bounds.

Chapter 6

Excess Risk Bounds with Envelope Complexity

In this chapter, we derive a novel risk bound in terms of the envelope complexity. This directly connects the Bayesian minimax regret with statistical risk and reveals a new relationship between the minimax-regret code length and batch prediction. First, we give a generic risk bound based on the PAC-Bayesian analysis and the envelope complexity, namely PAC-Bayesian-Envelope (PAC-BE) bound. Then, we present an instance of the PAC-BE bound for practical models.

6.1 Motivation

The envelope complexity, which we introduced in Chapter 5, is a complexity measure of online-learning problems in terms of the minimax regret achievable with Bayesian predictors. Since Bayesian predictors are a powerful yet relatively easy-to-compute class of predictors, the envelope complexity is an important quantity in its own right in the online-learning scenario.

On the contrary, sometimes one may want to guarantee the instantaneous risk of predictions at a specific future time point instead of the cumulative/averaged loss over time. Such a learning regime is addressed within the framework of the *batch learning problem.* In the batch learning scenario, the data X^{n+1} (note that the future data X_{n+1} is included here) are assumed to be subject to some unknown probability distribution. Since one cannot make any meaningful claims on the worst-case behavior on the instantaneous loss for unseen data, the *excess risk*

$$r(h|\mathcal{H}) \overset{\text{def}}{=} R(h) - \inf_{h' \in \mathcal{H}} R(h')$$

may be considered instead. Here, $R(h) \overset{\text{def}}{=} \mathbb{E}\ell_{X_{n+1}}(h)$ denotes the expected loss or risk of h.

In fact, online predictors induce certain excess-risk bounds independent of the data-generating distribution if their worst-case regrets are bounded (Littlestone, 1989; Cesa-Bianchi et al., 2004; Cesa-Bianchi and Lugosi, 2006). Specifically, if X^{n+1} are i.i.d and the loss functions is logarithmic, the averaged predictor associated with online predictor h

$$Q^{\text{ave}}(x) = \frac{1}{n} \sum_{i=1}^{n} P(x|h(X^{i-1}))$$

achieves the following excess risk bound:

$$r(Q^{\mathrm{ave}}|\mathcal{H}) \leq \frac{1}{n} \sup_{X^n} \mathrm{REG}(h|X^n, \mathcal{H}).$$

Therefore, good predictors in the online-learning scenario can be utilized for constructing good batch predictors. However, the averaged predictor is computationally expensive as it requires n different predictions to be computed for every single batch prediction, which is unacceptable for high-dimensional and complex models.

In this study, we consider plug-in predictors rather than averaged predictors. That is, we want to find a predictor $h^\star$ in the given model $\mathcal{H}$ that (approximately) achieves small excess risks

$$\hat{h} \approx \operatorname*{argmin}_{h \in \mathcal{H}} r(h|\mathcal{H}),$$

where $\hat{h}$ can be stochastically chosen. Then, we establish a novel relationship between Bayesian online and plug-in batch predictions combining the envelope complexity with PAC-Bayesian analysis (McAllester, 1999; Catoni, 2007). In particular, we develop a novel (excess) risk bound, namely the PAC-Bayesian-Envelope (PAC-BE) bound. The significance of our approach is summarized as follows:

- It enable us to systematically compute risk bounds for a wide range of complex models via the monotonicity of the envelope complexity.
- Its PAC-Bayesian nature allows us to drop the logarithmic assumption of the loss function.

The remainder of this chapter is organized as follows. In Section 6.2, we review previous research on controlling (excess) risks in terms of information theory and clarify the differences and our contributions relative to them. Then, we proceed to the main results on the PAC-BE bound in Section 6.3. In Section 6.4, we develop some instances of the PAC-BE bound that will be useful in practice. Then, we compare our risk bounds to the existing risk bounds and discuss the advantages and disadvantages of our approach. Finally, in Section 6.6, we conclude this chapter.

6.2 Related Work

One of the earliest results on utilizing information-theoretic complexity to bound risks is Barron and Cover (1991). They proposed the notion of the index of resolvability, which bounds the risk (not excess risk). The first PAC-Bayesian risk bound was proposed by McAllester (1999) and later numerous variants were proposed (e.g.,Catoni (2007); Seldin et al. (2012); Bégin et al. (2016)). The combination of the PAC-Bayesian analysis and the index of resolvability was developed by Zhang (2004). Barron and Luo (2008); Chatterjee and Barron (2014) also extended the index of resolvability applicable to the models with uncountable cardinality, including sparse regression and sparse graphical modeling. The resulting risk bounds achieve fast convergence of the excess risk if the model $\mathcal{H}$ is correctly specified, i.e., the data-generating distribution is given by $P(\cdot|h_0)$ and $h_0 \in \mathcal{H}$. However, if the model is misspecified as $h_0 \notin \mathcal{H}$, then the excess risk does not necessarily converge to zero.

Xu and Raginsky (2017); Asadi et al. (2018) also studied upper bounds of excess risk in terms of mutual information. Their results can be seen as the optimal upper bound based on PAC-Bayesian analysis. However, the resulting upper bounds are completely dependent on the data-generating distribution, which we do not know. Therefore, it does not straightforwardly provide us any principles for designing predictors without knowing the true distribution.

The PAC-BE bound, which we propose in this chapter, differs from these existing approaches in that it is able to work with misspecified models and assumes nothing about the data-generating distribution.

6.3 PAC-Bayesian Analysis with Envelope Complexity

In the following, we give risk bounds with the envelope complexity. First, we give a straightforward bound exploiting the standard reduction technique from risks to regrets. This requires the computation of posterior predictors at each iteration. Secondly, under some light-tail assumption on data, we give a risk bound called the PAC-BE bound. The PAC-BE bound suggests that we may predict with stochastic point predictors, which are easier to compute than Bayesian posteriors.

Let $x \in \mathcal{X}$ be a random variable generated from a stochastic source $\mathcal{S}$ and $\theta \in \mathbb{R}^d$ be a parameter of predictors for x. Let $\ell_x : \theta \mapsto \mathbb{R}$ be the loss incurred for making predictions with θ over actual outcomes x. Let $\ell_{\mathcal{S}}(\theta) \overset{\text{def}}{=} R(\theta) = \mathbb{E}_x[\ell_x(\theta)]$ be the risk of predictor θ. Assume that the lower tail probability of the loss function ℓ_x is σ-subGaussian,

$$\mathbb{E}_x\left[e^{-\beta(\ell_x(\theta)-\ell_{\mathcal{S}}(\theta))}\right] \leq \exp\left(\frac{\sigma^2\beta^2}{2}\right) \tag{6.1}$$

for all $\beta > 0$ and all predictors $\theta \in \mathbb{R}^d$.

Let f_x be an arbitrary surrogate function of ℓ_x, where $\ell_x(\theta) \leq f_x(\theta)$ for all $\theta \in \mathbb{R}^d$. Here, ℓ_x may be difficult to optimize due to the non-convexity of the loss with respect to the parameter θ, whereas f_x can be considered as easier to minimize.

Let $X = x^n = (x_1, \dots, x_n) \in \mathcal{X}^n$ be an n-sequence of i.i.d. copies of x and define the empirical risk function $\bar{\ell}_X(\theta) = \sum_{i=1}^n \ell_{x_i}(\theta)/n$ and the empirical surrogate risk function $\bar{f}_X(\theta) = \sum_{i=1}^n f_{x_i}(\theta)/n$. Also, let $\mathcal{F} = \{\bar{f}_X \mid X \in \mathcal{X}^n\}$ be the entire set of $\bar{f}_X$. We are allowed to exploit $\bar{f}_X \in \mathcal{F}$ for choosing a good predictor θ, which minimizes the risk $\ell_{\mathcal{S}}(\theta)$. To control the complexity of the predictors θ, also consider penalty functions $\gamma : \theta \mapsto [0, \infty]$.

To choose a good predictor, we consider prior and posterior distributions given as follows. Let $\mathcal{M}_+(\mathbb{R}^d)$ be all nonnegative Borel measures over $\mathbb{R}^d$. For all $w \in \mathcal{M}_+(\mathbb{R}^d)$, we denote by $w[\cdot]$ the integration with respect to $w(\mathrm{d}\theta)$. Let $\mathcal{P}(\mathbb{R}^d)$ be the set of all probability measures π over $\mathbb{R}^d$, which means that $\pi[1] = 1$. Let $P \in \mathcal{P}(\mathbb{R}^d)$ be the priors given by $P(\mathrm{d}\theta) \propto e^{-\alpha\gamma(\theta)}w(\mathrm{d}\theta)$ for some $w \in \mathcal{M}_+(\mathbb{R}^d)$ and $\alpha > 0$. Then, the corresponding Gibbs posteriors are given by

$$Q_X(\mathrm{d}\theta) \propto P(\mathrm{d}\theta)\exp\left[-\beta\bar{f}_X(\theta)\right], \quad \beta > 0, \tag{6.2}$$

which we make a prediction based on in the following sections.

