博士論文

Mathematical Structure of Coordination between Individual Learning and Populational Evolution and its Applications

（個の学習と集団の進化が協調する数理構造とその応用）

by

So Nakashima

中島蒼

A Doctor Thesis

Submitted to
the Graduate School of the University of Tokyo
on Decenber 3, 2021
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Information Science and Technology
in Mathematical Informatics

Thesis Supervisor: Tetsuya J. Kobayashi　小林徹也
Associate Professor of Mathematical Informatics

ABSTRACT

This thesis mathematically analyzes the systems in biology and informatics that individual learning and populational evolution coordinate to make decisions by solving an optimization problem. Individual learning and populational evolution are two methods to solve optimization problems. In individual learning, individual agents update a candidate of a solution by processing information. An example of individual learning in biology is the decision-making of individual organisms, whereas an example in information systems is iterative optimization algorithms such as gradient descent and Newton ' s method. In populational evolution, by contrast, a population of replicating agents solves the optimization problem without processing information by themselves: If an agent with a better candidate of a solution has more daughters and the daughters inherit the candidate, the share of better candidates expands in the population. An example in biology of populational evolution is biological evolution by natural selection. An example in information systems is the evolutionary algorithms. In each field, the coordination of individual learning and populational evolution has been considered. This thesis aims to solve the following two problems about coordination in biology and information systems.

The first problem is the acceleration of the evolutionary process by learning from experience. In biology, researchers have considered the evolution of agents that can process information to understand the fitness value of information processing. In this line of research, some studies indicated the possibility that learning from ancestors ' experiences accelerates the evolutionary process. However, it is still unclear whether learning can accelerate the evolutionary process. Also, we do not know what information is sufficient for learning to accelerate the evolutionary process. Furthermore, we do not have a method to quantify the acceleration.

In this thesis, we solve these problems. We first propose ancestral learning and numerically validate that a population of agents with ancestral learning acquires the optimal strategy faster than the conventional evolution. We next theoretically clarify the relationship between ancestral learning and the fitness gradient and prove that learning can accelerate the evolutionary process without communication between agents. To quantify the acceleration, we finally extend Fisher ' s fundamental theorem (FF-thm) of natural selection, which quantifies the speed of the evolutionary process. The extended FF-thm helps our understanding of when and why ancestral learning is beneficial for organisms.

The second problem is theoretical guarantee of the evolutionary algorithms with iterative optimization algorithms. In information systems, researchers have attempted to improve the evolutionary algorithms by incorporating a possibly stochastic iterative optimization algorithm, which is called the memetic algorithm. However, it is difficult to show theoretical guarantees of the memetic algorithm.

In this thesis, we propose a theoretical framework to analyze the memetic algorithm. To focus on the effect of populational evolution, we analyze the memetic algorithm without cross-over, which we call the Branching Algorithm (BA). We first extend FF-thm to the BA, which states that the BA always performs better than the parallel execution of an iterative optimization algorithm.

Although FF-thm gives us a general result that is applicable to all iterative optimization algorithms, it is difficult to calculate the difference in the performance in concrete examples. To resolve this problem, we introduce a more concise framework than FF-thm by introducing the retrospective process from population dynamics. We demonstrate the usefulness of the retrospective process by calculating the performance of the BA with Stochastic Gradient Descent (SGD). We show that the BA with SGD achieves a faster convergence rate for certain not strongly convex functions than the usual SGD because of populational evolution.

The contributions of the thesis are 1-1) the numerical and theoretical validation that learning from experience accelerates the evolutionary process of acquiring an optimal strategy, 1-2) the quantification of the acceleration by the extended FF-thm, and 2) a framework to analyze the memetic algorithms without cross-over by the extended FF-thm and the retrospective process.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Tetsuya J. Kobayashi for his advice and encouragement. He advised not only about my study but also about the attitude towards research. His advice was always to the point and improved my research, writing, and presentation. I have learned much how to make a good research question, how to solve it, and how to present it. I would also thank to the other members of the laboratory for quantitative biology. Discussion with them has deepened my thesis.

I would like to express my gratitude to my collaborators, Takanori Maehara and Yuraro Yamaguchi. Takanori Maehara kindly hired me as a research assistant and financially supported me. I have learned the attitude toward research from a different point of view than my supervisor.

I am also thankful to Japan Society for the Promotion of Science (JSPS) and Japan Science and Technology Agency (JST) for financial support. JST has also provided me opportunities to communicate with other researchers. I would like to express my gratitude to the adopters of the ACT-X “Frontier of mathematics and information science” for a warm and fruitful discussion. I would also thank to my research area advisor, Ichiro Hasuo, for his advice and encouragement.

I am also thankful to the members of the thesis committee, Prof. Hitoshi Iba, Prof. Taiji Suzuki, Prof. Taro Toyoizumi, and Prof. Hiromichi Nagao for kindly participating and discussion. They helped me to improve the thesis via the thesis defense.

Contents

Chapter 1

Introduction

Decision-making appears both in science and in engineering. In biological science, organisms change their traits to survive in surrounding environments [121, 122, 118, 59]. In engineering, decision-making is of course an important topic of information systems to address scheduling, routing, resource allocation, portfolio design, and so on [97, 21, 17]. In addition, we can regard statistical estimation as a decision of parameters [9]. Decision making in each field is often modeled as mathematical optimization by employing the framework of decision theory [9]. In biology, we can sometimes explain the behavior of organisms by the properties of the optimal solutions. In information systems, such formulation helps us to design algorithms to solve the desired problems.

To make an optimal decision, we have two approaches to solve the mathematical optimization: *individual learning* and *populational evolution* (Figure 1.1). In individual learning, an individual agent solves the mathematical optimization to update a candidate of a solution by processing information. In biology, individual learning corresponds to the phenomenon that organisms actively change their traits to increase fitness by using signals from the surrounding environment and its experience. For example, bacteria can sense chemical substances and move towards their source, which is called chemotaxis [106]. In information systems, individual learning is iterative optimization algorithms, that recursively update a candidate of the solution, like gradient descent and Newton's method [17]. In populational evolution, by contrast, a population of replicating agents solves the optimization problem without processing information by themselves: When an agent with a better candidate of a solution has more daughters and the daughters inherit the candidate, the share of better candidates expands in the population. An example in biology of populational evolution is biological evolution by natural selection. An example in information systems is the evolutionary algorithms. In addition, *Sequential Monte-Carlo (SMC) methods* utilized populational evolution of replicating particles to solve filtering problems [14].

Researchers have considered the coordination between individual learning and populational evolution in each field. The aim of this thesis is to theoretically analyze the coordination and solve the problem in each field from a unified view point. We first review individual learning, populational evolution, and their coordination in each field (Table 1.1; See also Chapter 2).

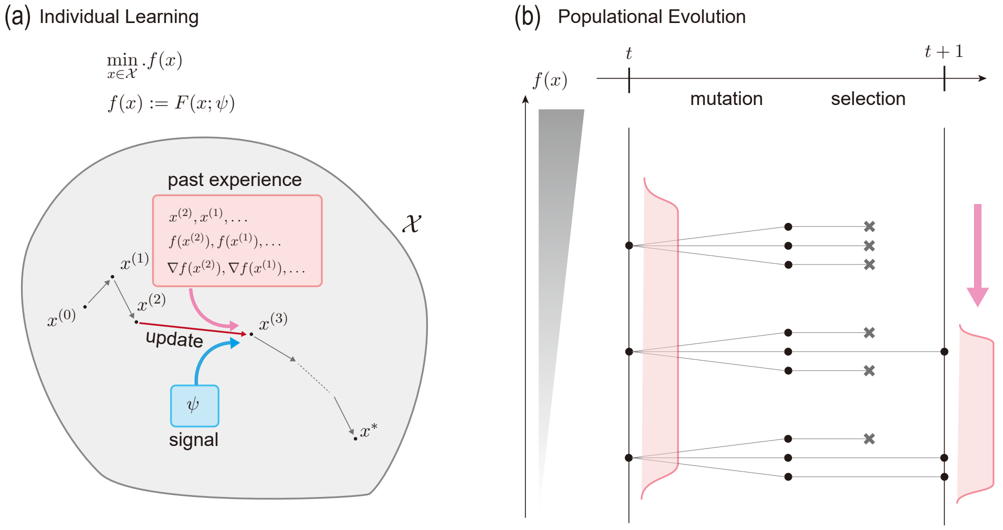

(a) Individual Learning
$\min_{x \in \mathcal{X}} . f(x)$
$f(x) := F(x; \psi)$
past experience
$x^{(2)}, x^{(1)}, \ldots$
$f(x^{(2)}), f(x^{(1)}), \ldots$
$\nabla f(x^{(2)}), \nabla f(x^{(1)}), \ldots$
$\mathcal{X}$
$x^{(0)}$
$x^{(1)}$
$x^{(2)}$
$x^{(3)}$
update
ψ
signal
x^*
(b) Populational Evolution
t
$t+1$
$f(x)$
mutation
selection

Figure 1.1: Schematic illustration of individual learning and populational evolution. Both methods aim to solve a mathematical optimization problem $\min . f(x)$ $(x \in \mathcal{X})$. An optimal solution is denoted by x^*. (a) Schematic illustration of Individual learning. An agent at time t has a candidate $x^{(t)}$ of a solution and update it to decrease the value of the objective function. To update a candidate of a solution, an agent can use one or both of the following two sources of information. One is the experiences, that is, $f(x^{(t')})$, $\nabla f(x^{(t')})$, and so on for $t' \leq t$. The other is the signal ψ that stochastically controls the objective function as $f(x) = F(x; \psi)$. (b) Populational evolution. We consider a population of agents each of which have a candidate of a solution. First, each agent update the solution by mutation. In mutation, an agent does not use the information of objective function. After that, the agents are replicated and selected depending on the value of the objective function: An agent with smaller value of the objective function have more daughters. By this replication competition, the share of solutions with smaller value of the objective function expands in the population. In the figure, red bell-shape objects represent the distribution of the value of the objective function in the population.

	biology	information systems
individual learning	information processing and decision-making by individual organisms	iterative optimization algorithms
populational evolution	biological evolution	evolutionary algorithms and Sequential Monte-Carlo (SMC) methods
coordination	fitness value of individual agent's information processing	memetic algorithms

Table 1.1: Examples of individual learning, populational evolution, and their coordination in each field.

1.1 Individual Learning

In individual learning, individual agents update a candidate of the solution by using two sources of information: *experience* and *signals* (Figure 1.1 (a)). The experience is information that is related to the past candidates $\{x^{(t')}\}_{t' \leq t}$ of a solution. Examples of the experience are of the value of an objective function $f(x^{(t')})$ and gradient $\nabla f(x^{(t')})$ for $t' < t$. An example of individual learning from experience is gradient descent. To explain what is the signal , we consider the case where the objective function depends on some random variable ψ as $f(x) = F(x; \psi)$ for some function F. Under this setting, the signal is the information related to random variable ψ. An example of individual learning from signals is chemotaxis. In chemotaxis, the signal is the sensing of chemical substances.

We add remarks on individual learning in information systems, that is, iterative optimization algorithms. It includes deterministic [17] and stochastic [124] algorithms. Since these algorithms uses the information of objective functions like $f(x^{(t)})$ and $\nabla f(x^{(t)})$, we can regard these algorithms as individual learning from experience. These algorithms are applied to many fields, including statistical estimation and machine learning [14]. Also, we can regard the maximum likelihood estimation as individual learning since it is point estimation instead of distribution estimation. Researchers have tried to show theoretical guarantees of such algorithms [124, 14].

1.2 Populational Evolution

The concept of populational evolution originates from biological evolution by natural selection. To understand and predict the evolutionary process, researchers have developed mathematical models [40, 25] and tools [43, 57, 71, 111, 103, 6, 104], including a *variational principle* and *retrospective process*. In this context, we are interested in how fitness increases by natural selection. A pioneering study by R. A. Fisher [33] proved that the increase of the fitness is proportional to the variance of the fitness. This result is called *Fisher's Fundamental theorem (FF-thm)* of natural selection. This simple relationship is insightful for understanding the impact and efficiency of natural selection.

In natural selection, organisms do not process any information about the surrounding environments by themselves. Even without information processing and according active change of the traits, natural selection can increase the fitness at the population level if the traits in the population are diversified by random mutations. Indeed, an organism with the trait that fits to surrounding environments has more daughters and the share of the trait expands in the population. In this sense, populational evolution is a collective and passive process. In other words, the replication competition of organisms biases the random mutations so that the fitness increases. This bias is a generalization of so-called *survivorship bias.* We use this term in the rest of the thesis.

Populational evolution is applied to information systems. An example is the *evolutionary algorithms* that solve mathematical optimization by using a population of candidates of the solution. The evolutionary algorithms simulate the evolution of the population of candidates of the solution by calculating their fitness from the value of an objective function. As in the biological evolution, the diversity of the candidates is generated by mutation. Several researchers have attempted to show theoretical guarantees of evolutionary algorithms [32]. Theoretical analysis is insightful for tuning hyperparameters of the algorithms. In addition, the theoretical analysis enables us to quantitatively compare the evolutionary algorithms to other optimization algorithms, especially to parallel algorithms, to select the best algorithm in advance of the execution.

Another example is SMC method. SMC method can solve the filtering problems of *hidden Markov models (HMM).* HMM is a time-series model that has a hidden state and an observation at each time. The former is unknown while the latter is observable. In the filtering problem, our objective is the calculation of the posterior distribution of the hidden state in an on-line manner given the observations. SMC methods solve filtering problems by simulating the populational evolution of particles in the hidden state space so that the population becomes an approximation of the posterior distribution.

1.3 Coordination between Individual Learning and Populational Evolution

1.3.1 Acceleration of Evolutionary Process by Learning from Ancestors' Experience

In theoretical biology, the coordination between individual learning and populational evolution has been considered to understand how fitness is increased by information processing by individual organisms. The gain of fitness is called a fitness value of information processing. For example, researchers have quantified the fitness gain by sensing the state of environment by using mutual information [38, 55, 90, 53]. We can regard this information processing as individual learning from signals.

In this context, some papers [115, 54] pointed out the possibility that learning from its ancestors' experience can accelerate evolutionary process. This information processing is an example of individual learning from experience. At first sight, the idea seems to violate the conventional assumption of evolutionary biology that offspring cannot genetically inherit acquired traits including what the parent experienced. However, there is accumulating evidence that an organism can transmit information to the offspring via nongenetic ways, for example, via epigenetic states and culture. The information transmission enables an organism to convey what it learned to the offspring. Therefore, the acceleration of evolutionary process by learning from ancestors' experience does not contradict to the conventional assumption of evolutionary biology. This point of view may be important to understand the variety of phenotypic traits and its inheritance measured by recently developed experimental techniques [105, 23, 72, 116, 113, 41, 66]. We might be able to explain these phenomena from the view point of individual learning from experience.

Since the coordination of learning from ancestors' experience and evolutionary process by natural selection is complicated, we first explain the situation that we consider and clarify individual learning from experience in this context. We consider a population of agents that replicates asexually. Each agent has type and expresses one type in one generation. The type and the state of the environment affect the number of daughters of an agent. Biologically, an agent models an organism and the type models the phenotypic trait. The type is not directly inherited by the daughters. Each agent also has a type-switching strategy that determines the stochastic expression of the types. The stochastic expression is known as bet-hegding and beneficial when the environment is stochastic [99, 27]. We assume that the strategy is heritable. Biologically, the strategy is genetic or epigenetic traits that can be transmitted to the daughters. The types of the daughter are correlated with those of the parent via the inherited strategy.

Since the strategy is heritable, it subjects to natural selection. Therefore, if we have a diversity of strategies in a population, an organism can acquire better strategies during the evolutionary process. In conventional biology, the diversity of strategies is generated by random (mutational or epigenetic) changes that occur when an agent inherits the strategy of the parent. As individual learning from experience, we consider the situation where the inherited strategy is biased by the information that the parent acquired. Specifically, we consider the bias that increases the fitness of the agent. We can regard the conventional random changes as the special case of individual learning from experience by which the daughter does not gain fitness. Therefore, we call the random changes the *zeroth-order* learning rules. Our main focus is the rules with which the daughter can gain fitness on average. To gain fitness, the bias of strategies should be in the direction of the gradient of fitness. Since the gradient is the first-order derivative, we call such rules the *first-order* learning rules.

Our research question is whether such first-order learning rules accelerate the process of acquiring the optimal strategy. A pioneering study by Xue and Leibler [115] considered the acceleration in a model of population dynamics. They showed that an organism can acquire the optimal strategy by a simple learning rule, which we call Xue's rule.

1.3.2 Memetic Algorithms

In information systems, several researchers have tried to improve the performance of the evolutionary algorithms by incorporating local search of solutions by iterative optimization algorithms. In usual evolutionary algorithms, the mutation of the solutions does not use the information of the objective function. However, we might improve the performance of the evolutionary algorithms by using such information. We can locally search a better candidate of the solution around the current candidate of the solution by, for example, exhausting search or gradient descent. We call such a generalized mutation an individual learning from experience in this context. Examples of the evolutionary algorithms with generalized mutation are memetic algorithm [69], covariance matrix adaptation evolutionary strategy (CMA-ES) [39], and information geometric optimization (IGO) [1]. In addition, a recent study [51] has attempted to integrate the evolutionary algorithms and reinforcement learning.

1.3.3 Model Estimation by Particle Methods

Although the typical use of SMC methods are filtering, we can use it to estimate parameters of the model in an on-line manner. There are two approaches to estimate the parameters by SMC methods. In the maximum likelihood approach, we estimate the gradient of the log-likelihood function by particles and then update the parameters to increase the log-likelihood function by gradient descent or expectation-maximization [50]. We can regard this approach as individual learning since we iteratively update a single set of parameters. In Bayesian approach, in contrast, we consider the particles in the parameter space in addition to the hidden state space and solve the parameter estimation as the filtration in the parameter space [50]. The dynamics of the particle in the parameter space is artificial: We use random mutations that do not use the information of likelihood. We can regard this approach as populational evolution since it is an extension of the conventional SMC method to the parameter space. We might improve the performance by combining both approaches. Let us consider the filtering in the parameter space, in which the dynamics of particles use the information of log-likelihood in the form of, for example, the gradient flow induced by the log-likelihood function. We can regard such a method as a coordination between individual learning and populational evolution.

1.4 Unsolved Problems in Each Field

Although the coordination between individual learning and populational evolution has been discussed in each field, there are several unsolved problems. In this section, we state the problems in biology and information systems.

1.4.1 Unsolved Problems in Biology

In evolutionary biology, we have at least three problems about the acceleration of biological evolution by the first-order learning rules. The first problem is whether the first-order rule can actually accelerate the biological evolution or not. Since biological agents cannot implement complicated learning rules by using chemical reactions or connections of neurons, we should investigate whether simple learning rules can accelerate the the evolutionary process. Previous study by Xue and Leibler [115] showed that agents can attain an optimal strategy by simple Xue's rule. However, since the zeroth-order learning rule with populational evolution can attain the optimal strategy, their result is insufficient to state that the first-order rule can accelerate the evolutionary process. We should check whether the population of agents which adopt the first-order rule attain the optimal strategy faster than that of agents which adopt the zeroth-order learning rule.

The second problem is whether an agent can estimate the fitness gradient. Specifically, we do not know what kind of information is sufficient to estimate the gradient. As discussed in Section 2.3.1, the first-order rule should update the strategy in the direction of the fitness gradient. Since the fitness gradient is a property of a population, it is nontrivial that an individual agent can estimate it from ancestors' experience. Such estimation might require communication among agents within the same generation.

The third problem is the quantification of how much the first-order learning rule accelerates the evolutionary process. Since the coordination between individual learning and populational evolution is more complicated than the conventional evolutionary process, a simple relationship like FF-thm in the conventional evolutionary biology is useful to understand when and why learning from experience is beneficial for organisms.

1.4.2 Unsolved Problems in Information Systems

In information systems, we have a problem in theoretical analysis about why the evolutionary algorithms work well. Several researchers have tackled this problem and proposed useful techniques [32]. Indeed, some of the techniques are the basis of our thesis. However, their results still have room for improvement (See Section 2.2.2). Some of their analysis do not predict quantitative behaviors like the convergence rate [45, 31, 24, 79, 96] and others focus on a specific objective function [89, 10, 12, 11].

The issues in the previous attempts are summarized by the following two points. The first issue is that the previous research tried to analyze the whole procedure of the evolutionary algorithms, that is, mutation, recombination (crossover), and selection. This approach is preferable to fully understand the power of the evolutionary algorithms. However, sustained but unsatisfactory previous attempts have revealed that the whole procedure is too difficult to analyze. We should separately analyze the effect of mutation-selection (populational evolution) and that of recombination.

The second issue is the way to compare the performance of the evolutionary algorithms to others. Previous research adopted the following approach: They first tried to show an upper bound of the value of the objective function after the execution of the evolutionary algorithm; They then compare the upper bound with those of other algorithms. However, this approach may underestimate the performance of the evolutionary algorithms. Evolutionary algorithms can incorporate an iterative optimization algorithm $\mathcal{L}$, like the gradient descent, as a substitute for mutation to improve its performance. Indeed, memetic algorithms use an iterative optimization algorithm $\mathcal{L}$ as mutation. We should therefore compare the performance of the memetic algorithm with other algorithms to avoid the underestimation. In this sense, the theoretical analysis of the evolutionary process is related to the coordination between individual learning and populational evolution.

1.5 Our Contribution

1.5.1 Contribution to Biology

In Chapter 4, we address the problems in Section 1.4.1. We solve the first problem by proposing *ancestral learning*, which is a generalization of Xue's rule (Figure 1.2 (a)). Ancestral learning is a simple learning rule that only utilizes the information from the ancestors, which we call *ancestral information*. The candidates of the information carrier are the abundance of proteins and mRNAs [105], epigenetic scars of social defeat stress [23], and the intergenerational effects of space-flight on epigenetic states [116]. We validate that ancestral learning can accelerate the evolutionary process via a numerical simulation. The numerical simulation showed that a population of agents with ancestral learning acquires the optimal strategy faster than that of agents with the zero-th order rules.

We next solve the second problem by characterizing sufficient information to estimate the fitness gradient. We show that ancestral information used in ancestral information is sufficient to estimate the fitness gradient. Since ancestral learning does not utilize communication among agents, this result indicates that an agent can estimate the gradient without communication. We in addition show that ancestral learning updates the strategy in the direction of the fitness gradient (Figure 1.2 (b)). This result implies that ancestral learning is indeed a first-order learning rule.

We finally quantify the acceleration of the evolutionary process by learning via extending FF-thm to ancestral learning. Extended FF-thm relates the acceleration to the variance of the fitness among ancestors. This theorem enables us to quantitatively predict the acceleration of the evolutionary process by ancestral learning. Therefore, the theorem is useful to understand when and why ancestral learning is beneficial.

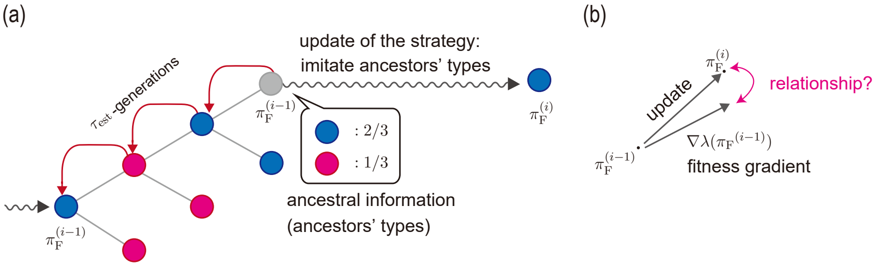


Figure 1.2: Schematic illustration of our contribution to biology. (a) Schematic illustration of ancestral learning. See also Section 4.2. Ancestral learning uses ancestors' types back to τ_{est}-generations. This information is called ancestral information. By using ancestral information, an agents update its strategy ${\pi_{\text{F}}}^{(i-1)}$ to ${\pi_{\text{F}}}^{(i)}$ by imitating the ancestors' types. (b) Remaining problem in this field. To determine whether ancestral learning accelerates the evolutionary process of acquiring optimal strategy, we need to clarify the relationship between the update ${\pi_{\text{F}}}^{(i)} - {\pi_{\text{F}}}^{(i-1)}$ of the strategy and the fitness gradient. In Section 4.4, we prove that the update is in the direction of the fitness gradient.

1.5.2 Contribution to Information Systems

In Chapter 5, we tackle the problems discussed in Section 1.4.2: We propose a theoretical framework to analyze the performance of the evolutionary algorithms by importing techniques from population dynamics (Figure 1.3). Our framework addresses two issues raised in Section 1.4.2. By using the framework, we present an approach to answer why the evolutionary algorithms perform well.

To address the first issue, we focus on the effect of mutation-selection (populational evolution) in this thesis. To focus on the effect of populational evolution, we introduce the memetic algorithm without recombination, which we call a *branching algorithm (BA)* in Section 5.1. To address the second issue, we adopt the following approach: We fix a stochastic iterative optimization algorithm $\mathcal{L}$; We then evaluate how much the performance of the BA with $\mathcal{L}$ differs from that of the parallel execution of $\mathcal{L}$. In other words, we examine whether populational evolution can accelerate individual learning $\mathcal{L}$. We call this approach a *relative evaluation* since we evaluate the difference of the performance. By adopting the relative evaluation, we can evaluate the effect of populational evolution separately from $\mathcal{L}$. We can therefore construct a unified theory that is applicable to all iterative algorithm $\mathcal{L}$. In addition, by setting $\mathcal{L}$ to conventional mutation, we can evaluate the performance of the conventional evolutionary algorithms.

By using this framework, we relatively evaluate the performance of the BA. As the previous approaches [45, 110], we assume that the size of the population is sufficiently large and approximate the BA as population dynamics. This approximation enables us to use techniques in population dynamics, in particular, FF-thm. We extend FF-thm for natural selection as in Chapter 4 and prove that the BA with learning rule $\mathcal{L}$ always performs better than the parallel execution of $\mathcal{L}$. In other words, the extended FF-thm reveals that populational evolution can accelerate individual learning.

Although the extended FF-thm is applicable to all learning rules, the difference of the performance predicted by FF-thm is difficult to calculate in concrete examples. FF-thm enables us to calculate the difference by using the distribution of the whole trajectories (paths) of the candidate of solutions updated by

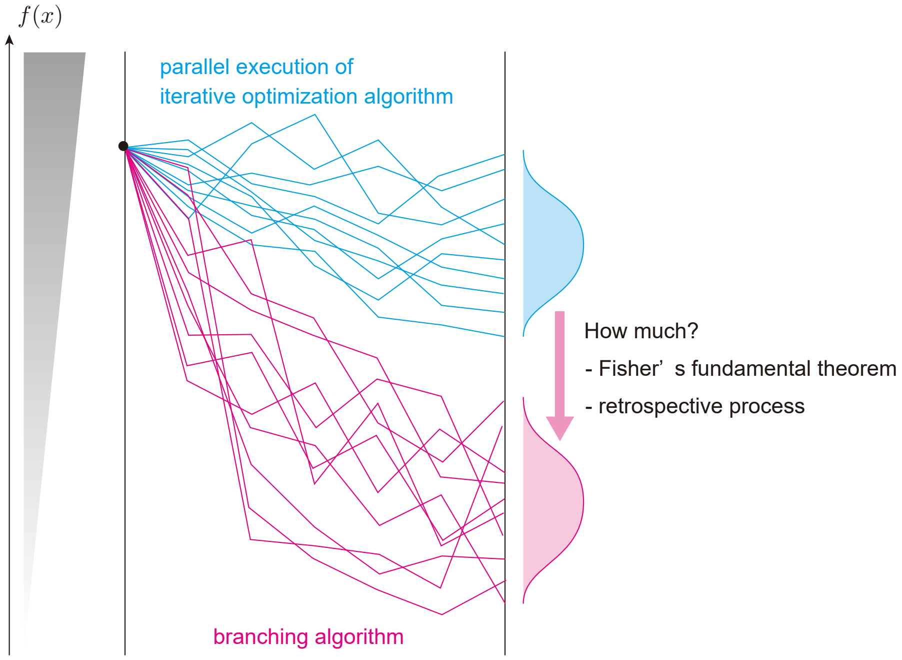


Figure 1.3: Schematic illustration of our contribution to informatics. In Chapter 5, we introduce a variant of the memetic algorithm without recombination, which we call a Branching Algorithm (BA). We compare the performance of the BA with stochastic iterative optimization algorithm $\mathcal{L}$ to that of the parallel execution (PA) of $\mathcal{L}$. For this purpose, we introduce two techniques from population dynamics: Fisher's fundamental theorem and the retrospective process. In the figure, the blue and the red curves show each trajectories of the value of objective function in PA and in BA, respectively. The blue and the red bell-shape represent the distribution of the value of the objective function at the end of the execution of the PA and the BA, respectively.

an iterative algorithm $\mathcal{L}$. However, the huge state space of paths makes the calculation difficult. To resolve this problem, we next introduce the retrospective process as another technique of population dynamics. The retrospective process $\mathcal{L}_{T,\mathrm{B}}^{(t)}$ in this context is a transition matrix of a Markov chain that describes the behavior of $\mathcal{L}$ biased by populational evolution. The retrospective process $\mathcal{L}_{T,\mathrm{B}}^{(t)}$ is relatively easier to calculate than the distribution of the path of the solutions. Moreover, the retrospective process still has sufficient information to evaluate the effect of populational evolution. As a demonstration, we apply the retrospective process to the BA with SGD since SGD is one of the most common stochastic optimization algorithms and that theoretical analysis is well developed. The retrospective process enables us to prove that the BA improves the convergence rate of SGD. The BA with SGD achieves $O(1/T)$-convergence even for not strongly convex functions like $f(x) = \|x\|_3^3$, while the conventional SGD does $O(1/\sqrt{T})$-convergence. We emphasize that the objective of this analysis is to demonstrate that the retrospective process is useful to analyze the evolutionary algorithm. We do not aim to propose the BA with SGD as a practically superior algorithm to the other optimization algorithms.

1.6 Structure of This Thesis

In Chapter 2, we review the relevant research in each field. In Chapter 3, we introduce the notation used in this thesis. In addition, we explain previous results that are necessary for later discussion. From Chapter 4 to Chapter 5, we explain our contribution. In Chapter 4, we solve the problem about the acceleration of the evolutionary process by learning presented in Section 1.4.1. In Chapter 5, we solve the problems about theoretical analysis of the evolutionary algorithms presented in Section 1.4.2. In Chapter 6, we summarize and conclude the thesis.

Chapter 2

Background

In this chapter, we explain the related work and its connection to the thesis.

2.1 Individual Learning

2.1.1 Examples in Biology

We add extra examples of individual learning in biology that we have not mentioned in Chapter 1. An example is an animals' decision making like when to sleep and to feed [64]. In addition, plants may optimize their life history [122]. To explain these phenomena, a viewpoint from mathematical optimization has played an important role. By assuming that the choice of an organism maximizes its fitness, we can explain these phenomena [64, 121, 122, 80, 86, 73, 74]. A stem cell decides the differentiating cell type [59]. The decision of fatal type called apoptosis is also known [118]. We also add extra explanations about the theoretical approach to individual learning in biology. Chemotaxis is modeled as a filtering problem and an optimal control [73, 74]. The division interval of bacteria is explained by optimal scheduling [86].

2.1.2 Examples in Information Systems

A typical example of individual learning in information systems is iterative optimization algorithms for an optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{X} \subseteq \mathbb{R}^d. \end{array} \tag{2.1}$$

An iterative algorithm recursively updates a candidate of the solution $x^{(t)}$ to $x^{(t+1)}$ by a certain rule. We call a candidate of the solution just the solution for simplicity in the rest of the thesis. When a certain stopping condition is satisfied, the algorithm outputs the current candidate $x^{(T)}$ of the solution. We call $x^{(T)}$ the output of the algorithm. Since the iterative optimization algorithm is characterized by the update rule, we sometimes identify the algorithm with the update rule. An example of iterative algorithms is *gradient descant* (GD), which updates the solution in the direction of the negative gradient $-\nabla f(x)$ (Algorithm 3.1 in Section 3.5.2) [17]. *Newton's method* is a variant of the gradient descent that additionally uses the second order derivative $\nabla^2 f(x)$ to update the solution [17].

When we cannot compute $f(x)$ but obtain its noisy estimator, we can use *stochastic optimization* methods. A typical situation is where the objective function is a summation with respect to the data d_i ($i \in [n]$), i.e.,

$$f(x) = \frac{1}{n}\sum_{i=0}^{n-1} l(x; d_i), \tag{2.2}$$

and the size n of the data is too much to compute $f(x)$. Another situation is where the objective function $f(x;\xi)$ depends on some random variable ξ and we want to minimize $\mathbb{E}[f(x;\xi)]$ [91]. Typically, the random variable ξ is the realization of a stochastic observation. An example of stochastic optimization methods is *stochastic gradient descent (SGD)*: We select a small subset of the data $S \subseteq [n]$ randomly and then compute an estimator g_S of the gradient from the selected data as

$$g_S = 1/|S| \cdot \sum_{i \in S} \nabla l(x; d_i). \tag{2.3}$$

The solution is then updated in the negative direction of g_S. Since the size of the selected data is small, we can save the computational time.

To resolve the problems that worsen the performance of SGD, many techniques are proposed. One problem is the variance of $x^{(t)}$ caused by the stochastic fluctuation of g_S with respect to the choice of S. The variance inhibits the solution from converging to a minimizer [1]. To suppress the variance, a technique called *averaging over time* is developed. Instead of the solution $x^{(T)}$ at the end time T, SGD with averaging over time outputs $\bar{x}^{(T)} = \sum_t s^{(t+1)} x^{(t)}$, where $s^{(t)}$ is a weight satisfying $s^{(t)} \geq 0$ and $\sum_{t=1}^{T+1} s^{(t)} = 1$. Another problem is the zigzag movement toward a minimizer around the critical point (i.e., $\nabla f(x) = 0$). When the solution is near the critical point, the gradient $\nabla f(x)$ is close to zero and the direction of $\nabla f(x)$ becomes unstable. This instability makes the convergence slow. Many techniques are proposed to resolve this problem. One is a *momentum method* [87, 93]. A momentum GD updates the solution in the negative direction of the average $1/k \cdot \sum_{s=t-k+1}^{t} \nabla f(x^{(s)})$ of the history of the gradient. The averaging makes the direction of the gradient stable. An sophisticated version of the momentum method is *Nesterov's acceleration* [78]. Nesterov's acceleration is originally proposed for GD and extended to SGD [46, 36, 37].

Many studies have tried to show the theoretical guarantee of the performance of GD and SGD (eg. [94, 3, 16, 70]). They typically try to bound the deviation $f(x^{(T)}) - f(x^*)$ of the value of the objective function at the end of the algorithm from the optimal value. If we have an upper bound $f(x^{(T)}) - f(x^*) = O(h(T))$, we say that the convergence rate is $O(h(T))$. For GD, the convergence rate is exponential $O(e^{-\mu T})$ for some constant $\mu > 0$ under mild assumptions [17]. For SGD, the convergence rate is polynomial instead of $O(e^{-\mu T})$ due to the stochastic fluctuation of the estimator g_S of $\nabla f(x)$. The upper bound is usually proven for the output $\bar{x}^{(T)}$ averaged over time. The convergence rate for convex functions is $O(1/\sqrt{T})$ [102, 77, 101]. Precisely, $\mathbb{E}[f(\bar{x}^{(T)})] - f(x^*) = O(1/\sqrt{T})$. If we can further assume the strong convexity, the convergence rate is $O(1/T)$ [56, 88]. These bounds are known to be optimal up to constant factor [1]. When the variance of the estimator of the gradient is sufficiently small, we can achieve almost $O(1/T^2)$-convergence by Nesterov's acceleration [78, 46, 36, 37].

The convergence rate is also proven for parallel SGD. Let us consider the situation where we execute SGD on N_{size}-computational nodes and then suppress the stochastic fluctuation of the outputs $x_i^{(T)}$ at each node by averaging

[1] A minimizer is the point $x \in \mathcal{X}$ satisfying $f(x) = \min_{x' \in \mathcal{X}} f(x')$.

$1/N_{\text{size}} \cdot \sum_{i=1}^{N_{\text{size}}} x_i^{(T)}$. We call this technique *averaging over paths.* Under this setting, the convergence rate is $O(1/TN_{\text{size}})$ for strongly convex functions when $N_{\text{size}} < \sqrt{T}$ [119]. Therefore, parallel execution on N_{size}-computational nodes makes convergence N_{size}-times faster.

More detailed properties of SGD than the convergence rate are analyzed under the assumption that the estimator g_S of the gradient follows a normal distribution. This assumption is justified when we use a mini-batch S with a sufficiently large size by the central limit theorem. For example, some authors [60, 120, 58] approximate SGD as an Ornstein-Uhlenbeck process under the normality assumption. We use the normality approximation as well in Section 5.4.

2.2 Populational Evolution

2.2.1 Populational Evolution in Biology

Biological evolution is one of the central topics in current biology and studied by both experimental and theoretical approaches. One of the pioneering theoretical results is FF-thm of natural selection [33]. Price generalized FF-thm as Price's equation [85] to incorporate the effect of recombination [84]. FF-thm and Price's equations are proven in various models of population dynamics.