We analyze the performance of such predictions on the basis of the Bayesian envelopes. Let $\mathcal{E}(\mathcal{G})$ be the Bayesian envelope of sets of functions $\mathcal{G} \ni g : \theta \mapsto \overline{\mathbb{R}}$, which is given by

$$\mathcal{E}(\mathcal{G}) \overset{\text{def}}{=} \bigcap_{g\in\mathcal{G}} \bigcap_{\theta\in\mathbb{R}^d} \left\{ w \in \mathcal{M}_+(\mathbb{R}^d) \,\middle|\, w\left[e^{-g}\right] \geq e^{-g(\theta)} \right\}.$$

We may consider $w \in \mathcal{E}(\mathcal{G})$ as envelope measures of $\mathcal{G}$. We also define the envelope complexity of measures $w \in \mathcal{M}_+(\mathbb{R}^d)$ with respect to the surrogate loss functions $\mathcal{F}$ and the penalty γ as

$$C(w) = C(w|\mathcal{F},\gamma) \overset{\text{def}}{=} \begin{cases} \ln w\left[e^{-\gamma}\right] & (w \in \mathcal{E}(\mathcal{F}+\gamma)) \\ \infty & (\text{otherwise}) \end{cases},$$

where we abbreviate γ and $\mathcal{F}$ when any confusion is unlikely. Moreover, let $C_\beta(w) = C_\beta(w|\gamma,\mathcal{F}) = C(w|\beta\gamma,\beta\mathcal{F})/\beta$ be the annealed envelope complexity with inverse temperature parameter $\beta > 0$.

6.3.1 Direct Reduction for Logarithmic Losses

Assume that, for now, ℓ_x is logarithmic with respect to x, i.e., $\ell_x(\theta) = -\ln p(x|\theta)$ for some probability density functions $p(\cdot|\theta)$. Moreover, for all priors $\pi \in \mathcal{P}(\mathbb{R}^d)$, denote the loss of the corresponding Bayesian predictors by $\ell_x(\pi) = -\ln \pi\left[p(x|\cdot)\right]$ and the risk by $\ell_\mathcal{S}(\pi) = \mathbb{E}_x \ell_x(\pi)$.

Now, we present the risk bounds on the Bayesian predictors via direct reduction from risks to regrets. Define the averaged posterior as

$$Q_X^{\text{ave}}(\mathrm{d}\theta) = \frac{1}{n}\sum_{t=1}^{n} Q_X^{(t-1)}(\mathrm{d}\theta), \tag{6.3}$$

$$Q_X^{(t)}(\mathrm{d}\theta) \propto \exp\left[-\sum_{i=1}^{t} f_{x_i}(\theta) + n\gamma(\theta)\right] w(\mathrm{d}\theta). \tag{6.4}$$

Below, we present the risk bound of the predictions based on the posteriors Q_X^{ave} with respect to the best risk.

Theorem 15 (Direct risk bound) Take an arbitrary $w \in \mathcal{M}_+(\mathbb{R}^d)$. Then, we have

$$\mathbb{E}\ell_\mathcal{S}(Q_X^{\text{ave}}) \le \mathbb{E}\bar{f}_X(\theta^*) + \gamma(\theta^*) + C_n(w)$$

for all $\theta^* \in \mathbb{R}^d$.

Proof It follows that

$$\begin{aligned}
\mathbb{E}\ell_\mathcal{S}(Q_X^{\text{ave}}) &\le \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ell_\mathcal{S}\left(Q_X^{(t-1)}\right) && (\text{Jensen's inequality}) \\
&= -\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ln Q_X^{(t-1)}\left[e^{-\ell_{x_t}}\right] \\
&\le -\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\ln Q_X^{(t-1)}\left[e^{-f_{x_t}}\right] \\
&= \frac{1}{n}\mathbb{E}\ln\frac{w\left[e^{-n\gamma}\right]}{w\left[e^{-n\bar{f}_X - n\gamma}\right]} && (\text{telescoping sum}) \\
&\le C_n(w) - \frac{1}{n}\mathbb{E}\ln w\left[e^{-n\bar{f}_X - n\gamma}\right] && (\text{definition of } C_n(w)) \\
&\le \mathbb{E}\bar{f}_X(\theta^*) + \gamma(\theta^*) + C_n(w). && (w \in \mathcal{E}(n\mathcal{F}+n\gamma) \text{ w.l.o.g.})
\end{aligned}$$

■

The key point here is the triplet $(\bar{f}_X, \gamma, w)$ and how they interact with each other through the envelope complexity $C_n(w) = C_n(w|\mathcal{F},\gamma)$. In particular, it is necessary for giving small upper bounds to design $\bar{f}_X$ such that there exists a good approximation $\bar{f}_X \approx \bar{\ell}_X$ and that it gives a low envelope complexity $C_n(w|\mathcal{F}, \gamma)$. If we take $f_x = \ell_x$ for all $x \in \mathcal{X}$, then the risk bound simplifies and we have an excess risk bound as follows.

Corollary 6 (Direct excess risk bound) Let $\mathcal{H}_B = \left\{\theta \in \mathbb{R}^d : \gamma(\theta) \leq B\right\}$ and take $f_x = \ell_x$. Take an arbitrary $w \in \mathcal{M}_+(\mathbb{R}^d)$. Then, we have

$$r(Q_X^{\mathrm{ave}}|\mathcal{H}_B) \leq B + C_n(w)$$

for all $\theta^* \in \mathbb{R}^d$.

We note that if $\gamma \equiv 0$, then the conventional regret-based excess risk bound is recovered:

$$r(Q_X^{\mathrm{ave}}|\mathcal{H}) \leq C_n(w),$$

where $\mathcal{H} = \left\{\theta \in \mathbb{R}^d\right\}$.

6.3.2 PAC-BE Bounds on Generalization Errors

The above risk bound guarantees the predictive performance of the Bayesian predictors π_T, which are not necessarily easy to compute. Moreover, it is only applicable to logarithmic loss functions. However, at the cost of an additional complexity, we can guarantee the performance of stochastic point predictors, which are much less computationally expensive than Bayesian predictors, with non-logarithmic loss functions.

Let $\beta > 0$ be an inverse temperature parameter and let P be a prior over predictors θ such that $P(\mathrm{d}\theta) \propto e^{-\beta\gamma(\theta)}w(\mathrm{d}\theta)$ for some envelope measures $w \in \mathcal{E}(\beta\mathcal{F}_\gamma)$. The Gibbs posteriors Q_X are now defined as

$$Q_X(\mathrm{d}\theta) \propto P(\mathrm{d}\theta)e^{-\beta\bar{f}_X(\theta)}. \tag{6.5}$$

Then, given observations X, we make a prediction with θ randomly drawn from the Gibbs posterior Q_X. The average risk is hence written as $Q_X[\ell_\mathcal{S}]$. Finally, define the gap function as

$$\Delta_X = \bar{f}_X - \bar{\ell}_X, \tag{6.6}$$

which is nonnegative. Below, we show that $Q_X[\ell_\mathcal{S} + \Delta_X]$, which is an upper bound of the risk $Q_X\ell_\mathcal{S}$, is bounded with respect to the best empirical penalized risk.

Theorem 16 (PAC-BE risk bound) Assume the σ-subGaussian property of ℓ_X as in (6.1). Let Δ_X be given by (6.6). For arbitrary $\beta > 0$ and $w \in \mathcal{M}(\mathbb{R}^d)$, define the posterior Q_X as in (6.5). Then, with probability $1 - \delta$ over the draw of $X \sim \mathcal{S}^n$, we have

$$Q_X[\ell_\mathcal{S} + \Delta_X] \leq \bar{f}_X(\theta^*) + \gamma(\theta^*) + C_\beta(w) + \frac{\ln\frac{1}{\delta}}{\beta} + \frac{\beta\sigma^2}{2n}$$

for all $\theta^* \in \mathbb{R}^d$. Moreover, we have a similar bound on the expectation:

$$\mathbb{E}Q_X[\ell_\mathcal{S} + \Delta_X] \leq \mathbb{E}\bar{f}_X(\theta^*) + \gamma(\theta^*) + C_\beta(w) + \frac{\beta\sigma^2}{2n}.$$

Proof Without loss of generality, we assume that $w \in \mathcal{E}(\beta\mathcal{F}_\gamma)$. According to the change of measure lemma (Donsker and Varadhan, 1975), we have

$$Q_X\left[\beta(\ell_\mathcal{S} - \bar{\ell}_X)\right] \leq KL(Q_X\|P) + \ln P\left[e^{\beta(\ell_\mathcal{S} - \bar{\ell}_X)}\right]$$

for all X. The definitions of measures P and Q_X yield

$$\begin{aligned} KL(Q_X \| P) &= Q_X \left[\ln \frac{\mathrm{d}Q_X}{\mathrm{d}P} \right] \\ &= -\beta Q_X \bar{f}_X + \ln \frac{w\left[e^{-\beta\gamma}\right]}{w\left[e^{-\beta(\bar{f}_X+\gamma)}\right]} \\ &\leq -\beta Q_X \bar{f}_X + \beta C_\beta(w) + \beta(\bar{f}_X(\theta^*) + \gamma(\theta^*)) && \because w \in \mathcal{E}(\beta\mathcal{F}_\gamma) \end{aligned}$$

for all $\theta^* \in \mathbb{R}^d$. Thus, combining both inequalities divided by β, we have

$$Q_X\left[\ell_\mathcal{S} + \Delta_X\right] \leq \bar{f}_X(\theta^*) + \gamma(\theta^*) + C_\beta(w) + \frac{1}{\beta} \ln P\left[e^{\beta(\ell_\mathcal{S} - \bar{\ell}_X)}\right].$$