Biological evolution is related to the adaptation (micro-evolution) of organisms. In this context, researchers are interested in how organisms can survive in harsh and randomly ever-changing environments. One strategy for survival is bet-hedging, which is a special form of strategies [99, 27]. In bet-hedging, an organism stochastically expresses types to keep the variety of types in a population. The variety is useful to avoid the extinction of the whole population.

To quantitatively understand the idea, we should calculate the fitness of the strategy. For this purpose, a pathwise formulation, variational principle, and retrospective processes are proposed [43, 57, 71, 111, 103, 6, 104]. In the pathwise formulation, we consider the history (path) of the ancestors' type to represent the fitness. The pathwise formulation is further simplified as a variational problem of the strategy. This representation is called variational principle and related to thermodynamics [103]. The maximizer of the variational problem is called the retrospective process. The retrospective process is useful to calculate the gradient of the fitness. These techniques are extended to various models of population dynamics. For example, the author and two collaborators extended these techniques to population dynamics with age-structure [104]. The retrospective process is useful to estimate the type-switching strategy of an organism from the data at the population level. Since the data at the population level is affected by survivorship bias, we should correct the bias. The retrospective process is useful for this correction. Indeed, the author [76] and other researchers [61, 62] proposed such methods.

2.2.2 Populational Evolution in Information Systems

Examples of populational evolution in information systems are the evolutionary algorithms and particle filters. We review these two topics.

Evolutionary algorithms

The evolutionary algorithms are an application of populational evolution to mathematical optimization. Examples of the evolutionary algorithms are the genetic algorithm and the evolutionary strategies. A simple form of the evolutionary algorithm is as follows: The algorithm uses a population of solutions $\{x_i^{(t)}\}_{i\in[N_{\text{size}}]}$. Each solution is stochastically updated to $x'^{(t)}_i$ depending on $x^{(i)}$ by some rule called *mutation*. Mutation is usually directionless: it does not use the information of the objective function $f(x)$. After that, the population of solutions is further updated to $\{x''^{(t)}_i\}_{i\in[N_{\text{size}}]}$ by some rule called *recombination (cross-over)*. The difference between mutation and recombination is the dependency on other solutions: When updating $x_i^{(t)}$, mutation depends only on $x_i^{(t)}$ while recombination can depend on other solutions $x_j^{(t)}$ $(j \neq i)$ in addition. Then, the population at the next time step is sampled from $\{x'^{(t)}_i\}_{i\in[N_{\text{size}}]}$ with weight $w_i(f(x'^{(t)}_i))$ that is monotonically decreasing with respect to the value of the objective function. This step is called *selection*. Due to populational evolution caused by w_i, we expect that the population will consists of the solutions with small $f(x)$ as t becomes large.

Several researchers have tried to show theoretical guarantees of evolutionary algorithms [32]. Holland [45] invented a schema theory and proved the fundamental theorem of genetic algorithms [18], which turned out to be a special case of Price's equation [2]. However, the schema theory cannot explain the behavior of the genetic algorithms in the limit $t \to \infty$. The Markov chain theory [31, 24, 79, 96] models the genetic algorithm as a Markov chain of the population $\{x^{(t)}\}_{i\in[N_{\text{size}}]} \in \mathcal{X}^{N_{\text{size}}}$ and tries to show the behavior in the limit $t \to \infty$ from the property of the Markov chain. The approach is too complicated to show a transient behavior like the convergence rate. The state space $\mathcal{X}^{N_{\text{size}}}$ of the Markov chain is too large to theoretically analyze. Another approach is direct calculations for specific functions. Rechenberg [89, 98] [2] showed a convergence rate of the evolutionary strategy when the size N_{size} of the population is one and the objective functions are limited to several specific functions, including linear and quadratic functions. Beyer [10, 11, 12] extended the result to the situation where the evaluation of the objective function is noisy and $N_{\text{size}} > 1$ by approximating the offspring distribution. Other approach assumes that the size of the population is infinite and approximates the behavior of the evolutionary algorithms as quadratic dynamical systems [110]. Mühlenbein and Voosen [63] used FF-thm to quantify the effect of mutation-selection, which is separated from the effect of recombination.

We use some of the ideas in the previous approaches in the thesis. We consider the effect of mutation-selection separately from that of recombination as [63]. We assume that the size of the population is infinite and approximate the BA as dynamical system [110]. The difference is that we model the evolutionary algorithm as a population dynamics (5.5) while the previous result models as quadratic dynamical systems. We extend FF-thm to the BA in a different formulation from [45, 18, 2, 63]. We apply the retrospective process to the BA, which is a Markov chain over $\mathcal{X}$ instead of the whole population $\mathcal{X}^{N_{\text{size}}}$ as done in [31, 24, 79, 96]. Although the state space is reduced, the retrospective process incorporates the effect of populational evolution and works usefully to analyze the evolutionary algorithms.

[2] Since this thesis is written in Germany and unavailable in Japan, I have never read it. Two papers [10, 98] pointed out that Rechenberg firstly proved this result in [89].

Sequential Monte-Carlo Methods

SMC methods (also called particle filters) solve filtering, smoothing, and parameter estimation of HMM by employing populational evolution [50]. In filtering of HMM, our aim is to estimate the hidden state $x^{(t)}$ that is not observable and evolves as a Markov chain that follows $\mathbb{P}[\mathbb{X}^{(T)}]$. For this estimation, we can use the set of observations $\mathbb{Y}^{(T)} = \{y^{(0)}, y^{(1)}, \ldots, y^{(T)}\}$ that follows a conditional distribution $K(x^{(t)} \mid y^{(t)})$. Naïve calculation of the conditional probability $\mathbb{P}[x^{(t)} \mid \mathbb{Y}^{(t)}]$ is intractable since the computation of the normalization factor requires $\Omega(\mathbb{X}^t)$-time. If $K(x^{(t)} \mid y^{(t)})$ is conjugate to the time evolution of $\mathbb{X}^{(t)}$, we can recursively compute the conditional probability [14]. However, in general settings, we need to use approximation methods. One of the approximation method is SMC method.

SMC methods use a population of particles, each of which has a value on the hidden state space $\mathcal{X}$. Each particle is updated by simulating the time evolution of $\mathbb{X}^{(T)}$. After that, the particles that are compatible with the current observation $y^{(t)}$ reproduces more daughter particles than the particles that are not compatible. Precisely, a particle with value $x^{(t)}$ has $K(x^{(t)} \mid y^{(t)})$-daughters on average. The total number N_{size} of the particles are kept constant by selection. By this populational evolution of particles, the empirical distribution $1/N_{\text{size}} \sum_{i \in [N_{\text{size}}]} \delta_{x, x^{(t)}}$ becomes an approximation of the conditional distribution $\mathbb{P}[x^{(t)} \mid \mathbb{Y}^{(t)}]$. A similar calculation is applicable to smoothing and parameter estimation of HMM [50].

The measure transformation from $\mathbb{P}[\mathbb{X}^{(T)}]$ to $\mathbb{P}[\mathbb{X}^{(T)} \mid \mathbb{Y}^{(T)}]$ is an example of Feynmann-Kac formula [68]. This interpretation is useful to design computationally efficient SMC methods. For example, an improved algorithm is proposed by using the history of the particles [28]. This technique is similar to the retrospective process in population dynamics. We deepen this connection in Chapter 4.

2.3 Coordination between Individual Learning and Natural Selection

2.3.1 Coordination in Biology

In theoretical biology, researchers have discussed the relationship between individual learning and populational evolution to understand the fitness value of information processing by organisms. An example is sensing of the state of environment. If an organism can sense the state and change its traits accordingly, the organism can survive more easily than without sensing. We can regard this information processing as individual learning from signals of environments. To quantify the fitness value of the sensing, researchers have calculated the difference of the fitness with and without sensing. The difference is characterized by the mutual information between the sensed signal and the state of environment [38, 55, 90, 53]. Although sensing is different from learning from experience that we consider in Chapter 4, these studies revealed an aspect of the coordination between individual learning and populational evolution.

In this context, several researchers [115, 54] pointed out the possibility of the acceleration of the evolutionary process by learning from ancestors' experience. A pioneering study by Xue and Leibler [115] proposed Xue's rule and showed that an agent can acquire the optimal strategy by Xue's rule. In Xue's rule, an organism chooses the phenotypic traits of the parent more frequently than the parent. Our ancestral learning is a generalization of Xue's rule and we deepen their argument in this thesis.

2.3.2 Coordination in Information Systems

One example of coordination in information systems is the memetic algorithms [69]. This kind of extensions are also used as CMA-ES [39] and IGO [1]. Indeed, at the mutation step, these algorithms use a distribution whose parameter is tuned by $f(x)$. Recently, the evolutionary algorithm is combined with reinforcement learning, which is another example of individual learning [51].

Chapter 3

Preliminary

In this chapter, we first introduce the notation used in this thesis. We then outline the previous results that are used later in the thesis.

3.1 Notation

We here summarize the notation used in this thesis.

Let $[n] := \{0, 1, 2, \ldots, n-1\}$. Also, let $[n : m] := \{n, n+1, \ldots, m-1\}$ for $n, m \in \mathbb{N}$ with $n \leq m$. Let $[s, t] := \{u \in \mathbb{R} \mid s \leq u \leq t\}$ and $(s, t) := \{u \in \mathbb{R} \mid s < u < t\}$. For a finite set A, let $|A|$ be its cardinality.

For $x = (x_0, x_1, \ldots, x_{d-1}) \in \mathbb{R}^d$, let $\|x\|_p := \left(\sum_{i=0}^{d-1} x_i^p\right)^{1/p}$ be the l_p-norm. For $p = 2$, we omit the subscript. For a $d \times d$ positive-definite matrix A and $x \in \mathbb{R}^d$, let $\|x\|_A := x^\top A x$. Let I be the identity matrix. For a square matrix A, let $\mathrm{Tr}\,(A)$ be its trace.

For two functions $f, g \colon \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R} \cup \{-\infty, \infty\}$, we say that $f(x) = O(g(x))$ in the limit $x \to a$ if there exists $\delta > 0$ and $M > 0$ such that $|f(x)| \leq Mg(x)$ for all $|x - a| \leq \delta$. We omit a when it is clear from the context. We say that $f(x) = o(g(x))$ if, for all $M > 0$, there exists $\delta > 0$ such that $|f(x)| \leq Mg(x)$ for all $|x - a| \leq \delta$. We say that $f(x) = \Omega(g(x))$ if $g(x) = O(f(x))$ and that $f(x) = \omega(g(x))$ if $g(x) = o(f(x))$. We say that $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$.

Let $t \in [T+1]$ be an index of time. For a set $\mathcal{X}$, a *path* $\mathcal{X}^{(T)}$ on $\mathcal{X}$ is a map $x^{(\cdot)} \colon [T+1] \to \mathcal{X}$ and denoted by $\mathbb{X} = \{x^{(0)}, x^{(1)}, \ldots, x^{(1)}\}$ or $\mathbb{X} = \{x^{(t)}\}_t$. For a path $\mathbb{X}^{(T)}$ and $t, t' \in [T+1]$ with $t \leq t'$, let $\mathbb{X}^{(t:t')} := \{x^{(t)}, x^{(t+1)}, \ldots, x^{(t')}\}$ be a *truncation* of $\mathbb{X}^{(T)}$. We note that $\mathbb{X}^{(0:T)} = \mathbb{X}^{(T)}$. For two paths $\mathbb{X}^{(t:t')} = \{x^{(s)}\}_s$ and $\mathbb{Y}^{(t:t')} = \{y^{(s)}\}_s$ on $\mathbb{R}$, we define an *inner product* of the paths by

$$\left\langle \mathbb{X}^{(t:t')}, \mathbb{Y}^{(t:t')} \right\rangle := \sum_{s=t}^{t'} x^{(s)} y^{(s)}. \tag{3.1}$$

In this paper, we sometimes abbreviate integration $\int_{x \in \mathcal{X}} f(x) \mathrm{d}x$ as $\sum_{x \in \mathcal{X}} f(x)$. For $\epsilon \in \mathbb{R}$ and $f(\epsilon)$ defined over $\epsilon > 0$, let $\lim_{\epsilon \to 0+} f(\epsilon)$ be the one-sided limit of f from the positive half line. We denote by $\delta_{x,y}$ both Kronecker's delta and the delta function.

For a set $\mathcal{X}$ and its σ-algebra Σ, we denote by $\mathcal{P}(\mathcal{X}, \Sigma)$ the set of the probability measures on $(\mathcal{X}, \Sigma)$. We omit Σ when it is clear from context. For a random variable X with distribution p, we denote the expectation and the variance of X by $\mathbb{E}_p[X]$ and $\mathbb{V}_p[X]$, respectively. We omit p from the subscript when it is clear from the context. Let $\mathcal{N}(\mu, \Sigma)$ be the multivariate normal distribution with mean vector μ and covariance matrix Σ. We abbreviate "independently identically distributed" to "i.i.d.". We write $X \sim_{\mathbb{P}} \nu$ if a random variable X follows a distribution ν under a measure $\mathbb{P}$. We omit $\mathbb{P}$ if it is clear from the context. For two distributions p and p' on $\mathcal{X}$, the *Kullback-Leibler divergence (KL-divergence)* is defined by

$$\mathcal{D}\left[p \| p'\right] := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p'(x)}, \tag{3.2}$$

if the support of p' contains that of p.

For two random variables X and Y whose joint distribution is $p(X, Y)$, we define a *log-covariance* and a *log-variance* by

$$\begin{aligned} \text{log-Cov}_p\left[X, Y\right] &:= \log \frac{\mathbb{E}_p[e^{X+Y}]}{\mathbb{E}_p[e^X]\mathbb{E}_p[e^Y]}, &(3.3)\\ \text{log-}\mathbb{V}_p[X] &:= \text{log-Cov}_p\left[X, X\right], &(3.4)\end{aligned}$$

respectively. We use the term "log-covariance" since

$$\text{log-Cov}_p\left[X, Y\right] = \log \mathbb{E}_p[e^X e^Y] - \log \mathbb{E}_p[e^X]\mathbb{E}_p[e^Y], \tag{3.5}$$

is similar to the definition of the covariance:

$$\text{Cov}_p\left[X, Y\right] = \mathbb{E}_p[XY] - \mathbb{E}_p[X]\mathbb{E}_p[Y]. \tag{3.6}$$

By a direct calculation, we can prove the relationship between the covairance and the log-covariance:

$$\text{log-Cov}_p\left[X, Y\right] = \log\left(1 + \frac{\text{Cov}_p\left[e^X, e^Y\right]}{\mathbb{E}_p[e^X]\mathbb{E}_p[e^Y]}\right). \tag{3.7}$$

3.2 Population Dynamics

We introduce a model of populational evolution that is the basis of the model in Chapter 4. We consider a population dynamics of asexual agents with discrete time $t \in \{0, 1, \ldots, \}$ (Figure 4.1). Let $x^{(t)} \in \mathcal{X}$ and $y^{(t)} \in \mathcal{Y}$ be the time of an agent and the state of the environment at time t. For simplicity, we call the state of the environment the environmental state. We assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite. Each agent has a strategy $\pi_{\mathrm{F}}(x) \in \mathcal{P}(\mathcal{X})$. The environmental state $y^{(t)}$ follows a distribution $Q(y^{(t)})$ independently. An agent first determines its type by $\pi_{\mathrm{F}}(x)$. An agent then reproduces the daughters depending on its type and the environmental state. An agent with type x reproduces $e^{k(x,y^{(t)})}$-daughters on average. We call the term $e^{k(x,y^{(t)})}$ an *individual fitness.* We define a path (history) of the type along a lineage and environmental states from time 0 to time t as $\mathbb{X}^{(t)}$ and $\mathbb{Y}^{(t)}$, respectively. Here, a lineage is the sequence of the ancestors of a specified agent at some time. We suppose that the path of the environmental states is $\mathbb{Y}^{(t)}$. Then, the number of agents in the population at time t is

$$N^{(t)}_{\pi_{\mathrm{F}}}[\mathbb{Y}^{(t-1)}] = \left[\sum_{x \in \mathcal{X}} e^{k(x,y^{(t-1)})} \pi_{\mathrm{F}}(x)\right] N^{(t-1)}_{\pi_{\mathrm{F}}}[\mathbb{Y}^{(t-2)}]. \tag{3.8}$$

Here, $N^{(0)}$ is the initial size of the population. We omit π_{F} and $\mathbb{Y}^{(t-1)}$ when it is clear from the context.

By using this dynamical system, we can define the "fitness" of strategy. A *cumulative fitness* of strategy π_{F} until time t is

$$\Lambda^{(t)}(\pi_{\mathrm{F}} \mid \mathbb{Y}^{(t-1)}) := \log \frac{N_{\pi_{\mathrm{F}}}^{(t)}[\mathbb{Y}^{(t-1)}]}{N_{\pi_{\mathrm{F}}}^{(0)}}. \tag{3.9}$$

A *time-averaged population fitness* of π_{F} is

$$\lambda(\pi_{\mathrm{F}}) := \lim_{t\to\infty} \frac{1}{t} \Lambda^{(t)}(\pi_{\mathrm{F}} \mid \mathbb{Y}^{(t-1)}), \tag{3.10}$$

which exists almost surely and independently of $\mathbb{Y}^{(t)}$ due to the ergodicity of the environmental state. For simplicity, we call $\lambda(\pi_{\mathrm{F}})$ the *population fitness.*

3.2.1 Pathwise Formulation and Variational Principle

To calculate the population fitness and its gradient, we use *pathwise formulation* and *variational principle* [103]. Suppose that the environmental states are $\mathbb{Y}^{(t)}$. We calculate the number of agents at time $t+1$ whose path of ancestors' types is $\mathbb{X}^{(t)}$. By recursively applying (3.8),

$$N_{\pi_{\mathrm{F}}}[\mathbb{X}^{(t)} \mid \mathbb{Y}^{(t)}] = e^{k(\mathbb{X}^{(t)},\mathbb{Y}^{(t)})} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(t)}] N^{(0)}, \tag{3.11}$$

where $k(\mathbb{X}^{(t)},\mathbb{Y}^{(t)}) = \sum_{t'=0}^{t} k(x^{(t')},y^{(t')})$ and $\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(t)}] = \prod_{t'=0}^{t} \pi_{\mathrm{F}}(x^{(t')})$ are the pathwise (historical) individual fitness and pairwise forward probability, respectively. We call (3.11) the pathwise formulation. Under the pathwise formulation, the cumulative population fitness is

$$\Lambda^{(t)}(\pi_{\mathrm{F}} \mid \mathbb{Y}^{(t-1)}) = \log \sum_{\mathbb{X}^{(t-1)} \in \mathcal{X}^t} e^{k(\mathbb{X}^{(t-1)},\mathbb{Y}^{(t-1)})} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(t-1)}]. \tag{3.12}$$

Since $y^{(t)}$ follows $Q(y^{(t)})$ independently, the population fitness is (cf. [100, 52])

$$\lambda(\pi_{\mathrm{F}}) = \mathbb{E}_{Q(y)}\left[\log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right]\right]. \tag{3.13}$$

Since the inside of the expectation in the right hand side is the scaled cumulant generating function, we have the following variational representation of the populational fitness:

Proposition 3.1 (Variational Principle)**.**

$$\lambda(\pi_{\mathrm{F}}) = \mathbb{E}_{Q(y)}\left[\max_{\pi\in\mathcal{P}(\mathcal{X})} \left\{\sum_{x\in\mathcal{X}} k(x,y)\pi(x) - \mathcal{D}\left[\pi \| \pi_{\mathrm{F}}\right]\right\}\right]. \tag{3.14}$$

In addition, the maximizer of the variational problem in the right hand side is

$$\pi_{\mathrm{B}}(x \mid y) = \frac{e^{k(x,y)}\pi_{\mathrm{F}}(x)}{\sum_{x'\in\mathcal{X}} e^{k(x',y)}\pi_{\mathrm{F}}(x')}. \tag{3.15}$$

The variational representation enables us to calculate the gradient of population fitness:

Proposition 3.2.

$$\frac{\partial \lambda(\pi_{\mathrm{F}}(x))}{\partial \pi_{\mathrm{F}}(x)} = \frac{\bar{\pi}_{\mathrm{B}}(x)}{\pi_{\mathrm{F}}(x)}, \tag{3.16}$$

where

$$\bar{\pi}_{\mathrm{B}}(x) := \sum_{y\in\mathcal{Y}} \pi_{\mathrm{B}}(x \mid y) Q(y), \tag{3.17}$$

For the completeness of the thesis, we give a proof in Section 3.7.

3.2.2 Retrospective Process

We give an interpretation of the maximizer (3.15). For this purpose, we define the empirical distribution $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}}$ of the ancestors' types. Let us choose an agent at time T uniformly at random. The empirical distribution $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}}$ of ancestors' type of the chosen agent is

$$j^{\pi_{\mathrm{F}}}_{\mathrm{emp}}(x) := \frac{1}{T}\sum_{t'=0}^{T-1} \delta_{x,x^{(t')}}, \tag{3.18}$$

where $\delta_{x,x^{(t)}}$ is the Kronecker's delta and $x^{(t')}$ is the type of the ancestor of the chosen agent at time t'. The empirical distribution is biased from π_{F} in that $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}} \approx \bar{\pi}_{\mathrm{B}} \neq \pi_{\mathrm{F}}$ when T is sufficiently large. This is an example of the survivorship bias and an interpretation of $\bar{\pi}_{\mathrm{B}}(x)$ and $\pi_{\mathrm{B}}(x \mid y)$. Since an agent with a type better fitted to the environmental state has more daughters, such types are emphasized in the empirical distribution. Therefore, $\bar{\pi}_{\mathrm{B}} \approx j^{\pi_{\mathrm{F}}}_{\mathrm{emp}}$ contains information about populational evolution and is useful to analyze population dynamics.

Let us prove that $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}} \approx \bar{\pi}_{\mathrm{B}}$ when T is sufficiently large. For simplicity, we first consider the following situation where the environment is constant $\mathcal{Y} = \{*\}$. In this case, we can write $e^{k(x,*)}$ as $e^{k(x)}$ since the individual fitness is independent of the environmental states. We also write $\pi_{\mathrm{B}}(x \mid y)$ as $\pi_{\mathrm{B}}(x)$ and specially define that $\bar{\pi}_{\mathrm{B}}(x) = \pi_{\mathrm{B}}(x)$. Since $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}} = 1/T \cdot \sum_{t'=0}^{T-1} \delta_{x,x^{(t')}}$ is a summation of random variables, the law of large number implies that

$$j_{\mathrm{emp}}(x) \approx \mathbb{E}\left[\delta_{x,x^{(t')}}\right], \tag{3.19}$$

when T is sufficiently large. We calculate $\mathbb{E}\left[\delta_{x,x^{(t')}}\right]$. If we choose an agent at time $t'+1$, then the number of the descendants of the chosen agent at time T is independent of $x^{(t')}$. We denote by $u^{(t'+1:T)}$ this constant. We also denote by $N^{(t')}$ the number of the agents in the population at time t'. By using these quantities, we can calculate the number of agents at time T whose ancestor at time t' expresses type x as

$$u^{(t'+1:T)} e^{k(x)} \pi_{\mathrm{F}}(x) N^{(t')}. \tag{3.20}$$

By a similar way, we can calculate the total number of agents at time T as

$$\sum_{x' \in \mathcal{X}} u^{(t'+1:T)} e^{k(x')} \pi_{\mathrm{F}}(x') N^{(t')}. \tag{3.21}$$

By using these equations, we know that

$$\begin{aligned}
\mathbb{E}\left[\delta_{x,x^{(t')}}\right] &= \frac{u^{(t'+1:T)} e^{k(x)} \pi_{\mathrm{F}}(x) N'^{(t')}}{\sum_{x\in\mathcal{X}} u^{(t'+1:T)} e^{k(x)} \pi_{\mathrm{F}}(x) N^{(t')}} \\
&= \frac{e^{k(x)} \pi_{\mathrm{F}}{}^{(i-1)}(x)}{\sum_{x\in\mathcal{X}} e^{k(x)} \pi_{\mathrm{F}}{}^{(i-1)}(x)} && (3.22) \\
&= \pi_{\mathrm{B}}(x). && (3.23)
\end{aligned}$$

This equation and (3.19) indicates that $j^{\pi_{\mathrm{F}}}_{\mathrm{emp}} \approx \pi_{\mathrm{B}}$ when T is sufficiently large. Owing to this fact, we call $\pi_{\mathrm{B}}(x)$ the retrospective process of π_{F} when the environment is constant [43, 6, 35, 103].

We next consider the case where the environment is not constant. We calculate $j_{\mathrm{emp}}^{\pi_{\mathrm{F}}}$ by a similar argument. Since the environmental state follows $Q(y)$ independently, the law of large number implies that

$$j_{\mathrm{emp}}(x) \approx \mathbb{E}_{Q(y)}\left[\mathbb{E}\left[\delta_{x,x^{(t')}} \mid y\right]\right], \tag{3.24}$$

where $\mathbb{E}\left[\delta_{x,x^{(t')}} \mid y\right]$ is the conditional expectation of random variable $\delta_{x,x^{(t')}}$ given that the environmental state $y^{(t)} = y$. We can calculate $\mathbb{E}\left[\delta_{x,x^{(t')}} \mid y\right]$ by a similar argument to (3.23):

$$\mathbb{E}\left[\delta_{x,x^{(t')}} \mid y\right] = \frac{e^{k(x,y)}\pi_{\mathrm{F}}(x)}{\sum_{x' \in \mathcal{X}} e^{k(x',y)}\pi_{\mathrm{F}}(x')} = \pi_{\mathrm{B}}(x \mid y). \tag{3.25}$$

This equation and (3.24) indicates that $j_{\mathrm{emp}}^{\pi_{\mathrm{F}}} \approx \bar{\pi}_{\mathrm{B}}$ when T is sufficiently large. Owing to this fact, we call $\pi_{\mathrm{F}}(x \mid y)$ the retrospective process and $\bar{\pi}_{\mathrm{B}}$ the averaged retrospective process.

3.2.3 Extension to Continuous-Time Age-Structured Models

The author and collaborators extended the variational principles and retrospective processes to continuous-time models. An important difference between discrete-time model (3.8) and continuous-time model are the synchronicity of the timing of replication. In discrete-time model, we implicitly assumed that the replication of all agents in the population synchronizes and occurs ones at every time step t. However, in real data (for example, [41]), the replication does not synchronize and we should consider more realistic models. The continuous-time model is an example of such models. In the model, we consider a waiting time τ for the next replication of each agent. In other words, we consider the age of an agent from its birth and τ is the age that the agent replicates. Therefore, we say this model is age-structured.

In continuous-time age-structured model, we consider the population dynamics defined as follows. We suppose that the environment is constant for simplicity. An agent has one type $x \in \mathcal{X}$ as in (3.8). The type is determined when an agent is born by using strategy $\pi_{\mathrm{F}} \in \mathcal{P}(\mathcal{X})$ [1]. We define an *age* of an agent as the time length from its birth. An agent replicates at age τ, which is called a *division time*. We suppose that τ independently follows a type-dependent distribution $\mu(\tau, x)$. The distribution $\mu(\tau, x)$ can be represented by using type-dependent rate $r(a, x)$ as follows [104]:

$$\mu(\tau) := r(\tau, x)e^{-\int_0^\tau r(a,x)\mathrm{d}a}. \tag{3.26}$$

We can interpret $r(a, x)\delta a$ as the probability that an agent with type x replicates at age $[a, a+\delta a]$ given that it has not replicated until age a. We also suppose that the average number of the offspring of an agent with type x and division time τ is $e^{k(\tau,x)}$.

Under this setting, we consider the time evolution of the number $N^{(t)}(a, x)$ of agent with age a and type x at time t. The number $N^{(t)}(a, x)$ follows *McKendric-von Voerster equation* [67, 109]:

$$\frac{\partial}{\partial t}N^{(t)}(a, x) = \left[-\frac{\partial}{\partial a} - r(a, x)\right] N^{(t)}(a, x), \tag{3.27}$$

[1] In our paper [103], we considered a generalized situation in which the strategy π_{F} depends on the age τ when the parent replicates. We consider a simplified model to make the discussion concise.

where the boundary condition is

$$N^{(t)}(0,x) = \pi_{\mathrm{F}}(x) \sum_{x\in\mathcal{X}} \int_0^\infty e^{k(\tau,x)} r(\tau,x) N^{(t)}(x,\tau)\mathrm{d}\tau. \tag{3.28}$$

The population fitness $\lambda(\pi_{\mathrm{F}})$ in this model is defined as

$$\lambda(\pi_{\mathrm{F}}) := \lim_{t\to\infty} \frac{1}{t} \log \frac{\sum_{x\in\mathcal{X}} \int_0^\infty N^{(t)}(a,x)\mathrm{d}a}{\sum_{x\in\mathcal{X}} \int_0^\infty N^{(0)}(a,x)\mathrm{d}a}. \tag{3.29}$$

The author and collaborators extended variational principle (3.14) to this model [104]. For this purpose, let $\mathcal{T}(\mathcal{X})$ be the set of measures $j(x;\tau,x')$ in $\mathcal{X}\times\mathbb{R}\times\mathcal{X}\to\mathbb{R}$ satisfying the normalization condition:

$$\sum_{x,x'\in\mathcal{X}} \int \tau' j(x;\tau',x')\mathrm{d}\tau' = 1, \tag{3.30}$$

and the *shift-invariance* condition:

$$\sum_{x\in\mathcal{X}} \int_0^\infty j(x;\tau',x')\mathrm{d}\tau' = \sum_{x\in\mathcal{X}} \int_0^\infty j(x';\tau',x)\mathrm{d}\tau' =: g_j(x). \tag{3.31}$$

We define a *rate function* $I_{\mathrm{F}}\colon \mathcal{T}(\mathcal{X})\to\mathbb{R}$ by

$$I_{\mathrm{F}}[j] = \sum_{x,x'\in\mathcal{X}} \int_0^\infty j(x;\tau',x') \log \frac{j(x;x',x')}{\pi_{\mathrm{F}}(x)\mu(\tau,x)g_j(x')}. \tag{3.32}$$

Under this setting, we have the following variational principle.

Proposition 3.3 (Variational Principle in age-structured models [104]).

$$\lambda(\pi_{\mathrm{F}}) = \max_{j\in\mathcal{T}(\mathcal{X})} \left\{ \sum_{x,x'\in\mathcal{X}} k(x;\tau,x')j(x;\tau',x') - I_{\mathrm{F}}[j] \right\}. \tag{3.33}$$

As in the case of discrete-time models, the optimal solution j^* is related to the empirical distribution of the ancestors' types and division times. Let us take an agent at time T uniformly at random. We define the empirical distribution j_{emp} of ancestors' types and division times of the chosen agent by generalizing (3.18):

$$j_{\mathrm{emp}}(\tau,x) := \frac{1}{n} \sum_{n'=0}^{n-1} \delta_{x,x^{(n')}} \delta_{\tau^{(n')}}(\tau), \tag{3.34}$$

where the chosen agent is the n-th generation, $x^{(n')}$ and $\tau^{(n')}$ are the type and the division time of the ancestor of the chosen agent at time n'-generation, respectively, and $\delta_{\tau^{(t')}}(\tau)$ is the delta function. From the explicit form of j^*, we proved the following convergence result.

Proposition 3.4 ([104]).

$$j_{\mathrm{emp}}(\tau,x) \to \mu_{\mathrm{B}}(\tau,x)\pi_{\mathrm{B}}(x), \tag{3.35}$$

as $T\to\infty$, where

$$\mu_{\mathrm{B}}(\tau,x) :\propto e^{k(\tau,x)-\lambda(\pi_{\mathrm{F}})\tau}\mu(\tau,x), \tag{3.36}$$

$$\pi_{\mathrm{B}}(x) :\propto \pi(x) \int_0^\infty e^{k(\tau,x)-\lambda(\pi_{\mathrm{F}})\tau}\mathrm{d}\tau. \tag{3.37}$$

Owing to this property, $j_{\mathrm{emp}}(\tau, x)$ is called the retrospective process of this model.

The retrospective process is useful to analyze the data of the growing population of cells. Such data is available for cells by using microfluidic devices [113, 41]. To analyze such data, we typically interested in statistics of the population all over the time. Suppose that there are N cells in the population in total from time 0 to time T. Let i be the label of cells and x_i and τ_i are the type and the division time of the i-th cell, respectively. Then, we consider the statistics of the population of the form:

$$\mathcal{S}_T[h] := \frac{1}{N} \sum_i h(\tau_i, x_i), \tag{3.38}$$

where h is a certain function. Marguet [61, 62] first considered this problem in a certain age-structured model and the author and collaborators [76] simplified the results to the model specialized but applicable to analyze existing data [41]. Marguet and we showed that

$$\mathcal{S}_T[h] \to \sum_{x \in \mathcal{X}} \int_0^\infty h(\tau, x) \pi(x) \mu_{\mathrm{B}}(\tau) \mathrm{d}\tau, \tag{3.39}$$

as $T \to \infty$ under certain regularity conditions.

3.3 Fisher's Fundamental Theorem of Natural Selection

We review FF-thm of natural selection. FF-thm relates the speed of evolution to the variance of individual fitness in the population. To see this clearly, we consider the following fixed-type population dynamics in a constant environment. The model is a modification of (3.8). We suppose that $\mathcal{Y} = \{*\}$. As in the derivation of (3.23), we write the individual fitness by $e^{k(x)}$ instead of $e^{k(x,*)}$. We suppose that the type of an agent equals that of its parent. Under this setting, the number $N^{(t)}(x)$ of the agents with type x at time t satisfies

$$N^{(t)}(x) = e^{k(x)} N^{(t-1)}(x). \tag{3.40}$$

Since we are interested in summary statistics of the population like the variance of the individual fitness, we introduce the fraction $p^{(t)}(x) := N^{(t)}(x) / \sum_{x' \in \mathcal{X}} N^{(t)}(x')$ of the agents with type x in the population at time t. By (3.40), the fraction $p^{(t)}$ satisfies

$$p^{(t)}(x) = \frac{e^{k(x)} p^{(t-1)}(x)}{\sum_{x' \in \mathcal{X}} e^{k(x)} p^{(t-1)}(x)}. \tag{3.41}$$

One way to measure the speed of evolution in this model is the increase in the averaged individual fitness $\mathbb{E}_{p^{(t)}}\left[e^{k(x)}\right]$. By (3.41), the increase satisfies the following:

Theorem 3.5 (Fisher's Fundamental Theorem of Natural Selection).

$$\begin{aligned} \Delta \mathbb{E}_{p^{(t)}}\left[e^{k(x)}\right] &:= \mathbb{E}_{p^{(t)}}\left[e^{k(x)}\right] - \mathbb{E}_{p^{(t-1)}}\left[e^{k(x)}\right] && (3.42) \\ &= \mathbb{V}_{p^{(t-1)}}\left[e^{k(x)}\right] / \mathbb{E}_{p^{(t-1)}}\left[e^{k(x)}\right]. && (3.43) \end{aligned}$$

The theorem reveals the relationship between the gain of the average and the variance of individual fitness in the population. For the completeness of the thesis, we give a proof in Section 3.7.

3.4 Hidden Markov Models and Sequential Monte-Carlo Methods

We here explain a basic fact about HMM and SMC methods. We first introduce HMM. In HMM, a hidden state $x^{(t)} \in \mathcal{X}$ is a realization of the Markov chain whose transition matrix is $\mathbb{T}_{\mathrm{F}}(x^{(t)} \mid x^{(t-1)})$. The hidden state is not observable. Instead, we observe $y^{(t)}$ which follows $K(x^{(t)} \mid y^{(t)})$ at each time. Let $\theta = \{\mathbb{T}_{\mathrm{F}}, K\}$ be the parameters of the model. The joint distribution of the model is

$$\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}, \mathbb{Y}^{(T)}] = \nu(x^{(0)}) \left(\prod_{t=0}^{T-1} \mathbb{T}_{\mathrm{F}}(x^{(t+1)} \mid x^{(t)}) \right) \prod_{t=0}^{T} K(x^{(t)} \mid y^{(t)}), \tag{3.44}$$

where ν is the initial distribution of $x^{(0)}$ and $\mathbb{X}^{(T)}$ and $\mathbb{Y}^{(T)}$ are the paths of $x^{(t)}$ and $y^{(t)}$, respectively. By definition, the conditional distribution is

$$\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)} \mid \mathbb{Y}^{(T)}] = \frac{\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}, \mathbb{Y}^{(T)}]}{\sum_{\mathbb{X}^{(T)} \in \mathcal{X}^{T+1}} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}, \mathbb{Y}^{(T)}]} \tag{3.45}$$

$$= \frac{\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}] \prod_{t=0}^{T} K(x^{(t)} \mid y^{(t)})}{\sum_{\mathbb{X}^{(T)} \in \mathcal{X}^{T+1}} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}] \prod_{t=0}^{T} K(x^{(t)} \mid y^{(t)})}. \tag{3.46}$$

The log-likelihood $l(\theta)$ to observe $\mathbb{Y}^{(T)}$ is

$$l(\theta) = \log \sum_{\mathbb{X}^{(T)} \in \mathbb{X}^{T+1}} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}, \mathbb{Y}^{(T)}]. \tag{3.47}$$

The gradient of log-likelihood satisfies the following identity:

Proposition 3.6 (Fisher's identity [30])**.**

$$\nabla l(\theta) = \mathbb{E}_{\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(t)} \mid \mathbb{Y}^{(t)}]} \left[\nabla \log \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(t)}] \right]. \tag{3.48}$$

Fisher's identity is useful to estimate the parameters θ by SMC methods. Indeed, the right hand side can be approximated by SMC methods. Therefore, we can update parameters by, for example, gradient descent or EM algorithms. When we use the EM algorithm and $\mathbb{T}_{\mathrm{F}}(x^{(t)} \mid x^{(t-1)})$ is simplified as $\pi_{\mathrm{F}}(x^{(t)})$, the parameter π_{F} is recursively updated by

$$\pi_{\mathrm{F}}{}^{(i)} \leftarrow \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{P}_{\mathrm{F}}[x^{(t)} \mid \mathbb{Y}^{(T)}], \tag{3.49}$$

where the right hand side is calculated for $\pi_{\mathrm{F}}{}^{(i-1)}$ [14].