The last term is further evaluated as

$$\begin{aligned} P\left[e^{\beta(\ell_\mathcal{S} - \bar{\ell}_X)}\right] &\leq \frac{1}{\delta} P\mathbb{E}\left[e^{\beta(\ell_\mathcal{S} - \bar{\ell}_X)}\right] && \text{Markov's inequality} \\ &= \frac{1}{\delta} P\mathbb{E}\left[\exp\left\{\frac{\beta}{n}\sum_{i=1}^n (\ell_\mathcal{S} - \ell_{x_i})\right\}\right] \\ &\leq \frac{1}{\delta}\exp\left(\frac{\sigma^2\beta^2}{2n}\right) && \text{subgaussian tail} \end{aligned}$$

with probability $1-\delta$. This yields the desired bound. As for the expectation bound, take the expectation of the last term and evaluate it as follows:

$$\begin{aligned} \mathbb{E} \ln P\left[e^{\beta(\ell_\mathcal{S} - \ell_X)}\right] &\leq \ln P\mathbb{E}\left[e^{\beta(\ell_\mathcal{S} - \ell_X)}\right] && \text{Jensen's inequality} \\ &\leq \frac{\sigma^2\beta^2}{2}. && \text{subgaussian tail} \end{aligned}$$

This completes the proof. ■

We call this bound the PAC-BE bound as it is derived from the PAC-Bayesian bound and the Bayesian envelope. A notable difference between the PAC-BE bound and the direct bound is that there is the additional term $\beta\sigma^2/2n$. Without this term, the PAC-BE bound is essentially the same as the direct bound taking $\beta = n$. Thus, the last term can be seen as the price for restricting the predictor θ to point predictors from Bayesian ones.

As a direct corollary, we have an excess risk bound if $f_x = \ell_x$.

Corollary 7 (PAC-BE excess risk bound) Assume the σ-subGaussian property of ℓ_X as in (6.1). Let $\Delta_X = 0$ for all $X \in \mathcal{X}^n$ and let $\mathcal{H}_B = \{\gamma(\theta) \leq B\}$. For arbitrary $\beta > 0$ and $w \in \mathcal{M}(\mathbb{R}^d)$, define the posterior Q_X as in (6.5). Then, we have

$$\mathbb{E}_{\theta\sim Q_X}\left[r(\theta|\mathcal{H}_B)\right] \leq B + C_\beta(w) + \frac{\beta\sigma^2}{2n},$$

where $\mathbb{E}_{\theta\sim Q_X}$ denotes the expectation with respect to both X and θ.

6.4 PAC-BE Instances

Now, for d-dimensional parametric models, we give concrete and ready-to-use risk bounds based on the PAC-BE bound and Bayesian envelopes. In particular, we give constructive examples for the posteriors Q_X. Below, let $\theta^+ = \theta^+(X) \in \mathbb{R}^d$ be the output of an arbitrary prediction algorithm. For simplicity, we assume that $\ell_x(\theta)$ is bounded to the unit interval $[0, 1]$. Thus, the subgaussian coefficient is given as $\sigma = (2\sqrt{n})^{-1}$.

6.4.1 Quadratic Surrogates with ℓ_2-Regularization

First, we consider ℓ_2-penalty functions $\gamma = \lambda \|\theta\|_2^2$. Assume that there exists the local approximation of $\bar{f}_X$ at θ^+ given by

$$\bar{f}_X^H(\theta) = \bar{f}_X(\theta^+) + \nabla f_X(\theta^+)^\top(\theta - \theta^+) + \frac{1}{2}(\theta - \theta^+)^\top H_X(\theta - \theta^+), \quad H_X \in \mathbb{R}^{d\times d},$$

such that $f_X^H \geq f_X$. Let $D_\lambda(X)$ be the *flat-minima dimension* of f_X defined as

$$D_\lambda(X) = \frac{1}{2}\ln\det(I_d + \lambda^{-1}H_X). \tag{6.7}$$

We note that D_λ is small if f_X^H is flat, i.e., the eigenvalues of H_X are small. Let $\{\mu_j(X)\}_{j=1}^d$ be the eigenvalues of H_X in descending order and let $\mu_1(X)$ be uniformly bounded. Then, we have

$$\begin{aligned} D_\lambda(X) &= \frac{1}{2}\sum_{i=1}^n \ln\left(1 + \frac{\mu_i(X)}{\lambda}\right) \\ &\leq Cd_0\ln(1+\lambda^{-1}) + \frac{1}{2\lambda}\sum_{i=d_0+1}^n \mu_i(X), \end{aligned}$$

where $C = \frac{1}{2}\ln(1 + \sup_X \mu_1(X))$. Then, if d_0 is taken such that $\sum_{i=k_0}^n \mu_i(X)/\lambda$ is small, we have $D_\lambda = O(d_0 \ln \lambda^{-1})$, where d_0 corresponds to the number of dimensions on which f_X^H is *not* flat. In other words, D_λ decreases as the number of 'flat' dimensions increases.

Let $\Delta_X^H(\theta) = f_X^H(\theta) - \ell_X(\theta)$ (≥ 0) be the corresponding gap function. Accordingly, let Q_X^H be the surrogate posterior given by

$$Q_X^H(\mathrm{d}\theta) = w(\mathrm{d}\theta)\exp\left\{-\beta\left(f_X^H(\theta) + \gamma(\theta)\right)\right\}.$$

Note that Q_X^H is easier to compute than Q_X. Take the prior as $e^{-\beta\gamma}dw \propto d\mathcal{N}_d[\mathbf{0}, (\beta\lambda)^{-1}I_d]$. Then, the posterior is given as

$$Q_X^H = \mathcal{N}_d[\theta^+, \beta^{-1}(\lambda I_d + H_X)^{-1}]. \tag{6.8}$$

Then, as a corollary of Theorem 16, we have the following risk bounds for flat minima.

Corollary 8 (Flat-minima risk bound) Let $\gamma(\theta) = \lambda \|\theta\|_2^2$ for $\lambda > 0$. Let Q_X^H be the posterior given by (6.8). Then, we have

$$Q_X^H \ell_{\mathcal{S}} \leq f_X(\theta^+) + \lambda \left\|\theta^+\right\|_2^2 + \sqrt{\frac{\tilde{D}_\lambda(X) + \ln^\star \tilde{D}_\lambda(X) + \ln\frac{c}{\delta}}{2n}},$$

where $c \approx 2.865064$ and $\tilde{D}_\lambda(X) = 1 + D_\lambda(X)$, with probability $1-\delta$ over the draw of data $X \sim \mathcal{S}$. Here, $\ln^\star x \overset{\text{def}}{=} \ln x + \ln\ln x + \ldots$, where the sum involves only the nonnegative terms.

Proof Let $w(\mathrm{d}\theta) = (\beta\lambda/2\pi)^{d/2} e^{D_\lambda(X)}\mathrm{d}\theta$. Let $\mathcal{F}^H$ be the set of f^H_X for all possible X. Then, we have

$$\ln w\left[e^{-\beta\gamma}\right] = D_\lambda(X)$$

and $w \in \mathcal{E}(\beta(f^H_X + \gamma))$. This, combined with Theorem 16, yields

$$\begin{aligned} Q^H_X \ell_{\mathcal{S}} &\le f^H_X(\theta^+) + \lambda \left\|\theta^+\right\|_2^2 + \frac{D_\lambda(X) + \ln\frac{1}{\delta}}{\beta} + \frac{\beta\sigma^2}{2}. \\ &= f_X(\theta^+) + \lambda \left\|\theta^+\right\|_2^2 + \frac{D_\lambda(X) + \ln\frac{1}{\delta}}{\beta} + \frac{\beta\sigma^2}{2}. \end{aligned}$$

Finally, taking union bound over the best choice of $\beta \in \mathbb{N}$ with the prior proposed by Rissanen (1983) completes the proof. ■

6.4.2 Quadratic Surrogates with ℓ_1-Regularization

Next, we consider ℓ_1-regularization. Take the prior as $e^{-\gamma(\theta)}dw(\theta) = d\pi^d_{\sqrt{\beta/L}\lambda}(\sqrt{\beta L}\theta)$ with (5.9), i.e., take the posterior as

$$Q^M_X(\theta) = \pi^d_{\sqrt{\beta/L}\lambda}\left(\sqrt{\beta L}\theta \,\middle|\, \sqrt{\beta L}\left[\theta^+ - \nabla f_X(\theta^+)\right]\right), \tag{6.9}$$

where $\pi^d_\lambda(\cdot|\cdot)$ is the d-th tensor product of $\pi_\lambda(\cdot|\cdot)$ such that

$$\begin{aligned} &\pi_\lambda(\cdot|x) \propto e^{-x^2/2}\delta_0 + \\ &\qquad \frac{\sqrt{2\pi}}{\lambda e}\left(\Phi(x - 2\lambda)e^{-\lambda(x-\lambda)}\mathcal{N}_{\ge\lambda}[x-\lambda, 1] + \Phi(-x-2\lambda)e^{\lambda(x+\lambda)}\mathcal{N}_{\le -\lambda}[x+\lambda, 1]\right). \end{aligned}$$

Here, $\mathcal{N}_{\ge\lambda}$ and $\mathcal{N}_{\le\lambda}$ denote the truncated normal distributions. Then, as a corollary of Theorem 16 and Corollary 5, we have the following risk bounds.