3.5 Mathematical Optimization

In this section, we summarize the previous results about mathematical optimization.

3.5.1 Convex functions

We here review the definition and the properties of convex functions. Convex functions [17, 124] form a broad class of objective functions that admit theoretical guarantees of the performance of optimization algorithms. Let $\mathcal{X} \subseteq \mathbb{R}^d$. A function $f\colon \mathcal{X} \to \mathbb{R}^d$ is *proper* if $\mathcal{X} \neq \emptyset$. A function $f\colon \mathcal{X} \to \mathbb{R}^d$ is *closed* if the *epigraph* $\{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t\}$ is closed. A set $\mathcal{X}$ is *convex* if $(1-t)x+ty \in \mathcal{X}$ for all $x, y \in \mathcal{X}$ and $t \in [0,1]$. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be *convex* if $\mathcal{X}$ is convex and

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \tag{3.50}$$

for all $x, y \in \mathcal{X}$ and $t \in [0,1]$. In this thesis, a convex function is supposed to be proper and closed unless we explicitly mention.

A convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be *α-strongly convex* ($\alpha \geq 0$) if

$$\frac{\alpha}{2}t(1-t)\|x-y\|^2 + f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \tag{3.51}$$

for all $x, y \in X$ and $t \in [0,1]$. We note that 0-strongly convexity coincides with the original convexity. We call an α-strongly convex function with $\alpha > 0$ just a strongly convex function when α is not important. When f is differentiable, the α-strong convexity is equivalent to the following condition [124]:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x-y\|^2 \leq f(y), \tag{3.52}$$

for any $x, y \in \mathcal{X}$. By adding this inequality to the symmetric condition

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2}\|x-y\|^2 \leq f(x), \tag{3.53}$$

we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha\|x-y\|^2. \tag{3.54}$$

A twice differentiable convex function f is α-strongly convex if and only if the minimum eigenvalue of $\nabla^2 f(x)$ is not less than α [17].

A convex function f is said to be γ-smooth if it is differential for any $x \in \mathbb{R}^d$ and

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma\|x-y\|. \tag{3.55}$$

3.5.2 Gradient Descent and its Variants

In this section, we review GD and its variants. Since Chapter 5 focuses on stochastic optimization, we intensively review SGD. Recall that we consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{X} \subseteq \mathbb{R}^d. \end{array} \tag{3.56}$$

Algorithm 3.1 Gradient Decent

```
Set initial solution $x^{(0)}$.
for $t = 0, 1, \dots,$ do
    $x^{(t+1)} \leftarrow \Pi_{\mathcal{X}}\left(x^{(t)} - \eta^{(t+1)}\nabla f(x^{(t)})\right)$.
    if the solution $x^{(t+1)}$ satisfies the stopping condition. then
        Exit the loop.
    end if
end for
Output $x^{(t+1)}$.
```

Algorithm 3.2 Stochastic Gradient Decent

```
Set initial solution $x^{(0)}$.
for $t = 0, 1, \dots,$ do
    Get an estimator $g^{(t+1)}$ of the gradient.
    $x^{(t+1)} \leftarrow \Pi_{\mathcal{X}} \left(x^{(t)} - \eta^{(t+1)} g^{(t+1)}\right)$.
    if the solution $x^{(t+1)}$ satisfies the stopping condition. then
        Exit the loop.
    end if
end for
Output $x^{(t+1)}$.
```

Algorithm 3.3 Nesterov's Acceleration for Stochastic Gradient Decent

```
Set initial solution $x^{(0)} = \mathbf{0}$.
Set $\mu^{(0)} = \zeta^{(0)} = x^{(0)}$.
for $t = 0, 1, \dots,$ do
    $\mu^{(t+1)} \leftarrow (1 - \omega^{(t+1)})\beta^{(t)} + \omega^{(t+1)}\zeta^{(t)}$.
    Get an estimator $g^{(t+1)}$ of the gradient.
    $\beta^{(t+1)} \leftarrow \Pi_{\mathcal{X}} \left(\beta^{(t+1)} - \eta^{(t+1)} \mu^{(t+1)}\right)$.
    $\zeta^{(t+1)} \leftarrow \zeta^{(t)} - (\eta^{(t+1)}\omega^{(t+1)} + \alpha)^{-1}[\eta^{(t+1)}(\mu^{(t+1)} - \beta^{(t+1)}) + \alpha(\zeta^{(t)} - \mu^{(t+1)})]$.
    if the solution $x^{(t+1)}$ satisfies the stopping condition. then
        Exit the loop.
    end if
end for
Output $x^{(t+1)}$.
```

For later convenience, we suppose that $\mathbf{0} \in \mathcal{X}$.

Gradient descent (GD) (Algorithm 3.1) is an iterative optimization algorithm that uses the gradient $\nabla f(x)$ to update the solution. GD first take an initial solution $x^{(0)}$. At each time step, the algorithm updates the solution by

$$x^{(t+1)} \leftarrow x^{(t)} - \eta^{(t+1)} \nabla f(x^{(t)}). \tag{3.57}$$

where $\eta^{(t)}$ is a constant called a *learning rate.* The learning rate is given as a hyperparameter in advance of the execution of GD or determined by a line search [17]. When we consider the convex constraint $\mathcal{X} \subseteq \mathbb{R}^d$ of the feasible domain, the update rule is modified as

$$x^{(t+1)} \leftarrow \Pi_{\mathcal{X}} \left(x^{(t)} - \eta^{(t+1)} \nabla f(x^{(t)})\right). \tag{3.58}$$

where $\Pi_{\mathcal{X}}$ is the projection onto $\mathcal{X}$ defined by

$$\Pi_{\mathcal{X}}(x) = \operatorname*{argmin}_{x' \in \mathcal{X}} \|x - x'\|. \tag{3.59}$$

The convergence rate of the gradient descent for convex functions is exponential $O(e^{-\mu T})$ under mild assumptions [17].

When the gradient $\nabla f(x)$ is unavailable but its stochastic estimator g is available, we use *Stochastic Gradient Descent (SGD)* (Algorithm 3.2) instead of GD. Typical situations are explained in Section 2.1.2. SGD updates the solution by substituting the gradient $\nabla f(x)$ with its estimator g in (3.58):

$$x^{(t+1)} \leftarrow \Pi_{\mathcal{X}} \left(x^{(t)} - \eta^{(t+1)} g^{(t+1)}\right), \tag{3.60}$$

where $g^{(t+1)}$ is the estimator of the gradient $\nabla f(x^{(t)})$ sampled from some probability measure $\mathbb{P}$. We suppose that $\mathbb{E}_{\mathbb{P}}[g^{(t+1)}] = \nabla f(x^{(t)})$ from the typical construction of g (ex. (2.3)). When we want to see the difference between g and $\nabla f(x)$ clearly, we use the following random variable called a *noise* of the estimator of the gradient:

$$\xi := g - \nabla f(x). \tag{3.61}$$

By using the noise, the update by SGD becomes

$$x^{(t+1)} \leftarrow \Pi_{\mathcal{X}} \left(x^{(t)} - \eta^{(t+1)} (\nabla f(x^{(t)}) + \xi^{(t+1)}) \right). \tag{3.62}$$

We summarize the theoretical guarantees of SGD. To prove theoretical guarantees of SGD, we usually use a technique called averaging over time (See Section 2.1.2). Let $\{s^{(t)}\}_{t=0,1,\dots,T} \in \mathbb{R}^{T+1}$ be a weight satisfying $s^{(t)} \geq 0$ for all t and $\sum_{t=1}^{T+1} s^{(t)} = 1$. In addition, let $\bar{x}^{(T)} := \sum_{t=0}^{T} s^{(t+1)} x^{(t)}$ be the averaged output and x^* be an optimal solution. We prove an upper bound for $\mathbb{E}[f(\bar{x}^{(T)})] - f(x^*)$.

To prove the bound, we introduce two assumptions. We suppose that

$$\mathbb{E}[\|g^{(t)}\|^2] \leq G^2, \tag{3.63}$$

for some constant G^2. In addition, we suppose that

$$\mathbb{E}[\|x^{(t)} - x^*\|^2] \leq D^2. \tag{3.64}$$

for some constant D^2.

Under the setting, we state the theoretical guarantees.

Theorem 3.7 ([102, 124])**.** Suppose that (3.63) and (3.64) hold. For a convex function f, let us choose $\eta^{(t)} = \eta/\sqrt{t}$ and $s^{(t)} = 1/(T+1)$. Then, Algorithm 3.2 yields $\{x^{(t)}\}_t$ such that

$$\mathbb{E}\left[f(\bar{x}^{(T)}) \right] - f(x^*) \leq \frac{\frac{D^2}{\eta} + G^2 \eta}{\sqrt{T+1}}. \tag{3.65}$$

We have a faster convergence rate for strongly convex functions. We first explain the convergence rate measured by l_2-norm from the optimal solution.

Theorem 3.8 ([88])**.** Suppose that (3.63) holds. We also assume that f is α-strongly convex. Let $\eta^{(t)} = 1/\alpha t$. Then, Algorithm 3.2 yields $\{x^{(t)}\}_t$ such that

$$\mathbb{E}[\|x^{(t)} - x^*\|^2] \leq \frac{4G^2}{\alpha^2 (t+1)}, \tag{3.66}$$

for all $t = 0, 1, \dots$.

We next explain the convergence rate of $f(\bar{x}^{(T)})$.

Theorem 3.9 ([56, 124])**.** Suppose that (3.63) holds. We also assume that f is α-strongly convex. Let

$$\eta^{(t)} = \frac{2}{\alpha(t+1)}, \tag{3.67}$$

$$s^{(t)} = \frac{2t}{(T+1)(T+2)}. \tag{3.68}$$

Then, Algorithm 3.2 yields $\{x^{(t)}\}_t$ such that

$$\mathbb{E}[f(\bar{x}^{(T)})] - f(x^*) \leq \frac{2G^2}{\alpha(T+1)}. \tag{3.69}$$

When the variance of the noise $\xi^{(t)}$ is small, then we can almost achieve $O(1/T^2)$-convergence by using Nesterov's acceleration [78] (Algorithm 3.3).

Theorem 3.10 ([46])**.** Let

$$\omega^{(t)} = \sqrt{s^{(t-1)} + \left(\frac{s^{(t-1)}}{2}\right)^2} - \frac{s^{(t-1)}}{2}, \tag{3.70}$$

$$s^{(t)} = \prod_{k=1}^{t}(1 - \omega^{(t)}), \tag{3.71}$$

$$\eta^{(t)} = \gamma + \frac{\alpha}{s^{(t-1)}}. \tag{3.72}$$

Suppose that f is α-strongly convex and γ-smooth. We also assume that (3.64) holds and

$$\mathbb{E}[\|g^{(t)} - \nabla f(x^{(t)})\|^2] \leq \sigma^2. \tag{3.73}$$

Under these assumptions, Algorithm 3.3 yields $\{x^{(t)}\}_t$ such that

$$\mathbb{E}[f(x^{(T)})] - f(x^*) \leq C\left(\frac{\sigma^2}{T\alpha} + \frac{D^2(\alpha+\gamma)}{T^2}\right). \tag{3.74}$$

When the coefficient σ^2 of the leading $O(1/T)$-term is sufficiently small, we can neglect it and the convergence rate is approximately $O(1/T^2)$. One way to reduce σ^2 is using a large mini-batch to calculate $g^{(t)}$ by (2.3). By using a mini-batch whose size is n-times larger, we can reduce σ^2 to σ^2/n.

We reformulate Theorem 3.9 to apply it in Chapter 5. Since we will consider a measure transformation of $\mathbb{P}(\xi^{(t)})$, we explicitly write it in the next theorem. In this theorem, we do not assume that $\mathbb{E}_{\mathbb{P}}[\xi^{(t)}] = 0$.

Theorem 3.11 (Reformulation of [124, 56])**.** Suppose that (3.63) holds for probability measure $\mathbb{P}$. We also assume that

$$\mathbb{E}_{\mathbb{P}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle\right] + \frac{C}{t+2} \geq \mathbb{E}_{\mathbb{P}}[f(x^{(t)})] - f(x^*) + \kappa^{-1}\mathbb{E}_{\mathbb{P}}\left[\|x^{(t)} - x^*\|^2\right], \tag{3.75}$$

for some non-negative constants C and κ. Let

$$\eta^{(t)} = \frac{\kappa}{t+1}, \tag{3.76}$$

$$s^{(t)} = \frac{2t}{(T+1)(T+2)}. \tag{3.77}$$

Then,

$$\sum_{t=0}^{T} s^{(t+1)}\mathbb{E}_{\mathbb{P}}[f(x^{(t)})] - f(x^*) \leq \frac{\kappa G^2 + 2C}{T+1}. \tag{3.78}$$

If $C = 0$ and $\kappa = 2\alpha^{-1}$, then we have Theorem 3.9 since the condition (3.75) follows by taking expectation of (3.52). For the completeness of the paper, we give a proof in Section 3.7.

3.6 Central Limit Theorem for Dependent Variables

In this section, we review the Central Limit Theorem (CLT) and its extension to dependent random variables, which we use in Chapter 5. The CLT states that the average $S_i := 1/T \cdot \sum_{i=0}^{T-1} X_i$ of i.i.d. random variables X_i converges to $\mathcal{N}(\mu, \Sigma)$ in distribution, where μ and Σ are the expectation and the covariance matrix of X_i, respectively. This theorem is extended to dependent random variables. An important extension is the martingale CLT [19]. In this thesis, we use the following two extensions.

The first extension is the Markov chain CLT. Let $\{X^{(t)}\}_{t=0,1,\ldots}$ be a Markov chain on $\mathcal{X}$. We suppose that this Markov chain has a stationary distribution π and $X^{(t)}$ follows π for all t. For a function $g\colon \mathcal{X} \to \mathbb{R}$, let

$$\mu = \mathbb{E}[g(X^{(0)})], \tag{3.79}$$

$$\sigma^2 = \sum_{t=0}^{\infty} \mathrm{Cov}\left[X^{(0)}, X^{(t)}\right]. \tag{3.80}$$

Also, let

$$\hat{\mu}^{(t)} = \frac{1}{t}\sum_{t'=0}^{t-1} g(X^{(t')}). \tag{3.81}$$

Under this setting, the following Markov chain CLT holds. See [47] for a review of the sufficient condition of the Markov Chain CLT.

Theorem 3.12 (Markov Chain Central Limit Theorem (Informal) [47])**.** Under a certain ergodicity condition,

$$\sqrt{t}\hat{\mu}^{(t)} \to \mathcal{N}\left(\mu, \sigma^2\right), \tag{3.82}$$

in distribution as $t \to \infty$.

The second extension is a mixingale CLT. A mixingale is a generalization of a martingale [29]. A sequence $\{X^{(t)}\}_{t=0,1,\ldots,}$ of random variables on $\mathbb{R}$ adapted to a filtration $\mathcal{F}^{(t)}$ is said to be a *mixingale* if it satisfies the following condition [2]: There exists a positive sequence ψ_k such that (1) $\psi_k \to 0$ as $k \to \infty$ and (2) for all $t \geq 1$ and $k \geq 0$,

$$\mathbb{E}[\|X^{(t)} \mid \mathcal{F}^{(t-k)}\|] \leq \psi_k. \tag{3.83}$$

If $\psi_k = 0$, then the mixingale coincides with the martingale difference sequence. When the above ψ_k is $O(n^{-p/2-\epsilon})$ for some $\epsilon > 0$, it is said to be *size* $-p$. Under this setting, we explain the mixingale CLT. Let

$$S^{(t)} = \frac{1}{t}\sum_{t'=0}^{t-1} X^{(t')}. \tag{3.84}$$

We assume that $\mathbb{E}[X^{(t)}] = 0$ and the existence of the variance $\mathbb{E}[S_t^2] = \sigma^2$. Formally, we assume that

$$\mathbb{E}\left[\left.\frac{\left(\sum_{t'=t}^{t+k-1} X_{t'}\right)^2}{k}\right| \mathcal{F}^{(t-m)}\right] \to \sigma^2, \tag{3.85}$$

as $\min(t, k, m) \to \infty$. We also assume that ψ_k in the above definition is size $-1/2$.

[2] Original condition in [65] is weaker. We present a simplified definition in this thesis.

Theorem 3.13 (Mixingale Central Limit Theorem [65])**.** Under the above conditions,

$$\sqrt{t} S^{(t)} \sim \mathcal{N}\left(0, \sigma^2\right). \tag{3.86}$$

3.7 Proofs

3.7.1 Proofs in Section 3.2

Proof of Proposition 3.1. The proof is a special case of [103, 53]. For the completeness of the thesis, we give the proof. For a fixed $y \in \mathcal{Y}$ and an arbitrary distribution π over $\mathcal{X}$,

$$\log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] = \log \sum_{x \in \mathcal{X}} \pi(x) \frac{\pi_{\mathrm{F}}(x)}{\pi(x)} e^{k(x,y)}. \tag{3.87}$$

By applying the Jensen's inequality, we have

$$\log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] \geq \sum_{x \in \mathcal{X}} \pi(x) \left[\log \frac{\pi_{\mathrm{F}}(x)}{\pi(x)} e^{k(x,y)}\right] \tag{3.88}$$

$$= \sum_{x \in \mathcal{X}} \pi(x) \left[k(x,y) - \log \frac{\pi(x)}{\pi_{\mathrm{F}}(x)}\right] \tag{3.89}$$

$$= \sum_{x \in \mathcal{X}} \pi(x) k(x,y) - \mathcal{D}\left[\pi \| \pi_{\mathrm{F}}\right]. \tag{3.90}$$

By substituting $\pi(x)$ with $\pi_{\mathrm{B}}(x \mid y)$, we can see that the equality is attained. Therefore,

$$\log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] = \max_{\pi \in \mathcal{P}(\mathcal{X})} \left\{ \sum_{x \in \mathcal{X}} k(x,y)\pi(x) - \mathcal{D}\left[\pi \| \pi_{\mathrm{F}}\right] \right\}, \tag{3.91}$$

and the maximizer is $\pi_{\mathrm{B}}(x \mid y)$. By averaging the equality with respect to $Q(y)$, we have (3.14). □

Proof of Proposition 3.2. The proof is essentially the same as [103]. Since the maximizer of the right hand side of Eq. (3.91) is $\pi_{\mathrm{B}}(x \mid y)$,

$$\log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] = \sum_{x \in \mathcal{X}} k(x,y)\pi_{\mathrm{B}}(x \mid y) - \mathcal{D}\left[\pi_{\mathrm{B}} \| \pi_{\mathrm{F}}\right]. \tag{3.92}$$

We differentiate both sides with respect to $\pi_{\mathrm{F}}(x)$ while taking into account of the dependence of π_{B} on π_{F}:

$$\frac{\partial}{\partial \pi_{\mathrm{F}}(x)} \log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] \tag{3.93}$$

$$= -\frac{\partial \mathcal{D}\left[\pi_{\mathrm{B}} \| \pi_{\mathrm{F}}\right]}{\partial \pi_{\mathrm{F}}(x)} + \sum_{x' \in \mathcal{X}} \frac{\partial \pi_{\mathrm{B}}(x' \mid y)}{\partial \pi_{\mathrm{F}}(x)} \frac{\partial F[\pi_{\mathrm{B}}]}{\partial \pi_{\mathrm{B}}(x' \mid y)}, \tag{3.94}$$

where $F[\pi] := \sum_{x \in \mathcal{X}} k(x,y)\pi(x) - \mathcal{D}\left[\pi \| \pi_{\mathrm{F}}\right]$. Since π_{B} is the maximizer of the F, the derivative of F by π_{B} is zero and consequently the second term vanishes. Therefore,

$$\frac{\partial}{\partial \pi_{\mathrm{F}}(x)} \log \mathbb{E}_{\pi_{\mathrm{F}}(x)}\left[e^{k(x,y)}\right] = \frac{\pi_{\mathrm{B}}(x \mid y)}{\pi_{\mathrm{F}}(x)}. \tag{3.95}$$

By taking the average with respect to $Q(y)$, we have Eq. (3.16). □

Proof of Theorem 3.5. By a direct calculation,

$$\begin{aligned}
\Delta\mathbb{E}_{p^{(t)}}\left[e^{k(x)}\right] &= \sum_{x\in\mathcal{X}} e^{k(x)}p^{(t)}(x) - \sum_{x\in\mathcal{X}} e^{k(x)}p^{(t-1)}(x) && (3.96)\\
&= \sum_{x\in\mathcal{X}} e^{k(x)} \frac{e^{k(x)}p^{(t-1)}(x)}{\sum_{x'\in\mathcal{X}} e^{k(x')}p^{(t-1)}(x')}\\
&\quad - \sum_{x\in\mathcal{X}} e^{k(x)}p^{(t-1)}(x) && (3.97)\\
&= \frac{\sum_{x\in\mathcal{X}}\left(e^{k(x)}\right)^2 p^{(t-1)}(x) - \left(\sum_{x\in\mathcal{X}} e^{k(x)}p^{(t-1)}(x)\right)^2}{\sum_{x'\in\mathcal{X}} e^{k(x')}p^{(t-1)}(x')} && (3.98)\\
&= \frac{\mathbb{V}_{p^{(t-1)}}\left[e^{k(x)}\right]}{\mathbb{E}_{p^{(t-1)}}\left[e^{k(x)}\right]}. && (3.99)
\end{aligned}$$

□

3.7.2 Proofs in Section 3.5

The following proof is essentially the same as [124, 56]. The following two lemmas are cited from [124].

Lemma 3.14 (Cauchy-Schwartz inequality)**.** For two vectors $a, b \in \mathbb{R}^d$ and $\mu > 0$,

$$|\langle a, b\rangle| \leq \frac{\|a\|^2}{2\mu} + \frac{\mu\|b\|^2}{2}. \quad (3.100)$$

Proof.

$$0 \leq \|a/\sqrt{\mu} - \sqrt{\mu}b\|^2 \leq \|a\|^2/\mu - 2\langle a, b\rangle + \mu\|b\|^2. \quad (3.101)$$

By rewriting it, we have

$$\langle a, b\rangle \leq \frac{\|a\|^2}{2\mu} + \frac{\mu\|b\|^2}{2}. \quad (3.102)$$

By substituting a with $-a$, we have

$$\langle a, b\rangle \geq -\frac{\|a\|^2}{2\mu} - \frac{\mu\|b\|^2}{2}. \quad (3.103)$$

We have the statement of the lemma from these two inequalities. □

Lemma 3.15. For $x, y, z \in \mathbb{R}^d$,

$$\langle x - y, y - z\rangle = -\frac{1}{2}\|x - y\|^2 - \frac{1}{2}\|y - z\|^2 + \frac{1}{2}\|x - z\|^2. \quad (3.104)$$

Proof.

$$\begin{aligned}
\frac{1}{2}\|x - z\|^2 &= \frac{1}{2}\|(x - y) + (y - z)\|^2 && (3.105)\\
&= \frac{1}{2}\|x - y\|^2 + \frac{1}{2}\|y - z\|^2 + \langle x - y, y - z\rangle. && (3.106)
\end{aligned}$$

By rewriting it, we have the statement of this lemma. □

Lemma 3.16. Let $\mathcal{X}$ be a convex set. For any $x, x' \in \mathcal{X}$ and $y \in \mathbb{R}^d$ with $x = \Pi_{\mathcal{X}}(y)$, we have

$$\left\langle x' - x, x - y \right\rangle < 0. \tag{3.107}$$

Proof. Suppose to the contrary that $\langle x' - x, x - y \rangle \geq 0$. By the definition of $\Pi_{\mathcal{X}}$, we have

$$x = \operatorname*{argmin}_{z \in \mathcal{X}} \|z - y\|^2. \tag{3.108}$$

Let $x_t = (1 - t)x + tx'$. Since $\mathcal{X}$ is convex, $x_t \in \mathcal{X}$ for $t \in [0, 1]$. By a direct calculation,

$$\frac{\mathrm{d}}{\mathrm{d}t} \|x_t - y\|^2 = 2 \left\langle x' - x, x - y \right\rangle. \tag{3.109}$$

In contradicts to the fact that x is the unique minimizer. □

Proof of Theorem 3.11. We follow the proof in [56, 124].

By the assumption (3.75),

$$\mathbb{E}_{\mathbb{P}}[f(x^{(t)})] - f(x^*) \leq \mathbb{E}_{\mathbb{P}} \left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \right] - \frac{1}{\kappa} \mathbb{E}_{\mathbb{P}}[\|x^{(t)} - x^*\|^2] + \frac{C}{t+2}. \tag{3.110}$$

We evaluate each terms in (3.110). By Lemma 3.14,

$$\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \tag{3.111}$$
$$\leq \left\langle g^{(t+1)}, x^{(t)} - x^{(t+1)} \right\rangle + \left\langle g^{(t+1)}, x^{(t+1)} - x^* \right\rangle \tag{3.112}$$
$$\leq \frac{\eta^{(t)} \|g^{(t+1)}\|^2}{2} + \frac{1}{2\eta^{(t)}} \|x^{(t)} - x^{(t+1)}\|^2 + \left\langle g^{(t+1)}, x^{(t+1)} - x^* \right\rangle, \tag{3.113}$$

for $t = 1, 2, \ldots, T - 1$. By a similar argument, we have

$$\left\langle g^{(T+1)}, x^{(T)} - x^* \right\rangle \leq \frac{\eta^{(T-1)} \|g^{(T+1)}\|^2}{2} + \frac{1}{2\eta^{(T-1)}} \|x^{(t)} - x^*\|^2. \tag{3.114}$$

Here, we allow $g^{(T+1)}$ that is not used in the algorithm. By Lemma 3.16,

$$0 \leq \frac{1}{\eta^{(t-1)}} \left\langle x^* - x^{(t)}, x^{(t)} - \left(x^{(t-1)} - \eta^{(t)} g^{(t)} \right) \right\rangle \tag{3.115}$$
$$\leq \frac{1}{\eta^{(t-1)}} \left\langle x^* - x^{(t)}, x^{(t)} - x^{(t-1)} \right\rangle + \frac{t}{t+1} \left\langle g^{(t)}, x^* - x^{(t)} \right\rangle \tag{3.116}$$
$$\leq \frac{1}{2\eta^{(t-1)}} \left(\|x^{(t-1)} - x^*\|^2 - \|x^{(t)} - x^*\|^2 - \|x^{(t)} - x^{(t-1)}\|^2 \right) \tag{3.117}$$
$$- \frac{t}{t+1} \left\langle g^{(t)}, x^{(t)} - x^* \right\rangle, \tag{3.118}$$

for $t = 1, 2, \ldots, T$. In the last transformation, we used (3.15). To simplify the notation, we define $\tilde{\eta}^{(T)} = \eta^{(T-1)}$ and $\tilde{\eta}^{(t)} = \eta^{(t)}$ otherwise. By using these two inequalities, (3.110) becomes

$$\mathbb{E}_{\mathbb{P}}[f(x^{(t)})] - f(x^*) - \frac{C}{t+2} \tag{3.119}$$

$$\leq \mathbb{E}_{\mathbb{P}}\left[\frac{\eta^{(t)}\|g^{(t+1)}\|^2}{2} + \frac{1}{2\eta^{(t)}}\|x^{(t)} - x^{(t+1)}\|^2 + \left\langle g^{(t+1)}, x^{(t+1)} - x^*\right\rangle\right. \tag{3.120}$$

$$+\frac{1}{2\eta^{(t-1)}}\left(\|x^{(t-1)} - x^*\|^2 - \|x^{(t)} - x^*\|^2 - \|x^{(t)} - x^{(t-1)}\|^2\right) \tag{3.121}$$

$$\left.-\frac{t}{t+1}\left\langle g^{(t)}, x^{(t)} - x^*\right\rangle - \frac{1}{\kappa}\|x^{(t)} - x^*\|^2\right] \tag{3.122}$$

$$= \mathbb{E}_{\mathbb{P}}\left[\frac{\tilde{\eta}^{(t)}\|g^{(t+1)}\|^2}{2} + \left(\frac{1}{2\tilde{\eta}^{(t-1)}}\|x^{(t-1)} - x^*\|^2 - \frac{1/\tilde{\eta}^{(t-1)} + 2/\kappa}{2}\|x^{(t)} - x^*\|^2\right)\right. \tag{3.123}$$

$$+\left(\frac{1}{2\tilde{\eta}^{(t)}}\|x^{(t+1)} - x^{(t)}\|^2 - \frac{1}{2\tilde{\eta}^{(t-1)}}\|x^{(t)} - x^{(t-1)}\|^2\right) \tag{3.124}$$

$$\left.+\left(\left\langle g^{(t+1)}, x^{(t+1)} - x^*\right\rangle - \frac{t}{t+1}\left\langle g^{(t)}, x^{(t)} - x^*\right\rangle\right)\right], \tag{3.125}$$

for $t = 1, 2, \ldots, T-1$. For $t = T$, we have

$$\mathbb{E}_{\mathbb{P}}[f(x^{(T)})] - f(x^*) - \frac{C}{T+2} \tag{3.126}$$

$$\leq \mathbb{E}_{\mathbb{P}}\left[\frac{\eta^{(T-1)}\|g^{(T)}\|^2}{2} + \frac{1}{2\eta^{(T-1)}}\|x^{(T)} - x^*\|^2 +\right. \tag{3.127}$$

$$+\frac{1}{2\eta^{(T-1)}}\left(\|x^* - x^{(T-1)}\|^2 - \|x^{(T)} - x^*\|^2 - \|x^{(T)} - x^{(T-1)}\|^2\right) \tag{3.128}$$

$$\left.-\frac{T}{T+1}\left\langle g^{(T)}, x^{(T)} - x^*\right\rangle - \frac{1}{\kappa}\|x^{(T)} - x^*\|^2\right] \tag{3.129}$$

$$\leq \mathbb{E}_{\mathbb{P}}\left[\frac{\tilde{\eta}^{(T)}\|g^{(T)}\|^2}{2} + \frac{1}{2\tilde{\eta}^{(T)}}\|x^{(T-1)} - x^*\|^2\right. \tag{3.130}$$

$$\left.-\frac{1}{2\tilde{\eta}^{(T)}}\|x^{(T)} - x^{(T-1)}\|^2 - \frac{T}{T+1}\left\langle g^{(T)}, x^{(T)} - x^*\right\rangle\right]. \tag{3.131}$$

For $t = 0$, we have

$$\mathbb{E}_{\mathbb{P}}[f(x^{(0)})] - f(x^*) - \frac{C}{2} \leq \mathbb{E}_{\mathbb{P}}\left[\left\langle g^{(1)}, x^{(0)} - x^*\right\rangle\right] - \frac{1}{\kappa}\mathbb{E}_{\mathbb{P}}[\|x^{(0)} - x^*\|^2]. \tag{3.132}$$

By summing up (3.125), (3.131), and (3.132) with weight $s^{(t+1)}$ and use the method of differences (telescoping sum), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T} s^{(t+1)}\left(f(x^{(t)}) - f(x^*)\right)\right] && (3.133)\\
&\le \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \frac{s^{(t+1)}\tilde{\eta}^{(t)}\|g^{(t+1)}\|^2}{2}\right] + \sum_{t=0}^{T}\frac{Cs^{(t+1)}}{t+2} && (3.134)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T-1}\frac{s^{(t+2)}/\tilde{\eta}^{(t)} - s^{(t+1)}(1/\tilde{\eta}^{(t-1)} + 2/\kappa)}{2}\|x^{(t)} - x^*\|^2\right] && (3.135)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(0)} - x^*\|^2 - \frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(1)} - x^{(0)}\|^2\right. && (3.136)\\
&\qquad \left. - \frac{s^{(2)}}{2}\left\langle g^{(1)}, x^{(1)} - x^*\right\rangle + s^{(1)}\left\langle g^{(1)}, x^{(0)} - x^*\right\rangle - \frac{s^{(1)}}{\kappa}\|x^{(0)} - x^*\|^2\right] && (3.137)\\
&\le \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \frac{s^{(t+1)}\tilde{\eta}^{(t)}\|g^{(t+1)}\|^2}{2}\right] + \sum_{t=0}^{T}\frac{Cs^{(t+1)}}{t+2} && (3.138)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T-1}\frac{s^{(t+2)}/\tilde{\eta}^{(t)} - s^{(t+1)}(1/\tilde{\eta}^{(t-1)} + 2/\kappa)}{2}\|x^{(t)} - x^*\|^2\right] && (3.139)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(0)} - x^*\|^2 - \frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(1)} - x^{(0)}\|^2\right. && (3.140)\\
&\qquad \left. - s^{(1)}\left\langle g^{(1)}, x^{(1)} - x^{(0)}\right\rangle - \frac{s^{(1)}}{\kappa}\|x^{(0)} - x^*\|^2\right] && (3.141)\\
&\le \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} \frac{s^{(t+1)}\tilde{\eta}^{(t+1)}\|g^{(t+1)}\|^2}{2}\right] + \sum_{t=0}^{T}\frac{Cs^{(t+1)}}{t+2} && (3.142)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T-1}\frac{s^{(t+2)}/\tilde{\eta}^{(t)} - s^{(t+1)}(1/\tilde{\eta}^{(t-1)} + 2/\kappa)}{2}\|x^{(t)} - x^*\|^2\right] && (3.143)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[-s^{(1)}\left\langle g^{(1)}, x^{(1)} - x^{(0)}\right\rangle - \frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(1)} - x^{(0)}\|^2\right]. && (3.144)
\end{aligned}
$$

Here, we specially define that $1/\eta^{(-1)} := 0$. By Lemma 3.14,

$$
\begin{aligned}
-s^{(1)}\left\langle g^{(1)}, x^{(1)} - x^{(0)}\right\rangle - \frac{s^{(2)}}{2\tilde{\eta}^{(0)}}\|x^{(1)} - x^{(0)}\|^2 &\le \frac{(s^{(1)})^2\tilde{\eta}^{(0)}}{2s^{(2)}}\|g^{(1)}\|^2 && (3.145)\\
&\le \frac{s^{(1)}\tilde{\eta}^{(0)}}{2}\|g^{(1)}\|^2. && (3.146)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T} s^{(t+1)}\left(f(x^{(t)}) - f(x^*)\right)\right] && (3.147)\\
&\le \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T} \frac{s^{(t+1)}\tilde{\eta}^{(t)}\|g^{(t+1)}\|^2}{2}\right] + \sum_{t=0}^{T}\frac{Cs^{(t+1)}}{t+2} && (3.148)\\
&\quad + \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T-1}\frac{s^{(t+2)}/\tilde{\eta}^{(t)} - s^{(t+1)}(1/\tilde{\eta}^{(t-1)} + 2/\kappa)}{2}\|x^{(t)} - x^*\|^2\right]. && (3.149)
\end{aligned}
$$

We evaluate three terms in the right hand side. For the first term, we have

$$\mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T}\frac{s^{(t+1)}\tilde{\eta}^{(t)}\|g^{(t+1)}\|^2}{2}\right] \leq \sum_{t=0}^{T}\frac{s^{(t+1)}\tilde{\eta}^{(t)}G^2}{2} \leq \frac{\kappa G^2}{T+1}. \tag{3.150}$$

For the second term, since $(t+1)/(t+2) \leq 1$, we have

$$\sum_{t=0}^{T}\frac{Cs^{(t+1)}}{t+2} \leq \frac{2C}{T+2} \leq \frac{2C}{T+1}. \tag{3.151}$$

For the third term, by the choice of $\eta^{(t)}$ and $s^{(t)}$, we have

$$s^{(t+2)}/\eta^{(t)} - s^{(t+1)}(1/\eta^{(t-1)} + 2/\kappa) = 0. \tag{3.152}$$

Therefore, the third term vanishes. All in all, we have

$$\mathbb{E}_{\mathbb{P}}[f(\bar{x}^{(T)}) - f(x^*)] \leq \frac{\kappa G^2 + 2C}{T+1}. \tag{3.153}$$

□

Chapter 4

Acceleration of Evolutionary Processes by Learning and Extended Fisher's Fundamental Theorem

In this chapter, we solve the problems presented in Section 1.4.1 and explain our contributions presented in Section 1.5.1. See Section 4.11 for the derivation of the equations that we omit from the main text. This chapter is published in Physical Review Research [75].