Corollary 9 (PAC-BE risk bound with ℓ_1-penalties) Let $d \ge 2$ and assume (6.1). Suppose that f_X is L-smooth and $\gamma(\theta) = \lambda\left\|\theta\right\|_1$ with $\lambda > 0$. Let Q^M_X be the Gibbs posterior given by (6.9) with $\beta = 2\lambda^{-2}L\ln d$. Then, we have

$$Q^M_X \ell_{\mathcal{S}} \le f_X(\theta^+) + \lambda\left\|\theta^+\right\|_1 + \frac{L\ln d}{4n\lambda^2} + \frac{\lambda^2}{2L\ln d}\ln\frac{e}{\delta}$$

with probability $1 - \delta$ over the draw of data $X \sim \mathcal{S}$.

Proof Let $\mathcal{F}^M$ be the set of f^M_X for all possible X. By Corollary 5, we have $w \in \mathcal{E}(\beta\mathcal{F}^M_\gamma)$ and

$$\begin{aligned} \ln w\left[e^{-\beta\gamma}\right] &= \sum_{j=1}^d \ln\left(1 + \frac{2\beta L}{e\beta^2\lambda^2}e^{-\frac{\beta^2\lambda^2}{2\beta L}}\right) \\ &= d\ln\left(1 + \frac{1}{ed\ln d}\right) && \beta = 2\lambda^{-2}L\ln d \\ &\le \frac{1}{e\ln d} \le 1. && d \ge 2 \end{aligned}$$

Meanwhile, since $f_X^M \geq f_X$ owing to the smoothness of f_X, we can safely replace f_X with f_X^M in Theorem 16. Therefore, we have

$$\begin{aligned} Q_X\left[\ell_{\mathcal{S}} + \Delta_X^M\right] &\leq f_X^M(\theta^+) + \lambda \left\|\theta^+\right\|_2^2 + \frac{\beta\sigma^2}{2} + \frac{1}{\beta}\ln\frac{e}{\delta} \\ &= f_X(\theta^+) + \lambda\left\|\theta^+\right\|_2^2 + \frac{\sigma^2 L \ln d}{\lambda^2} + \frac{\lambda^2}{2L\ln d}\ln\frac{e}{\delta}. \quad f_X^M(\theta^+) = f_X(\theta^+) \end{aligned}$$

This, together with $\sigma = (2\sqrt{n})^{-1}$, completes the proof. ■

Comparing Corollary 8 and Corollary 9, one can see that there is a trade-off between the dependencies on the dimensionality d and the penalty weight λ. The bound for ℓ_2-regularization grows logarithmically to λ, but linearly to d in the worst case (considering all the eigenvalues μ_j as large). On the other hand, the ℓ_1-regularization bound grows polynomially to λ, but logarithmically to d. Therefore, the latter bound can be non-vacuous even if $d \gg n$.

6.5 Discussion

The PAC-Bayesian bound has been extensively utilized to bound the (excess) risk for the last couple of decades in the literature of statistical learning theory. Specifically, in the context of deep learning, Dziugaite and Roy (2017, 2018) showed that the PAC-Bayesian bound can give a non-vacuous risk bound on based on flat-minima phenomena. Neyshabur et al. (2017) also analyzed the generalization of deep neural networks with a PAC-Bayesian scheme under the large-margin assumption, where another type of flatness is assumed. Compared to these results, the flat-minima bound (Corollary 8) can be seen as yet another PAC-Bayesian interpretation of the generalization power of flat minima. The major difference of our bounds to previous work is that the flatness is characterized with the eigenspectrum of the curvature matrix H_X. The effective dimensionality associated with the curvature also plays an important role in the theory of the kernel method (Shawe-Taylor et al., 2005; Steinwart and Christmann, 2008; Suzuki, 2018). From this viewpoint, our flat-minima bound can be seen as an extension of such curvature characterization to the PAC-Bayesian framework.

On the other hand, the ℓ_1-penalized risk bound given in Corollary 9 is more closely based on the ordinary large-margin assumption. This is because with ℓ_1-penalty, the prior distribution $e^{-\gamma}w(\mathrm{d}\theta)$ depends on L, and hence the degree of flatness L cannot vary in response to data X. However, λ is easily optimized owing to the constant property of L, and we have

$$Q_X^M \ell_{\mathcal{S}} \leq f_X(\theta^+) + \left(\frac{2B^2 L \ln d}{n}\right)^{1/3} + O\left(n^{-2/3}\ln\frac{1}{\delta}\right),$$

where $\left\|\theta^+\right\|_1 \leq B$. By comparing this to the margin-based bounds, even though the exponent 1/3 is less attractive, it is independent of the margin assumption, which was difficult to guarantee beforehand, and instead, it bounds the risk with respect to the Lipschitz smoothness L. It is also advantageous that the bound is simple, easy to compute given L, and has no implicit large constant in the major term.

6.6 Conclusion

In this chapter, we have introduced the PAC-BE bound, which is derived from the PAC-Bayesian analysis and the envelope complexity. It has revealed that the Bayesian minimax regret actually bounds above the risk of plug-in batch prediction. Moreover, the proposed bound can be systematically evaluated and is applicable to non-logarithmic losses. We have also presented two instances of the PAC-BE bound that are readily applicable to actual instances of machine learning.

Chapter 7

Conclusion

7.1 Concluding Remarks

In this thesis, we have addressed fundamental challenges that occur when the MDL principle is applied to high-dimensional models. We have considered two major tasks of the MDL principle: model selection and prediction. The key challenge throughout this study was how to approximate the NML distribution in a non-asymptotic manner, and we have consistently utilized the LNML distribution, which is a relaxation of the NML distribution, to this end.

In Chapter 3 and Chapter 4, we developed two different approximation methods for high-dimensional model selection, namely the stochastic gradient approximation and smoothness-based analytic approximation. In Chapter 3, we demonstrated that the stochastic gradient of LNML is useful for selecting sparse high-dimensional models from an exponential number of candidates. The major advantage of this method is its wide applicability. It can be applied to any models as long as the stochastic gradients of their likelihoods are analytically tractable. On the other hand, as it is an iterative sampling-based method, its computational cost relies on the mixing speed of the sampling and the stopping criterion is not clearly given. In Chapter 4, we developed the analytic approximation method as well as theoretical guarantees on the approximation gap and the convergence of the local minimizer. The strengthes of this method, compared to the previous one, are the theoretical guarantees and the deterministic optimization behavior backed by the smoothness assumption. However, the applicability to non-smooth models, including ReLU neural networks, remains as future work.

In the following two chapters, we turned to the novel notion of complexity, i.e., the envelope complexity, to facilitate high-dimensional prediction based on the MDL principle. We have considered two prediction scenarios, namely online prediction and batch prediction. In Chapter 5, we studied high-dimensional online prediction and presented adaptive minimax predictors based on the analysis of the envelope complexity. In Chapter 6, we studied the batch learning scenario and revealed a novel relationship between the online and batch regimes through the generic risk bound called the PAC-Bayesian-envelope bound.

We believe that these results contribute to the foundation of the MDL principle in high-dimensional settings from both the perspectives of model selection and prediction.

7.2 Future Perspective

The strengthes of the MDL principle, in our view, are in the generality owing to its information-theoretic nature (information is everywhere!) and the existence of the closed-form optimal predictor, i.e., the NML distribution. This may be the reason why the

MDL principle has prospered, albeit in diverse formulations, in every corner of machine learning and data mining literature, and this is why we believe it is key to achieve the unified understanding of inference systems.

However, the scope of the MDL analysis has been limited to large-sample settings to date due to the lack of non-asymptotic analyses. This thesis provides three distinct methods, i.e., the stochastic gradient approximation, the smoothness-based analytic approximation, and the Bayesian minimax-regret approximation, to deal with small-sample settings.

Currently, we are aware of at least two promising future extension of these results. First, it is important to be able to deal with infinite-dimensional models. Numerous high-dimensional models, including neural networks and Gaussian mixture models, can be regarded as the finite approximation of their infinite limit. Applying the MDL principle to the infinite-dimensional models is not trivial, but it is beneficial for understanding the behavior of such large-scale models in view of information theory. We presume that the smoothness-based approximation would be fit to this end because the smoothness of some infinite-dimensional models has already been well-studied in the language of reproducing kernel Hilbert space.

Secondly, minimax regret analysis is also important in the fields of reinforcement learning and adversarial learning, where we must deal with black-box, nonstationary, and interactive environments. In this setting, the model tends to be high-dimensional as we consider complex environments or adversaries. Therefore, our approximation method of the high-dimensional minimax-regret optimal solution could be useful for improving efficiency of the reinforcement.

In either case, the approximation of the NML distribution is one of the key problems and can be addressed using our results.

Acknowledgements

I am deeply grateful to my supervisor, Prof. Kenji Yamanishi, whose comments and advice have been invaluable throughout the course of my study. He was also my supervisor for my Masters course and kindly invited me to his PhD course. His engineering philosophy on research has always been inspiring to me, and this thesis would not have been possible without it. I would also like to thank all of my thesis committee members, Prof. Fumiyasu Komaki, Prof. Tomonari Sei, Prof. Taiji Suzuki, Prof. Akiko Takeda, and my supervisor, who took their time and provided helpful feedback and suggestions to improve my thesis.