4.1 Setup

We first introduce a model of population dynamics with learning. In the case where agents learn, we use the following dynamical system instead of (3.8) [1]. The number $N^{(t)}(\pi)$ of the agents with strategy π at time t is

$$N^{(t)}(\pi) = \left[\sum_{x \in \mathcal{X}, \pi' \in \mathcal{P}(\mathcal{X})} \mathcal{L}(\pi \mid \pi') e^{k(x, y^{(t-1)})} \pi'(x) \right] N^{(t-1)}(\pi'). \tag{4.1}$$

Here, $\mathcal{L}$ is a possibly stochastic learning rule which is represented as a transition matrix of Markov chain on $\mathcal{P}(\mathcal{X})$.

The learning rule can depend on the information which organisms can use to learn. In this thesis, we suppose the following source of information. We assume that an agent can access to the types of its ancestors. This assumption models the information transmission to daughters via epigenetic states or culture. Although we do not explicitly consider communication among agents in the same generation, we will show that agents can acquire the acceleration of the evolutionary process without communication in Section 4.4. In addition, we do not assume that the organism can access the environmental states. For further generalization, see the discussion in Section 4.10.

Under this setting, we consider how agents gradually acquire the optimal strategy,

$$\pi_{\mathrm{F}}^{*} = \underset{\pi}{\operatorname{argmax}}\, \lambda(\pi), \tag{4.2}$$

which is unique due to the strong concavity of $\lambda(\pi)$ [2]. The concavity easily follows from (3.13)

[1] Although the order of type-switching, replication, and learning in (4.1) is different from the BA in Chapter 5, the difference is not so important.

[2] If there exist $x, x' \in \mathcal{X}$ such that $x \neq x'$ and $k(x, y) = k(x', y)$ for all $y \in \mathcal{Y}$, then the optimal solution is not unique. We call this situation a degeneration. However, we do not consider the degeneration since we can resolve it by identifying x with x'.

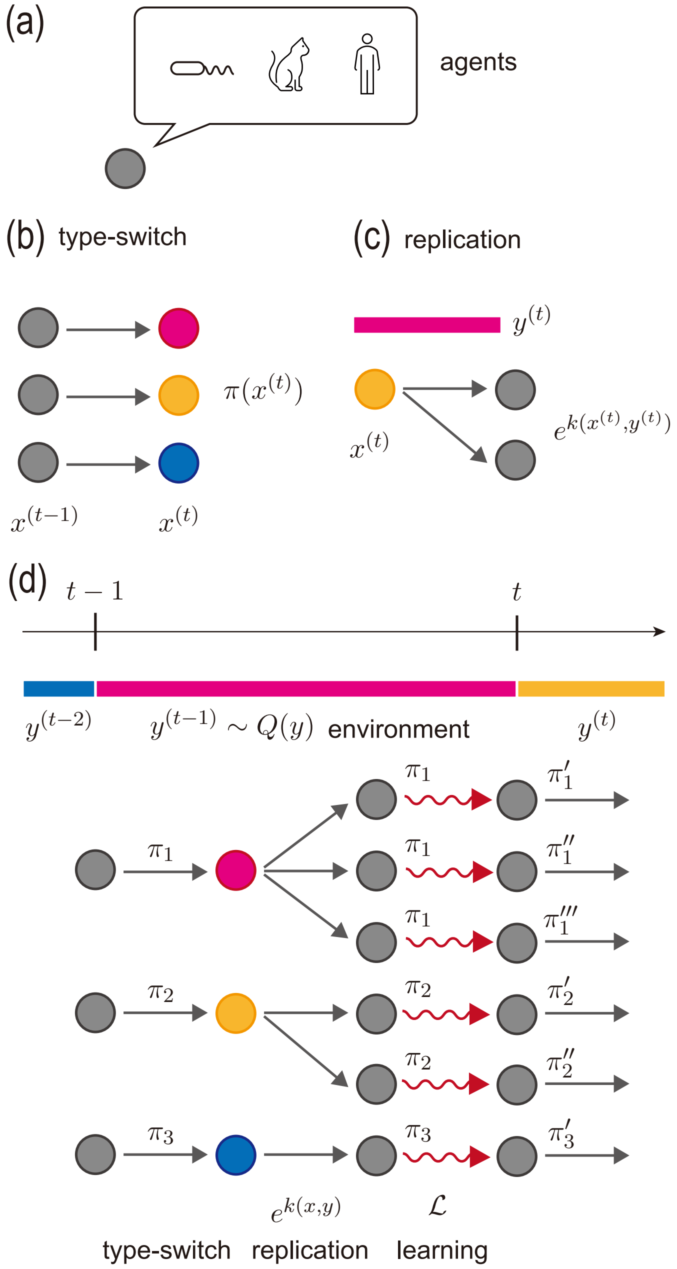


Figure 4.1: Schematic representation of population dynamics of agents that learn. The figure is adopted from [75]. (a) Examples of agents that can replicate and learn. The examples are microbes, animals, and humans. (b–d) Schematic illustration of the model. (b) Type-switching of an agent. Each agent expresses one type $x \in \mathcal{X}$ in one generation. The type of agent is determined by its own strategy $\pi \in \mathcal{P}(\mathcal{X})$. In this figure, the color other than gray represents the type of an agent. (c) Replication of an agent. Each agent at time $t-1$ reproduces $e^{k(x,y^{(t-1)})}$-daughters on average if its type is x and the environmental state is $y^{(t-1)}$. (d) Learning of an agent. After the replication, each agent updates its strategy by using some given learning rule $\mathcal{L}$. The learning rule is a Markov chain of the strategy. The daughter agents inherit the learned strategy.

4.2 Ancestral Learning

We introduce *ancestral learning* and validate that learning can accelerate the evolutionary process. Ancestral learning updates the strategy every τ_{est}-generations. Here, τ_{est} is a hyperparameter called an *update interval.* For simplicity, we suppose that the i-th update occurs at time $t = i\tau_{\mathrm{est}} - 1$ $(i = 1, 2, \dots)$. We specially regard that the initial strategy ${\pi_{\mathrm{F}}}^{(0)}$ is acquired by the 0-th update at time -1. After an agent acquires ${\pi_{\mathrm{F}}}^{(i-1)}$ by the $(i-1)$-th update, the descendants of the agent at time $(i-1)\tau_{\mathrm{est}} \le t \le i\tau_{\mathrm{est}} - 1$ have the same strategy. At time $i\tau_{\mathrm{est}} - 1$, i.e., at the next update, each descendant at time $i\tau_{\mathrm{est}}$ calculates the empirical distribution j_{emp} of the types of its ancestors back to time $(i-1)\tau_{\mathrm{est}}$. Precisely, the empirical distribution is defined by

$$j_{\mathrm{emp}}^{{\pi_{\mathrm{F}}}^{(i-1)}}(x) := \frac{1}{\tau_{\mathrm{est}}} \sum_{t'=(i-1)\tau_{\mathrm{est}}}^{i\tau_{\mathrm{est}}-1} \delta_{x,x^{(t')}}, \tag{4.3}$$

$x^{(t')}$ is the type of the ancestor at time t'. Each descendant then updates the strategy by

$${\pi_{\mathrm{F}}}^{(i)} = (1-\alpha){\pi_{\mathrm{F}}}^{(i-1)} + \alpha j_{\mathrm{emp}}, \tag{4.4}$$

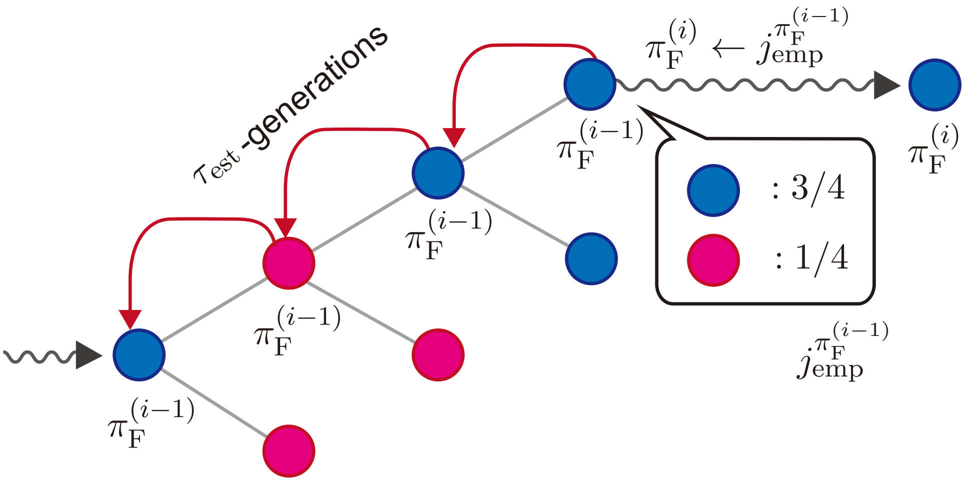
where α is a hyperparameter called a *learning rate.* In this update rule, the updated strategy ${\pi_{\mathrm{F}}}^{(i)}$ is the mixture of ${\pi_{\mathrm{F}}}^{(i-1)}$ and j_{emp}. If $\alpha \approx 1$, then the update strategy almost equals the empirical distribution of the ancestors' type. If α is small, then the empirical distribution j_{emp} is gradually assimilated to the strategy. Ancestral learning coincides with Xue's rule if $\tau_{\mathrm{est}} = 1$.

Ancestral learning is a biologically reasonable learning rule. The update rule only utilizes the empirical distribution j_{emp} of ancestors' types, which can be transmitted via epigenetic states or culture. Owing to this fact, we call j_{emp} *ancestral information.* Also, the memory to store j_{emp} is sufficiently small. An agent stores $j_{\mathrm{emp}} \in \mathcal{P}(\mathcal{X})$, whose size is substantially smaller than the whole path $\mathbb{X}^{(t)}$ of ancestors' types. We will see in Section 4.4 that this reduced information is sufficient for ancestral learning to accelerate the evolutionary process. The update rule of ancestral learning is natural since it is similar to Hebb's rule [44] as pointed out in [115]. Hebb's rule is a reinforcement of the synaptic connection between activated and coactivated neurons. Both Hebb's rule and ancestral learning are positive feedbacks. Indeed, with the updated strategy ${\pi_{\mathrm{F}}}^{(i)}$, an agent is more likely to express type x that the ancestor expresses more frequently.

The intuitive reason why an agent can acquire the optimal strategy by ancestral learning is that replicating the types with which the ancestors survived is likely to contribute to the survival of the descendants. Due to populational evolution, the empirical distribution $j_{\mathrm{emp}}^{{\pi_{\mathrm{F}}}^{(i-1)}}$ is biased from ${\pi_{\mathrm{F}}}^{(i-1)}$. The empirical distribution $j_{\mathrm{emp}}^{{\pi_{\mathrm{F}}}^{(i-1)}}$ seen as a strategy has greater populational fitness than ${\pi_{\mathrm{F}}}^{(i-1)}$. This survivorship bias is the driving force of ancestral learning as well as the conventional evolutionary process by natural selection.

To make the discussion clear, we first consider a simple case where the environment is constant $\mathcal{Y} = \{*\}$, the learning rate $\alpha = 1.0$, and the update interval τ_{est} is sufficiently long. In this case, the optimal strategy is

$$\begin{cases} \pi_{\mathrm{F}}^{*}(x^{*}) = 1 & (x^{*} = \operatorname*{argmax}_{x \in \mathcal{X}} k(x)), \\ \pi_{\mathrm{F}}^{*}(x) = 0 & (\text{otherwise}), \end{cases} \tag{4.5}$$



τ_{est} -generations
$\pi_{\mathrm{F}}^{(i)} \leftarrow j_{\mathrm{emp}}^{\pi_{\mathrm{F}}^{(i-1)}}$
$\pi_{\mathrm{F}}^{(i-1)}$
$\pi_{\mathrm{F}}^{(i-1)}$
$\pi_{\mathrm{F}}^{(i-1)}$
$\pi_{\mathrm{F}}^{(i-1)}$
$\pi_{\mathrm{F}}^{(i)}$
: 3/4
: 1/4
$j_{\mathrm{emp}}^{\pi_{\mathrm{F}}^{(i-1)}}$

Figure 4.2: Schematic representation of ancestral learning. The figure is adopted from [75]. The color of an agent represents its type. After an agent acquires the strategy ${\pi_{\mathrm{F}}}^{(i-1)}$ by the $(i-1)$-th update, its descendants have the same strategy for τ_{est}-generations. After τ_{est}-generations, the descendant calculate the empirical distribution $j_{\mathrm{emp}}^{\pi_{\mathrm{F}}^{(i-1)}}$ of ancestors' types defined by (4.3). Then, the strategy is updated by the rule (4.4). The figure corresponds to the case where $\tau_{\mathrm{est}} = 4$ and $\alpha = 1.0$.

which means that the optimal strategy π_{F}^* only selects the optimal type x^* that maximizes the individual fitness. To see ${\pi_{\mathrm{F}}}^{(i)}$ converges to π_{F}^* as $i \to \infty$, let us calculate $j_{\mathrm{emp}}^{\pi_{\mathrm{F}}^{(i-1)}}$. By the assumption $\alpha = 1.0$ and $\tau_{\mathrm{est}} \approx \infty$, we know that $j_{\mathrm{emp}}^{\pi_{\mathrm{F}}^{(i-1)}}$ is close to retrospective process ${\pi_{\mathrm{B}}}^{(i-1)}$ of ${\pi_{\mathrm{F}}}^{(i-1)}$. The retrospective process (3.23) is biased so that ${\pi_{\mathrm{B}}}^{(i-1)}(x^*)$, the probability selecting the optimal type, is larger then ${\pi_{\mathrm{F}}}^{(i-1)}(x^*)$ due to the factor $e^{k(x)}$ in the numerator. Therefore, ${\pi_{\mathrm{B}}}^{(i-1)}$ seen as a strategy has greater population fitness. By recursively applying the update ${\pi_{\mathrm{F}}}^{(i)} \leftarrow {\pi_{\mathrm{B}}}^{(i-1)}$ of the strategy, we know that ${\pi_{\mathrm{F}}}^{(i)}(x) \propto e^{ik(x)}{\pi_{\mathrm{F}}}^{(0)}(x)$. This equation implies that ${\pi_{\mathrm{F}}}^{(i)}$ converges to the optimal strategy (4.5) as $i \to \infty$.

We next consider the case where the environment is not constant. By a similar argument, we know that ${\pi_{\mathrm{F}}}^{(i)} = \bar{\pi}_{\mathrm{B}}^{(i-1)}$, where $\bar{\pi}_{\mathrm{B}}^{(i-1)}(x) = \mathbb{E}_{Q(y)}[\pi_{\mathrm{B}}(x \mid y)]$ is the averaged retrospective process (3.17) of ${\pi_{\mathrm{F}}}^{(i-1)}$. The retrospective process ${\pi_{\mathrm{B}}}^{(i-1)}(x \mid y)$ is better fitted to the environmental state y. Therefore, the updated strategy ${\pi_{\mathrm{F}}}^{(i)}$ is the mixture of $\pi_{\mathrm{B}}(x \mid y)$, each of which is better fitted to y than ${\pi_{\mathrm{F}}}^{(i-1)}$. In the following, we numerically (Section 4.3) and theoretically (Section 4.4) show that the recursive updates to such a mixed strategy lead to the optimal strategy.

4.3 Ancestral Learning can Accelerate Evolutionary Process

We next numerically validate that ancestral learning can accelerate the evolutionary process. Specifically, we numerically show that the optimal strategy π_{F} is acquired by ancestral learning faster than the zeroth-order rules.

We simulated the evolutionary process as a multitype branching process in random environments [40, 114]. In other words, we simulated the dynamical system (4.1) while taking into account of the individuality and the finite size of the population. In this simulation, we set $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ (Figure 4.3 (a)). We set $Q(0) = 0.6$ and $Q(y) = 0.2$ otherwise. Each agent with type x at time t reproduces four daughters if $x = y^{(t)}$ and one daughter otherwise. In short, $e^{k(x,y)} = 4$ if $x = y$ and $e^{k(x,y)} = 1$ otherwise. We represent π_{F} in the form $(\pi_{\mathrm{F}}(0), \pi_{\mathrm{F}}(1), \pi_{\mathrm{F}}(2))$. The optimal strategy in this setting is approximately $\pi_{\mathrm{F}}^* = (0.92, 0.04, 0.04)$ since it satisfies the optimality condition (Karush-Kuhn-Tucker condition) with a small error [22, Theroem 16.2.1]. The optimal strategy has greater $\pi_{\mathrm{F}}(0)$ than $\pi_{\mathrm{F}}(1)$ and $\pi_{\mathrm{F}}(2)$. We started the simulation from a single agent with an initial strategy ${\pi_{\mathrm{F}}}^{(0)} = (1/3, 1/3, 1/3)$. We limited the number of agents in the population to $N_{\mathrm{size}} = 30$ to avoid the situation where exponential growth of the population makes the simulation intractable. If the number of agents exceeded N_{size}, then we selected N_{size}-agents uniformly at random.

We investigated three learning rules. Each learning rule updates the strategy at every generation, i.e., $\tau_{\mathrm{est}} = 1$. The first rule was ancestral learning with $\alpha = 0.01$. The second and the third rules were the zeroth-order rules. Since there were innumerable zeroth-order rules, we selected two rules as representative to perform the control experiments for ancestral learning. The second learning rule was $\pi_{\mathrm{F}} \leftarrow (1 - \alpha)\pi'_{\mathrm{F}} + \alpha\delta_{x,x_{\mathrm{rand}}}$, where π'_{F} and π_{F} were the strategy before and after the update and x_{rand} was selected uniformly at random from $\mathcal{X}$. In biological systems, this update rule can be seen as a random mutation of π_{F} with a constant mutation rate. The trajectory of π_{F} updated by this rule is a random walk over strategies if no growth occurs, that is, $e^{k(x,y)} = 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Therefore, we call this rule a random walk. The third rule was $\pi_{\mathrm{F}} \leftarrow (1 - \alpha)\pi'_{\mathrm{F}} + \alpha\delta_{x,x_{\mathrm{samp}}}$, where x_{samp} was sampled from π'_{F}. In biological systems, this rule can be seen as a mutation of π_{F} whose rate is dependent on current π_{F}. The change of the mutation rate is known as adaptive mutation [92]. Therefore, we call this rule an adaptive random walk. The adaptive random walk coincides with the ancestral learning if no growth occurs. In this sense, the adaptive random walk is a control to see the effect of populational evolution on ancestral learning.