I would like to thank Dr. Shin Matsushima all of his help and encouragement throughout my Masters and PhD courses. I have learned much from him about conducting research, from problem setting to experimentation. The discussions with the members of Yamanishi laboratory and Suzuki laboratory, including Mr. Atsushi Suzuki, Dr. Atsushi Nitanda, and Dr. Taichi Kiwaki, were always exciting and an enormous help to me when I was lost with my study. It was a great pleasure to be able to share time with such nice people. I thank Dr. Hiroshi Kajino and Dr. Takayuki Osogami for mentoring my internship at IBM Research Tokyo in 2017. The experience of the internship and discussions there dramatically improved my writing and presentation skills. I have learned much from Dr. Hiroshi Kajino about his research philosophy in a practical manner.

I would also like to express my gratitude to my family for their moral support and warm encouragements. Finally, the responsibility for the final formulation and any errors contained within are entirely mine.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alquier, P., Lounici, K., et al. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145.

Asadi, A. R., Abbe, E., and Verdú, S. (2018). Chaining mutual information and tightening generalization bounds. *arXiv preprint arXiv:1806.03803*.

Barron, A. and Luo, X. (2008). MDL procedures with l1 penalty and their statistical risk. In *Proceedings of the 2008 Workshop on Information Theoretic Methods in Science and Engineering*.

Barron, A., Roos, T., and Watanabe, K. (2014). Bayesian properties of normalized maximum likelihood and its fast computation. In *IEEE International Symposium on Information Theory - Proceedings*.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444.

Catoni, O. (2007). PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

Cesa-Bianchi, N. and Gentile, C. (2008). Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge university press.

Chatterjee, S. and Barron, A. (2014). Information theoretic validity of penalized likelihood. In *IEEE International Symposium on Information Theory - Proceedings*, pages 3027–3031. IEEE.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Cover, T. M. (2011). Universal portfolios. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pages 181–209. World Scientific.

Danskin, J. M. (1966). The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292.

Dhillon, P. S., Foster, D., and Ungar, L. H. (2011). Minimum description length penalization for group and multi-task sparse learning. *Journal of Machine Learning Research*, 12(Feb):525–564.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over ℓ_p-balls for ℓ_p-error. *Probability Theory and Related Fields*, 99(2):277–303.

Donsker, M. D. and Varadhan, S. S. (1975). Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Duchi, J., Gould, S., and Koller, D. (2012). Projected subgradient methods for learning sparse Gaussians. *arXiv preprint arXiv:1206.3249*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Dziugaite, G. K. and Roy, D. (2018). Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pages 1376–1385.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Flajolet, P. and Odlyzko, A. (1990). Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Geisser, S. (2017). *Predictive Inference*. Routledge.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373.

Gerchinovitz, S. and Yu, J. Y. (2014). Adaptive and optimal online linear regression on 1-balls. *Theoretical Computer Science*, 519:4–28.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Grünwald, P. (1999). Viewing all models as "probabilistic". In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 171–182. ACM.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT press.

Grünwald, P. D. and Mehta, N. A. (2017). A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. *arXiv preprint arXiv:1710.07732*.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Hedayati, F. and Bartlett, P. L. (2012). The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 7–1.

Hirai, S. and Yamanishi, K. (2011). Efficient computation of normalized maximum likelihood coding for Gaussian mixtures with its applications to optimal clustering. In *IEEE International Symposium on Information Theory - Proceedings*, pages 1031–1035. IEEE.

Hirai, S. and Yamanishi, K. (2013). Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering. *IEEE Transactions on Information Theory*, 59(11):7718–7727.

Hirai, S. and Yamanishi, K. (2017). An upper bound on normalized maximum likelihood codes for Gaussian mixture models. *arXiv preprint arXiv:1709.00925.*

Ito, Y., Oeda, S., and Yamanishi, K. (2016). Selecting ranks and detecting their changes for non-negative matrix factorization using normalized maximum likelihood coding. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 720–728.

Kakade, S. M., Seeger, M. W., and Foster, D. P. (2006). Worst-case bounds for Gaussian process models. In *Advances in Neural Information Processing Systems*, pages 619–626.

Kawakita, M. and Takeuchi, J. (2016). Barron and Cover’ s theory in supervised learning and its application to lasso. In *International Conference on Machine Learning*, pages 1958–1966.

Kleinberg, R., Li, Y., and Yuan, Y. (2018). An alternative view: When does SGD escape local minima? *arXiv preprint arXiv:1802.06175.*

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 369–376.

Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457. Springer.

Komatu, Y. (1955). Elementary inequalities for mills’ ratio. *Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs*, 4:69–70.

Kontkanen, P. (2009). *Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering.* PhD thesis.

Kontkanen, P. and Myllymäki, P. (2007). A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233.

Koolen, W. M., Malek, A., and Bartlett, P. L. (2014). Efficient minimax strategies for square loss games. In *Advances in Neural Information Processing Systems*, pages 3230–3238.

Krichevsky, R. and Trofimov, V. (1981). The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207.

Larsen, J., Hansen, L. K., Svarer, C., and Ohlsson, M. (1996). Design and regularization of neural networks: the optimal use of a validation set. In *Proceedings of the 1996 IEEE Signal Processing Society Workshop*, pages 62–71. IEEE.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation.* Springer Science & Business Media.

Littlestone, N. (1989). From on-line to batch learning. In *Proceedings of the second annual workshop on Computational learning theory*, pages 269–284.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125.

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170. ACM.

Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J., and van Eeuwijk, F. A. (2010). Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. *PloS One*, 5(12):e14147.

Miyaguchi, K. (2017). Normalized maximum likelihood with luckiness for multivariate normal distributions. *arXiv preprint arXiv:1708.01861.*

Miyaguchi, K., Matsushima, S., and Yamanishi, K. (2017). Sparse graphical modeling via stochastic complexity. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 723–731. SIAM.

Miyaguchi, K. and Yamanishi, K. (2018a). Adaptive minimax regret against smooth logarithmic losses over high-dimensional ℓ_1-balls via envelope complexity. *arXiv preprint* *arXiv:1810.03825.*

Miyaguchi, K. and Yamanishi, K. (2018b). High-dimensional penalty selection via minimum description length principle. *Machine Learning*, 107:1283–1302.

Mockus, J., Eddy, W., and Reklaitis, G. (2013). *Bayesian Heuristic Approach to Discrete and Global Optimization: Algorithms, Visualization, Software, and Applications*, volume 17. Springer Science & Business Media.

Mononen, T. and Myllymäki, P. (2007). Fast NML computation for naive Bayes models. In *International Conference on Discovery Science*, pages 151–160. Springer.

Mononen, T. and Myllymäki, P. (2008). Computing the multinomial stochastic complexity in sub-linear time. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models.*

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564.*

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

RA Fisher, M. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368.

Rafiei, M. H. and Adeli, H. (2015). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Rish, I. and Grabarnik, G. (2014). *Sparse Modeling: Theory, Algorithms, and Applications.* CRC press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, pages 416–431.

Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543.

Rissanen, J. (2012). *Optimal estimation of parameters.* Cambridge University Press.

Rissanen, J. and Roos, T. (2007). Conditional NML universal models. In *Proceedings of the 2007 Information Theory and Applications Workshop*, pages 337–341. IEEE.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roos, T. (2008). On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering 2008.*

Roos, T., Myllymaki, P., and Rissanen, J. (2009). MDL denoising revisited. *IEEE Transactions on Signal Processing*, 57(9):3347–3360.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical association*, 46(253):55–67.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Shawe-Taylor, J., Williams, C. K., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522.

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9. ACM.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17.

Steinwart, I. and Christmann, A. (2008). *Support vector machines.* Springer Science & Business Media.

Suzuki, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1397–1406.

Takeuchi, J. and Barron, A. R. (1998). Asymptotically minimax regret by Bayes mixtures. In *IEEE International Symposium on Information Theory - Proceedings.*

Takeuchi, J. and Barron, A. R. (2013). Asymptotically minimax regret by Bayes mixtures for non-exponential families. In *Proceedings of the 2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE.

Takimoto, E. and Warmuth, M. K. (2000). The last-step minimax algorithm. In *International Conference on Algorithmic Learning Theory*, pages 279–290. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tieleman, T. (2008). Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1064–1071. ACM.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838.

Wald, A. (1950). Statistical decision functions.

Watanabe, K. and Roos, T. (2015). Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies. *The Journal of Machine Learning Research*, 16(1):2357–2375.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688.

Wu, T., Sugawara, S., and Yamanishi, K. (2017). Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM.

Xie, Q. and Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533.

Yamanishi, K. (1992). A learning criterion for stochastic rules. *Machine Learning*, 9(2-3):165–203.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44(4):1424–1439.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15(4):915–936.

Zhang, T. (2004). On the convergence of MDL density estimation. In *International Conference on Computational Learning Theory*, pages 315–330. Springer.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

A

Technical Results and Lemmas

A.1 Derivation of Derivatives in Relaxed Stochastic Complexity

In this section, we give proofs for the formulae of derivatives (3.14) and 3.17.

A.1.1 Derivation of (3.14)

By differentiating both sides of (3.13), we have

$$\begin{aligned}
&\frac{\partial}{\partial \Lambda_{ij}} RSC(X; v) \\
&= \frac{n}{2} \frac{\partial}{\partial \Lambda_{ij}} h(S, \Lambda) + \frac{\int \left[-\frac{n}{2} \frac{\partial}{\partial \Lambda_{ij}} h(S', \Lambda) \right] e^{-\frac{n}{2} h(S', \Lambda)} dX'}{\int e^{-\frac{n}{2} h(S', \Lambda)} dX'} \\
&= \frac{n}{2} \left\{ \frac{\partial}{\partial \Lambda_{ij}} h(S, \Lambda) - \mathbb{E}_q \left[\frac{\partial}{\partial \Lambda_{ij}} h(S', \Lambda) \right] \right\}.
\end{aligned}$$

Then, by the following theorem, we have

$$\begin{aligned}
\frac{\partial}{\partial \Lambda_{ij}} h(S, \Lambda) &= \frac{\partial}{\partial \Lambda_{ij}} \min_{\Theta \succ 0} \left\{ \operatorname{tr}[S\Theta] - \ln \det \Theta + \sum_{k,l} \Lambda_{kl} |\Theta_{kl}| \right\} \\
&= \frac{\partial}{\partial \Lambda_{ij}} \left\{ \operatorname{tr}[S\bar{\Theta}] - \ln \det \bar{\Theta} + \sum_{k,l} \Lambda_{kl} |\bar{\Theta}_{kl}| \right\} \\
&= 2 |\bar{\Theta}_{ij}|,
\end{aligned}$$

given $\bar{\Theta} = \bar{\Theta}(S, \Lambda)$. This implies (3.14).

Theorem 17 (Danskin (Danskin, 1966)) Suppose that $f(y, \theta)_{\theta \in \Omega}$ is a continuous function and C^1 with respect to y. Further, assume that Ω is compact and $f(y, \cdot)$ has a unique minimizer. Then, $\bar{f}(y) \overset{\text{def}}{=} \min_{\theta \in \Omega} f(y, \theta)$ is C^1, and $\frac{\partial}{\partial y} \bar{f}(y) = \frac{\partial}{\partial y} f(y, \bar{\theta})$ given $\bar{\theta} = \operatorname{argmin}_{\theta \in \Omega} f(y, \theta)$.

Now, in order to apply the theorem, we confirm that the assumed conditions are satisfied. Let $f(y, \theta)$ be the objective function of the graphical LASSO, where $y = S \otimes \Lambda$, $\theta = \Theta$, and $\Omega = \{\Theta \in \mathbb{R}^{m \times m} \mid \Theta \succ 0\}$. Then, the continuity and y-differentiability of the objective function and the existence of a unique minimizer $\theta = \bar{\Theta}(S, \Lambda)$ hold immediately. Since the objective function is strictly convex with respect to θ, there exist for every y, a closed ε-ball centered at $\bar{\theta}$, say $B_\varepsilon(\bar{\theta})$, and an open δ-ball centered at y, say $\Gamma_\delta(y)$, such that

$$\forall y_1 \in \Gamma_\delta(y),\ \operatorname*{argmin}_{\theta\in\Omega} f(y_1,\theta) \in B_\varepsilon(\bar\theta) \subset \Omega, \tag{A.1}$$

for $\varepsilon, \delta > 0$. Therefore, we can restrict the range of minimization to $B_\varepsilon(\bar\theta)$, which is compact, in a point-wise manner.

A.1.2 Derivation of (3.17)

With trivial calculation, we have

$$\begin{aligned}\frac{\partial}{\partial S_{ij}} \ln q(S) &= \frac{\partial}{\partial S_{ij}} \left\{ \frac{n-m-1}{2} \ln\det S - RSC(X;v) \right\} \\ &= (n-m-1)S_{ij}^{-1} - \frac{n}{2}\frac{\partial}{\partial S_{ij}} h(S,\Lambda).\end{aligned}$$

Then, again by Theorem 17, we have

$$\begin{aligned}\frac{\partial}{\partial S_{ij}} h(S,\Lambda) &= \frac{\partial}{\partial S_{ij}} \min_{\Theta\succ 0} \left\{ \operatorname{tr}[S\Theta] - \ln\det\Theta + \sum_{k<l} \Lambda_{kl} |\Theta_{kl}| \right\} \\ &= \frac{\partial}{\partial S_{ij}} \left\{ \operatorname{tr}\left[S\bar\Theta\right] - \ln\det\bar\Theta + \sum_{k<l} \Lambda_{kl} \left|\bar\Theta_{kl}\right| \right\} \\ &= 2\bar\Theta_{ij},\end{aligned}$$

given $\bar\Theta = \bar\Theta(S,\Lambda)$. This implies (3.17).

A.2 Proof of Utility Lemma of Strong Convexity

In this section, we present a useful property of strongly convex functions. The following lemma introduces a useful lower bound of $\underline{H}$-strongly convex functions.

Lemma 18 Let $\hat\theta = \operatorname{argmin}_{\theta\in\Omega} f(\theta)$, and suppose that $\hat\theta \in \Omega^{\circ}$. Then, for all $\underline{H}$-strongly convex functions $f : \mathbb{R}^p \to \overline{\mathbb{R}}$, we have

$$f(\psi) \geq f(\hat\theta) + \frac{1}{2}\left\|\psi - \hat\theta\right\|_{\underline{H}}^2, \quad \forall\psi \in \mathbb{R}^p.$$

Proof Choose α_0 and define $\mu(\alpha) \overset{\text{def}}{=} \alpha\hat\theta + (1-\alpha)\psi$ ($\alpha_0 \leq \alpha < 1$) such that $\mu(\alpha) \in \Omega$ for all $\alpha \in [\alpha_0, 1)$. Note that this is possible because $\hat\theta$ resides in the interior. By the strong convexity of f and the inequality $f(\hat\theta) \leq f(\mu(\alpha))$, we have

$$\begin{aligned}f(\psi) - f(\hat\theta) \geq f(\psi) - f(\mu(\alpha)) &\geq \alpha\left\langle \xi(\mu(\alpha)), \psi - \hat\theta \right\rangle + \frac{\alpha^2}{2}\left\|\psi - \hat\theta\right\|_{\underline{H}}^2, \\ 0 \geq f(\hat\theta) - f(\mu(\alpha)) &\geq -(1-\alpha)\left\langle \xi(\mu(\alpha)), \psi - \hat\theta \right\rangle + \frac{(1-\alpha)^2}{2}\left\|\psi - \hat\theta\right\|_{\underline{H}}^2.\end{aligned}$$

Then, adding the upper inequality to the lower one multiplied with $\frac{\alpha}{1-\alpha}$ yields

$$f(\psi) - f(\hat\theta) \geq \frac{\alpha}{2}\left\|\psi - \hat\theta\right\|_{\underline{H}}^2.$$

Therefore, taking $\alpha \to 1$ completes the proof. ■

A.3 Lower Bounds and Gaps on Bayesian Minimax Regret

In this supplementary section, we present some technical lemmas and theorems for completeness.

A.3.1 Asymptotic Lower Bound of Shtarkov Complexity for Standard Normal Location Models

We show an asymptotic lower bound of the Shtarkov complexity of standard normal location models.

Lemma 19 Consider the d-dimensional standard normal location model given by $f_X(\theta) = \frac{1}{2}\left\|X-\theta\right\|_2^2 + \frac{d}{2}\ln 2\pi$, where $X \in \mathcal{X} = \mathbb{R}^d$. Let $\gamma = \lambda \left\|\theta\right\|_1$ for $\lambda \geq 0$. Then, we have

$$S(\gamma) \geq d \ln\left(1 + \frac{e^{-\lambda^2/2}}{\sqrt{2\pi}\lambda^3}(1+o(1))\right).$$

Proof By definition of $S(\gamma)$, we have

$$\begin{aligned}
S(\gamma) &= \ln \int e^{-m(f_X+\gamma)}\nu(\mathrm{d}X) \\
&= d\ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \sup_{t\in\mathbb{R}} \exp\left[-\frac{1}{2}(x-t)^2 - \lambda\left|t\right|\right]\mathrm{d}x \\
&= d\ln \frac{1}{\sqrt{2\pi}}\left[\int_{-\infty}^{-\lambda} e^{-\lambda(-\lambda-x)-\frac{\lambda^2}{2}}\mathrm{d}x + \right. \\
&\qquad\qquad \left. \int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\mathrm{d}x + \int_{\lambda}^{\infty} e^{-\lambda(x-\lambda)-\frac{\lambda^2}{2}}\mathrm{d}x\right] \\
&= d\ln\left[2\Phi(\lambda) - 1 + \frac{2e^{-\lambda^2/2}}{\sqrt{2\pi}}\int_0^{\infty} e^{-\lambda x}\mathrm{d}x\right] \\
&= d\ln\left[2\Phi(\lambda) - 1 + \sqrt{\frac{2}{\pi}}\frac{e^{-\lambda^2/2}}{\lambda}\right],
\end{aligned}$$

where $\Phi(\lambda)$ denotes the standard normal distribution function. Now, by Komatu (1955), $\Phi(\lambda)$ is bounded below with $\Phi(\lambda) > 1 - 2\phi(\lambda)/(\sqrt{2+x^2}+x)$ for $\phi(\lambda)$ being the standard normal density, which yields the lower bound of interest after a few lines of elementary calculation. ■

A.3.2 Lower Bound on Minimax Regret of Smooth Models

We describe how we adopt the minimax risk lower bound to show the minimax-regret lower bound.

The concept of the proof is based on Donoho and Johnstone (1994). First, the so-called three-point prior is constructed to approximate the least favorable prior. Then, since the approximate prior violates the ℓ_1-constraint, the degree of the violation is shown to be appropriately bounded to derive a valid lower bound.

The goal of our proof is to establish a lower bound on the minimax regret with respect to logarithmic losses, whereas their proof is about the minimax risk with respect to ℓ_q-loss. Therefore, below we present the proof highlighting (i) an approximate least favorable prior for *logarithmic losses* over ℓ_1-balls, and (ii) the way to bound *regrets* on the basis of risk bounds.

Let $\mathcal{H} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le B\}$ be an ℓ_1-ball. Let $X \sim \mathcal{N}_d[\theta, I_d/L]$ be a d-dimensional normal random variable with mean $\theta \in H$ and precision $L > 0$. We denote the distribution simply by $X \sim \theta$ where any confusion is unlikely. Let $h \in \hat{\mathcal{H}}$ be a predictor associated with any sub-probability distribution $P(\cdot|h) \in \mathcal{M}_+(\mathbb{R}^d)$. For notational simplicity, we may write $f_X(\theta) = \frac{L}{2}\|X-\theta\|_2^2 + \frac{d}{2}\ln\frac{2\pi}{L}$ and $f_X(h) = \ln\frac{\mathrm{d}P(X|h)}{\mathrm{d}\nu}$, where ν is the Lebesgue measure over $\mathbb{R}^d$.

Consider the risk function

$$R_d(h,\theta) \overset{\mathrm{def}}{=} \mathbb{E}_{X\sim\theta}\left[f_X(h) - f_X(\theta)\right]$$

and the Bayes risk function

$$R_d(h,\pi) \overset{\mathrm{def}}{=} \mathbb{E}_{\theta\sim\pi}\left[R_d(h,\theta)\right],$$

where $\pi \in \mathcal{P}(\mathcal{H})$ denotes prior distributions on $\mathcal{H}$. Then, the minimax Bayes risk bounds below the minimax regret:

$$\begin{aligned}
\mathrm{REG}^\star(\mathcal{H}) &= \inf_{h\in\hat{\mathcal{H}}} \sup_{\theta\in\mathcal{H}} \sup_{X\in\mathbb{R}^d} f_X(h) - f_X(\theta) \\
&\ge \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{X\sim\theta}\left[f_X(h) - f_X(\theta)\right] \\
&= \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(h,\pi).
\end{aligned}$$

The minimax theorem states that there exists a saddle point (h^*, π_*) such that

$$R_d(h^*,\pi_*) = \inf_{h\in\hat{\mathcal{H}}} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(h,\pi) = \sup_{\pi\in\mathcal{P}(\mathcal{H})} \inf_{h\in\hat{\mathcal{H}}} R_d(h,\pi) \overset{\mathrm{def}}{=} \sup_{\pi\in\mathcal{P}(\mathcal{H})} R_d(\pi),$$

and π_* is referred to as the least favorable prior. We want to approximate π_* to give an analytic approximation of $R_d(\pi_*)$, which is a lower bound of $\mathrm{REG}^\star(\mathcal{H})$.

Let $F_{\epsilon,\mu} \in \mathcal{P}(\mathbb{R})$ be the three-point prior defined by

$$F_{\epsilon,\mu} = (1-\epsilon)\delta_0 + \frac{\epsilon}{2}\left(\delta_{-\mu} + \delta_\mu\right)$$

for $\epsilon, \mu > 0$. We show that the corresponding achievable Bayes risk $R_1(F_{\epsilon,\mu})$ tends to be the entropy of the prior $F_{\epsilon,\mu}$ in some limit of small ϵ.

Lemma 20 Take $\mu = \mu(\epsilon) = \sqrt{2L^{-1}\ln\epsilon^{-1}}$. Let $H_\epsilon = H(F_{\epsilon,\mu}) = (1-\epsilon)\ln(1-\epsilon)^{-1} + \epsilon\ln 2\epsilon^{-1}$ be the entropy of the prior. Then, we have

$$R_1(F_{\epsilon,\mu}) \sim H_\epsilon \sim \epsilon\ln\frac{1}{\epsilon}$$

as $\epsilon \to 0$. Here, $x \sim y$ denotes the asymptotic equality such that $x/y \to 1$.

Proof First, we show the famous inequality on the entropy given by $R_1(F_{\epsilon,\mu}) \le H_\epsilon$. Let $P(\cdot|h) = \mathbb{E}_{\theta\sim F_{\epsilon,\mu}}P(\cdot|\theta) = (1-\epsilon)P(\cdot|0) + \frac{\epsilon}{2}(P(\cdot|-\mu) + P(\cdot|\mu))$ be the Bayes marginal distribution with respect to $F_{\epsilon,\mu}$. Then, we have

$$\begin{aligned}
H_\epsilon - R_1(F_{\epsilon,\mu}) &= H_\epsilon - R_1(h, F_{\epsilon,\mu}) \\
&= H_\epsilon - \mathbb{E}_{\theta\sim F_{\epsilon,\mu}}\mathbb{E}_{X\sim\theta} \ln \frac{\mathrm{d}P(X|\theta)}{\mathrm{d}P(X|h)} \\
&= H_\epsilon - (1-\epsilon)\mathbb{E}_{P(X|0)} \ln \frac{\mathrm{d}P(X|0)}{\mathrm{d}P(X|h)} - \epsilon\mathbb{E}_{P(X|\mu)} \ln \frac{\mathrm{d}P(X|\mu)}{\mathrm{d}P(X|h)} \\
&= (1-\epsilon)\mathbb{E}_{P(X|0)} \ln \left(1 + \frac{\epsilon}{1-\epsilon}\frac{\mathrm{d}P(X|\mu) + \mathrm{d}P(X|-\mu)}{2\mathrm{d}P(X|0)}\right) \\
&\quad + \epsilon\mathbb{E}_{P(X|\mu)} \ln \left(1 + \frac{1-\epsilon}{\epsilon}\frac{2\mathrm{d}P(X|0) + \mathrm{d}P(X|-\mu)}{\mathrm{d}P(X|\mu)}\right) \\
&\geq 0.
\end{aligned}$$

Now, we show that with the specific value of $\mu = \mu(\epsilon)$, the gap is negligible compared to the entropy itself. Applying Jensen's inequality, we have

$$\begin{aligned}
H_\epsilon - R_1(F_{\epsilon,\mu}) &\leq \epsilon + \epsilon\mathbb{E}_{P(X|\mu)} \ln\left(1 + (1-\epsilon)\left(2e^{-L\mu X} + \epsilon^3 e^{-2L\mu X}\right)\right) \\
&\leq \epsilon(1 + \ln 4 + \mathbb{E}_{P(X|\mu)} \max\{0,\ -2L\mu X\}) \\
&= \epsilon\left(1 + \ln 4 + \mathbb{E}_{Z\sim\mathcal{N}[0,1]} \max\left\{0,\ 2\sqrt{L}\mu(Z - \sqrt{L}\mu)\right\}\right) \\
&\qquad (\because -\sqrt{L}(X-\mu) = Z) \\
&\leq \epsilon\left(1 + \ln 4 + 2\sqrt{L}\mu\epsilon\right) \\
&= \epsilon\left(1 + \ln 4 + 2\epsilon\sqrt{2\ln\frac{1}{\epsilon}}\right) = o(H_\epsilon).
\end{aligned}$$

Thus, we obtain $H_\epsilon \sim R_1(F_{\epsilon,\mu})$. ■

Now we show that the d-th Kronecker product of $F_{\epsilon,\mu}$, $F^d_{\epsilon,\mu}$, can be used to bound the Bayes minimax risk $R_d(\pi_*)$ with appropriate choices of ϵ and μ. To this end, let $\pi_+ = F^d_{\epsilon,\mu} \mid \mathcal{H}$ be the conditional prior restricted over the ℓ_1-ball $\mathcal{H}$.

Lemma 21 Take $\epsilon\mu = (1-c)B/d$ and $\mu = \sqrt{2L^{-1}\ln\epsilon^{-1}}$ for $0 < c < 1$. Then, if $\epsilon \to 0$ and $d\epsilon \to \infty$, we have

$$R_d(\pi_*) \geq R_d(\pi_+) \sim R_d(F^d_{\epsilon,\mu}) \sim d\epsilon \ln\frac{1}{\epsilon}.$$

Proof First, the inequality is trivial from the definition of $R_d(\pi)$. Moreover, the second asymptotic equality immediately follows from Lemma 20.

Now, we consider the first asymptotic equality. Let h be the Bayes minimax predictor with respect to the prior $F_{\epsilon,\mu}$ and h^+ be the one with respect to the conditional prior π_+. Then, we have

$$\begin{aligned}
R_d(F^d_{\epsilon,\mu}) &= R_d(h, F^d_{\epsilon,\mu}) \\
&= \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h,\theta)\right] \\
&= F^d_{\epsilon,\mu}(\mathcal{H})R_d(h,\pi_+) + \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h,\theta)\cdot \mathbb{1}\left\{\theta \notin \mathcal{H}\right\}\right] \\
&\geq F^d_{\epsilon,\mu}(\mathcal{H}) \cdot R_d(\pi_+)
\end{aligned}$$

and

$$\begin{aligned} R_d(F^d_{\epsilon,\mu}) &\le R_d(h^+, F^d_{\epsilon,\mu}) \\ &= \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\right] \\ &= F^d_{\epsilon,\mu}(\mathcal{H})\cdot R_d(\pi_+) + \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\cdot \mathbb{1}\left\{\theta\notin\mathcal{H}\right\}\right]. \end{aligned}$$

Let N be the number of nonzero elements in $\theta \sim F^d_{\epsilon,\mu}$. Then, N is subject to the Binomial distribution $\mathrm{Bin}(d,\epsilon)$. On the other hand, the event $\theta \in \mathcal{H}$ is equal to $\{\|\theta\|_1 \le B\} = \{N \le B/\mu = \mathbb{E}N/(1-c)\}$. Therefore, applying Chebyshev's inequality, we obtain

$$P_d \overset{\text{def}}{=} F^d_{\epsilon,\mu}(\mathcal{H}^c) = \Pr\left\{\frac{N-\mathbb{E}N}{\mathbb{E}N} > \frac{c}{1-c}\right\} \le \frac{(1-c)^2}{c^2 d\epsilon} \to 0.$$

Similarly, we have $\mathbb{E}\left|N-\mathbb{E}N\right|/\mathbb{E}N \to 0$. Now, observe that

$$\begin{aligned} \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^+,\theta)\cdot\mathbb{1}\left\{\theta\notin\mathcal{H}\right\}\right] &\le \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\mathbb{E}_{\varphi\sim\pi_+}\left[R_d(\varphi,\theta)\cdot\mathbb{1}\left\{\theta\notin\mathcal{H}\right\}\right]. \\ &\le 2L\mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\mathbb{E}_{\varphi\sim\pi_+}\left[\left(\|\varphi\|_2^2+\|\theta\|_2^2\right)\cdot\mathbb{1}\left\{\theta\notin\mathcal{H}\right\}\right] \\ &\le 2L\mu^2\mathbb{E}\left[P_dN + N\cdot\mathbb{1}\left\{N>B/\mu\right\}\right] \\ &\qquad (\because \|\theta\|_2^2 = \mu^2 N) \\ &\le 2L\mu^2\mathbb{E}N\left(2P_d + \frac{\mathbb{E}\left|N-\mathbb{E}N\right|}{\mathbb{E}N}\right) \\ &= 4d\epsilon\ln\frac{1}{\epsilon}\left(2P_d + \frac{\mathbb{E}\left|N-\mathbb{E}N\right|}{\mathbb{E}N}\right). \\ &= o(R_d(F^d_{\epsilon,\mu})). \end{aligned}$$

Thus, combining all of the above, we get

$$\begin{aligned} (1+o(1))R_d(\pi_+) &= (1-P_d)R_d(\pi_+) \\ &\le R_d(F^d_{\epsilon,\mu}) \\ &\le (1-P_d)\cdot R_d(\pi_+) + \mathbb{E}_{\theta\sim F^d_{\epsilon,\mu}}\left[R_d(h^*,\theta)\cdot\mathbb{1}\left\{\theta\notin\mathcal{H}\right\}\right]. \\ &= (1-o(1))R_d(\pi_*) + o(R_d(F^d_{\epsilon,\mu})), \end{aligned}$$

which implies the desired asymptotic equality $R_d(F_{\epsilon,\mu}) \sim R_d(\pi_+)$. ■

Summing these, we have an asymptotic lower bound on the minimax regret, which is the same as the upper bound given by the ST prior within a factor of two (see Theorem 14). This implies that both the regret of the ST prior and the Bayes risk of the prior π_+ are tight with respect to the minimax-regret rate except with a factor of two.

Theorem 22 (Lower bound on minimax regret) Suppose that $\omega(1) = \ln(d/\sqrt{L}) = o(L)$. Then, we have

$$\mathrm{REG}^\star(\mathcal{H}) \gtrsim \frac{B}{2}\sqrt{2L\ln\frac{d}{\sqrt{L}}},$$

where $x \gtrsim y$ means that there exists $y' \sim y$ such that $x \ge y'$.

Proof The assumptions of Lemma 21 are satisfied for all $0 < c < 1$, since

$$\begin{aligned}\epsilon \lesssim \epsilon\sqrt{\ln\frac{1}{\epsilon}} &= \frac{1-c}{d}\sqrt{\frac{L}{2}} \to 0,\\ d\epsilon = (1-c)\sqrt{\frac{L}{2\ln\frac{1}{\epsilon}}} &\sim (1-c)\sqrt{\frac{L}{2\ln\frac{d}{\sqrt{L}}}} \to \infty.\end{aligned}$$

Thus, we have

$$\mathrm{REG}^\star(\mathcal{H}) \geq R_d(\pi_*) \gtrsim d\epsilon\ln\frac{1}{\epsilon} \sim (1-c)\,\frac{B}{2}\sqrt{2L\ln\frac{d}{\sqrt{L}}}$$

for all $0 < c < 1$. Slowly moving c towards zero completes the theorem. ■

A.3.3 Existence of Gap between $\mathrm{LREG}^\star$ and $\mathrm{LREG}^{\mathrm{Bayes}}$ under ℓ_1-Penalty

Below we show that under standard normal location models, the Bayesian luckiness minimax regret is strictly larger than the non-Bayesian luckiness minimax regret if γ is nontrivial and has a non-differentiable point. Here, we refer to γ as *trivial* when there exists θ_0 such that $\gamma(\theta) = \infty$ for all $\theta \neq \theta_0$.

Lemma 23 Let $f_X(\theta) = \frac{1}{2}(X-\theta)^2 + \frac{1}{2}\ln 2\pi$ for $X \in \mathbb{R}$ and $\theta \in \mathbb{R}$. Then, for all nontrivial, convex, and non-differentiable penalties $\gamma : \mathbb{R} \to \overline{\mathbb{R}}$,

$$\mathrm{LREG}^\star(\gamma) < \mathrm{LREG}^{\mathrm{Bayes}}(\gamma).$$

Proof Let $\mathcal{F} = \{f_X \mid X \in \mathbb{R}\}$ and recall that $\mathrm{LREG}^{\mathrm{Bayes}}(\gamma) = \inf_{w\in\mathcal{E}(\mathcal{F}_\gamma)} \ln w\left[e^{-\gamma}\right]$ by Theorem 8. Let $\|\cdot\|_\gamma$ be the metric of pre-priors $w \in \mathcal{M}_+(\mathbb{R})$ given by $\|w\|_\gamma = w\left[e^{-\gamma}\right]$. Owing to the continuity of $w \mapsto \ln w\left[e^{-\gamma}\right]$ and the completeness of $\mathcal{E}(\mathcal{F}_\gamma) \subset \mathcal{M}_+(\mathbb{R})$, it suffices to show that there exists no pre-prior $w \in \mathcal{E}(\mathcal{F}_\gamma)$ such that $\ln w\left[e^{-\gamma}\right] = S(\gamma)$. Let us prove this by contradiction. Assume that $\ln w\left[e^{-\gamma}\right] = S(\gamma)$. Observe that

$$\begin{aligned}0 &= w\left[e^{-\gamma}\right] - \exp S(\gamma)\\ &= w\left[\int e^{-f_X-\gamma}\nu(\mathrm{d}X)\right] - \int e^{-m(f_X+\gamma)}\nu(\mathrm{d}X)\\ &= \int\left\{w\left[e^{-f_X-\gamma}\right] - e^{-m(f_X+\gamma)}\right\}\nu(\mathrm{d}X),\end{aligned}$$

which means $w\left[e^{-f_X-\gamma}\right] = e^{-m(f_X+\gamma)}$ for almost every X, since $w \in \mathcal{E}(\mathcal{F}_\gamma)$. Note that $f_X(\theta)$ is continuous with respect to X, and then we have $w\left[e^{-f_X-\gamma}\right] = e^{-m(f_X+\gamma)}$ for all X. After some rearrangement and differentiation, we have

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}X}w\left[e^{-f_X-\gamma+m(f_X+\gamma)}\right]\\ &= w\left[\frac{\mathrm{d}e^{-f_X-\gamma+m(f_X+\gamma)}}{\mathrm{d}X}\right]\\ &= w_\theta\left[(\theta-\theta_X^*)\,e^{-f_X-\gamma+m(f_X+\gamma)}\right],\end{aligned} \tag{A.2}$$

where $\theta^*_X = \arg m(f_X + \gamma)$. Here, we exploited Danskin's theorem at the last equality. One more differentiation gives us

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}X} w_\theta \left[(\theta - \theta^*_X)\, e^{-f_X - \gamma + m(f_X + \gamma)} \right], \\ &= w_\theta \left[\left\{ (\theta - \theta^*_X)^2 - \frac{\mathrm{d}\theta^*_X}{\mathrm{d}X} \right\} e^{-f_X - \gamma + m(f_X + \gamma)} \right] \end{aligned}$$

for all $X \in \mathbb{R}$.

Note that we have $\frac{\mathrm{d}\theta^*_X}{\mathrm{d}X}|_{X=t} = 0$ for any non-differentiable points t of γ. Then, this implies that $w = c\delta_{\theta^*_t}$, where δ_s denotes the Kronecker delta measure. Then, according to (A.2), we have

$$\begin{aligned} 0 &= w_\theta \left[(\theta - \theta^*_X)\, e^{-f_X - \gamma + m(f_X + \gamma)} \right]. \\ &= c\,(\theta^*_t - \theta^*_X)\, e^{-f_X(\theta^*_t) - \gamma(\theta^*_t) + m(f_X + \gamma)}, \end{aligned}$$

which means that $\theta^*_X = \theta^*_t$ is a constant independent of X. However, this contradicts the assumption that γ is nontrivial. ■

As a remark, we note that this lemma is easily extended to a multidimensional exponential family of distributions.