Figure 4.3 showed the simulation of the three learning rules. We ran the simulation until $t = 50$. The population fitness of the population with ancestral learning increased faster than those with the other learning rules along the lineage of the agent that had the greatest population fitness among the population at the end, which we call the most successful agent in the following (Figure 4.3 (b)). We observed the faster increase in the population fitness of the population with ancestral learning at the lineage tree level (Figure 4.3 (d–f)). In (g–i), we showed the trajectories of the strategies along the lineage of the most successful agent with each learning rule. We observed that the strategy moved faster towards the optimal strategy $\pi_{\mathrm{F}}^* \approx (0.92, 0.04, 0.04)$ when agents adopt ancestral learning.

To see whether the optimal strategy was acquired by ancestral learning, we ran another simulation until $t = 1500$. We first checked that the strategy converged and then verified that the converged strategy was optimal. The strategy converged since the population fitness along the lineage of the most successful agent with ancestral learning reached a ceiling (Figure 4.3). We observed the convergence directly from the trajectories of the strategies along the lineage (Figure 4.3 (j–l)). The converged strategy was closed to the optimal one $\pi_{\mathrm{F}}^* \approx (0.92, 0.04, 0.04)$.

From these results, we conclude that ancestral learning accelerates the evolutionary process. The population with ancestral learning acquired the optimal strategy faster than those with the other learning rules. Since ancestral learning does not utilize communication among agents, these results indicate that learning can accelerate the evolutionary process even without communication.

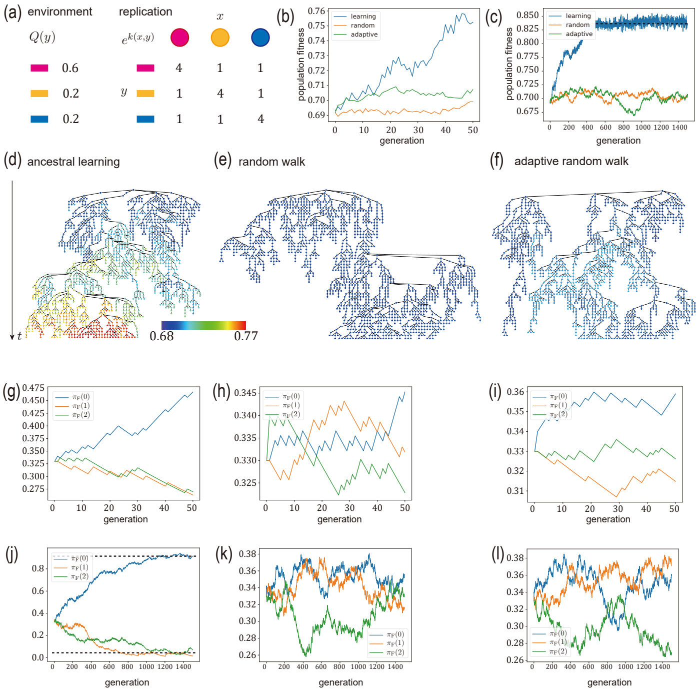


Figure 4.3: Numerical experiments of ancestral learning. This figure is adopted from [75]. (a) The parameters of the model we simulate. In this panel, $0, 1, 2 \in \mathcal{Y}$ are represented by red, yellow, and blue, respectively. The red environmental state occurs more frequently than the others. An agent reproduces more daughters when its type equals the environmental state. (b) The trajectories of the population fitness until $t = 50$ along the lineage of the agent whose population fitness was the greatest among the population at the end of the simulation, which we call the most successful agent in the following. The blue, green, and orange curves represent ancestral learning, the random walk, and the adaptive random walk, respectively. Ancestral learning increased the population fitness the best among the three learning rules. (c) The same plot as (b) until $t = 1500$. The dotted line showed the population fitness of the most successful agent with ancestral learning at $t = 1500$. We observed the convergence of the population fitness. (d–f) The simulated lineage trees of the population of agents that adopt ancestral learning (d), the random walk (e), and the adaptive random walk (f), respectively. Each dot corresponds to an agent, and a parent and its daughters are connected by lines. The color of each dot represents the population fitness of the corresponding agent. We observed that ancestral learning increases the population fitness the best at the lineage tree level. (g–i) The trajectories of the strategy until $t = 50$ along the linage of the most successful agent that adopts ancestral learning (g), random walk (h), adaptive random walk (i), respectively. (j–l) The same plot as (g–i) until $t = 1500$. In (j), the above dotted line showed $\pi_{\mathrm{F}}(0)$ and the below showed the average of $\pi_{\mathrm{F}}(1)$ and $\pi_{\mathrm{F}}(2)$ at the end of the lineage. In (j), we can see that the strategy converged. The converged strategy was approximately $(0.92, 0.04, 0.04)$, which satisfied the optimality condition with a small error [22, Theroem 16.2.1]. In (k) and (l), the strategies do not converge.

4.4 Ancestral Information is Sufficient to Estimate Fitness Gradient

We next address the second problem: We investigate whether an agent can estimate the gradient of the population fitness from the accessible information. Although we numerically showed that ancestral learning accelerates the evolutionary process, the relationship between the ancestral information and the gradient is not clear. The ancestral information might be insufficient to estimate the gradient and communication among agents might be required. In this section, we theoretically show that an agent can estimate the gradient of the population fitness from the ancestral information.

Since π_{F} has a constraint $\pi_{\mathrm{F}} \in \mathcal{P}(\mathcal{X})$, we use a the following quantity instead of the gradient $\partial\lambda/\partial\pi_{\mathrm{F}}(x)$:

$$\lim_{\epsilon\to^{+}0} \underset{\substack{\delta\pi \\ \pi_{\mathrm{F}}+\delta\pi\in D_{\epsilon}(\pi_{\mathrm{F}})}}{\operatorname{argmax}} \left\{\lambda(\pi_{\mathrm{F}}+\delta\pi)\right\}, \tag{4.6}$$

where $\delta\pi \in \mathbb{R}^{\mathcal{X}}$ and $D_{\epsilon}(\pi_{\mathrm{F}})$ is the sphere around π_{F} with radius ϵ. To define the sphere, we use the KL-divergence as a natural distance on $\mathcal{P}(\mathcal{X})$ as

$$D_{\epsilon}(\pi_{\mathrm{F}}) := \{\pi \in \mathcal{P}(\mathcal{X}) \mid \mathcal{D}\left[\pi_{\mathrm{F}}\|\pi\right] < \epsilon\}. \tag{4.7}$$

We note that (4.6) is similar to the representation of the gradient by a proximal operator [82]. Indeed, if we consider no constraints and we use the l_2-norm instead of the KL-divergence, then (4.6) coincides with the gradient. We therefore call (4.6) a *proximal gradient* of the population fitness. Intuitively, the proximal gradient is the direction where the population fitness increases the most among the alternatives that satisfy the constraint and that is the same distance from π_{F}.

$P(\{x^{(t)}\}_t \mid \{y^{(t)}\}_t) = \frac{e^{k(\{x^{(t)}\}_t,\{y^{(t)}\}_t)}P(\{x^{(t)}\}_t)}{\sum_{\{x^{(t)}\}_t} e^{k(\{x^{(t)}\}_t,\{y^{(t)}\}_t)}P(\{x^{(t)}\}_t)}$

We can calculate the proximal gradient from Proposition 3.2.

Theorem 4.1. If π_{F} is interior [3] of $\mathcal{P}(\mathcal{X})$, then

$$\lim_{\epsilon\to^{+}0} \underset{\substack{\delta\pi \\ \pi_{\mathrm{F}}+\delta\pi\in D_{\epsilon}(\pi_{\mathrm{F}})}}{\operatorname{argmax}} \left\{\lambda(\pi_{\mathrm{F}}+\delta\pi)\right\} \propto \bar{\pi}_{\mathrm{B}} - \pi_{\mathrm{F}}. \tag{4.8}$$

The theorem addresses the second problem. To estimate the proximal gradient of the population fitness, an agent must estimate $\bar{\pi}_{\mathrm{B}}$. By the discussion in Section 4.2, we know that the ancestral information j_{emp} is an unbiased estimator of $\bar{\pi}_{\mathrm{B}}$. Therefore, an agent can estimate the proximal gradient without communication among agents at the same generation. The theorem also implies that ancestral learning updates the strategy in the direction of the proximal gradient. The direction ${\pi_{\mathrm{F}}}^{(i)} - {\pi_{\mathrm{F}}}^{(i-1)}$ of the update is proportional to the proximal gradient. In particular, ancestral learning can attain the optimal strategy if the learning rate α is sufficiently small since the population fitness $\lambda(\pi_{\mathrm{F}})$ is concave.

4.5 Fisher's Fundamental Theorem of Ancestral Learning

We next solve the third problem, the quantification of the acceleration of the evolutionary process by learning. The acceleration may depend on the property

[3] We exclude the case where π_{F} is on the boundary of $\mathcal{P}(\mathcal{X})$ since the boundary is measure zero. By considering the KKT-condition, we can generalize the theorem for such a case.

$Q(y)$ of the environment and the hyperparameters (α and τ_{est}) of ancestral learning. We can quantitatively understand such dependency by extending FF-thm of natural selection to ancestral learning.

To see the connection of the original FF-thm and ancestral learning, we first extend FF-thm to the population fitness in the fixed-type constant environment model (3.40). Although the original FF-thm quantifies the increase in the average individual fitness, we are not interested in the average individual fitness but the population fitness. The population fitness at time t in the fixed-type model is defined by

$$\lambda^{(t)} := \log \frac{\sum_{x \in \mathcal{X}} N^{(t)}(x)}{\sum_{x \in \mathcal{X}} N^{(t-1)}(x)}. \tag{4.9}$$

We measure the speed of the evolutionary process by the increase of the population fitness. We can characterize the increase by using the log-variance (3.4).

Lemma 4.2 (Fisher's Fundamental Theorem of Population Fitness)**.**

$$\Delta\lambda^{(t)} := \lambda^{(t)} - \lambda^{(t-1)} = \text{log-}\mathbb{V}_{p^{(t-1)}}\left[k(x)\right] \tag{4.10}$$

FF-thm of the population fitness relates the increase of the population fitness to the log-variance of the individual fitness in the population.

FF-thm of the population fitness has a close connection with ancestral learning. To see this, let us consider a simple situation where the environment is constant $\mathcal{Y} = \{*\}$, the learning rate $\alpha = 1.0$, and the update interval $\tau_{\mathrm{est}} \approx \infty$. In this situation, we define the acceleration of the evolutionary process by ancestral learning learning as $\Delta\lambda^{(i)} := \lambda({\pi_{\mathrm{F}}}^{(i)}) - \lambda({\pi_{\mathrm{F}}}^{(i-1)})$, where ${\pi_{\mathrm{F}}}^{(i-1)}$ and ${\pi_{\mathrm{F}}}^{(i)}$ are the strategies before and after the update by ancestral learning, respectively. Since the gain $\Delta\lambda^{(i)}$ is independent of populational evolution, we can regard it as the acceleration of evolutionary process by ancestral learning. The gain $\Delta\lambda^{(i)}$ is equivalent to the left hand side of (4.10) if we identify ${\pi_{\mathrm{F}}}^{(i)}$ with $p^{(t)}$. In addition, we showed in Section 4.2 that the update of ancestral learning is ${\pi_{\mathrm{F}}}^{(i)} \leftarrow {\pi_{\mathrm{B}}}^{(i-1)}$ under this setting, where ${\pi_{\mathrm{B}}}^{(i-1)}$ is the retrospective process with respect to ${\pi_{\mathrm{F}}}^{(i-1)}$. The update rule is equivalent to the time evolution of the fraction $p^{(t)}(x)$ of organisms with type x in the fixed-type model (3.41) if we identify ${\pi_{\mathrm{F}}}^{(i)}$ with $p^{(t)}$ again. By these equivalences, we have FF-thm of ancestral learning.

Theorem 4.3 (Fisher's Fundamental Theorem of Ancestral Learning in Constant Environments)**.**

$$\Delta\lambda^{(i)} := \lambda({\pi_{\mathrm{F}}}^{(i)}) - \lambda({\pi_{\mathrm{F}}}^{(i-1)}) = \text{log-}\mathbb{V}_{{\pi_{\mathrm{F}}}^{(i-1)}}[k(x)]. \tag{4.11}$$

The theorem relates the gain of the population fitness by ancestral learning to the log-variance of the individual fitness of the strategy. The theorem also reveals the trade-off between the acceleration $\Delta\lambda^{(i)}$ and the population fitness $\lambda({\pi_{\mathrm{F}}}^{(i)})$. When the log-variance $\text{log-}\mathbb{V}_{{\pi_{\mathrm{F}}}^{(i-1)}}[k(x)]$ is large, an agent can acquire information about which type is fitted to the environment by expressing a variety of types. Therefore, the gain of the population fitness by ancestral learning is large as the extended FF-thm indicates. We call such a situation exploratory. In contrast, when ${\pi_{\mathrm{F}}}^{(i)}$ is close to the optimal and $\lambda({\pi_{\mathrm{F}}}^{(i)})$ is large, the log-variance of the strategy is small since the optimal strategy only expresses the optimal type x^* (4.5). Therefore, the gain $\lambda({\pi_{\mathrm{F}}}^{(i)})$ is small by the extended FF-thm. We call such a situation exploitation. In all, we can see the so-called exploratory-exploitation trade-off in this setting.

We can further extend FF-thm to the case where the environment is not constant.

Theorem 4.4 (Fisher's Fundamental Theorem of Ancestral Learning)**.**

$$\begin{aligned}\Delta\lambda^{(i)} = \mathbb{E}_{Q(y)Q(y')}\left[\text{log-Cov}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[k(x,y),k(x,y')\right]\right] \\ + \mathcal{D}\left[Q(y)Q(y')\middle\|\bar{Q}^{(i)}(y' \mid y)Q(y)\right],\end{aligned} \tag{4.12}$$

where

$$\bar{Q}^{(i)}(y' \mid y) :\propto \sum_{x\in\mathcal{X}} e^{k(x,y)}\pi_{\mathrm{B}}{}^{(i-1)}(x \mid y')Q(y'). \tag{4.13}$$

We note that Theorem 4.4 is reduced to Theorem 4.3 when the environment is constant. We also note that Theorem 4.4 is different from the FF-thm of natural selection in random environments. Indeed, the time evolution of $p^{(t)}$ in random environments differs from the update of by ancestral learning. The time evolution of $p^{(t)}$ is stochastic and satisfies

$$p^{(t)}(x) = \frac{e^{k(x,y)}p^{(t-1)}(x)}{\sum_{x'\in\mathcal{X}} e^{k(x',y)}p^{(t-1)}(x')}, \tag{4.14}$$

with probability $Q(y)$.

4.6 Measures to Characterize Ancestral Learning

Since FF-thm in general environments (4.12) has the second term that does not appear in (4.11), we give an interpretation why two terms appear in (4.12). For this purpose, we define *actual gain* $\Delta_{\mathrm{ac}}\lambda^{(i)}$ and *expected gain* $\Delta_{\mathrm{ex}}\lambda^{(i)}$ as the left hand and the right hand side of (4.12), respectively:

$$\Delta_{\mathrm{ac}}\lambda^{(i)} := \lambda(\pi_{\mathrm{F}}{}^{(i)}) - \lambda(\pi_{\mathrm{F}}{}^{(i-1)}), \tag{4.15}$$

and

$$\Delta_{\mathrm{ex}}\lambda^{(i)} := \tilde{\Sigma}^{(i)} + \mathrm{KL}^{(i)}, \tag{4.16}$$

where $\tilde{\Sigma}^{(i)}$ and $\mathrm{KL}^{(i)}$ are the variance and the KL terms of $\Delta_{\mathrm{ex}}\lambda^{(i)}$ defined respectively as

$$\tilde{\Sigma}^{(i)} := \mathbb{E}_{Q(y)Q(y')}\left[\text{log-Cov}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[k(x,y),k(x,y')\right]\right], \tag{4.17}$$

and

$$\mathrm{KL}^{(i)} := \mathcal{D}\left[Q(y)Q(y')\middle\|\bar{Q}^{(i)}(y' \mid y)Q(y)\right]. \tag{4.18}$$

The reason the additional KL term appears in (4.12) is attributed to the existence of two representative strategies: *specialist* and *generalist.* Each term ((4.17) and (4.18)) corresponds to one of the representative strategies and measures the gain of the population fitness to acquire the corresponding strategy by ancestral learning. Specialist is defined as a situation where an agent expresses a few types that are fitted to the environment. Formally, a strategy π_{F} is specialized to $\mathcal{X}' \subseteq \mathcal{X}$ if $\pi_{\mathrm{F}}(x) > 0$ for all $x \in \mathcal{X}'$ and $\pi_{\mathrm{F}}(x) = 0$ otherwise. An example of specialist is the optimal strategy (4.5) when the environment is constant. Specialization is beneficial when the environment is constant or the environmental states are similar. In such situations, an agent can survive by expressing a few types. However, if the environmental states are dissimilar, an agent cannot survive by expressing a few types because concentrated types might not be fitted to some environmental states. Therefore, an agent should stochastically choose types from a variety of alternatives to reduce the risk of specialization. The probability to express a type should be set so that the population fitness is maximized. Even if the strategy is concentrating on $\mathcal{X}' \subsetneq \mathcal{X}$ with $|\mathcal{X}'| > 1$, the probability $\pi_{\mathrm{F}}(x)$ for $x \in \mathcal{X}'$ should be set to maximize the population fitness. We call such a situation a generalist. Formally, a strategy is generalized in $\mathcal{X}'$ if $\pi_{\mathrm{F}}(x) > 0$ for all $x \in \mathcal{X}'$ and the probabilities $\pi_{\mathrm{F}}(x)$ are set so that the population fitness is maximized. An example of generalization is the optimal strategy in the model we used in Section 4.3 (Figure 4.3 (a)). In general, the optimal strategy is the combination of specialist and generalist. For example, let us examine the optimal strategy $\pi_{\mathrm{F}}^* = (0.72, 0.0, 0.28)$ of the model shown in Figure 4.4 (j). The optimal strategy is specialized to and generalized in $\{0, 2\} \subsetneq \mathcal{X}$.

During the evolutionary process with learning, an agent attains the optimal strategy by acquiring two representative strategies. The gain of the population fitness by acquiring each representative strategy is measured by the variance term $\tilde{\Sigma}^{(i)}$ and the KL term $\mathrm{KL}^{(i)}$ of the expected gain (4.16). The variance term measures the fitness gain by acquiring a specialist strategy and the KL term does the gain by acquiring a generalist strategy. To see this interpretation, we rewrite $\bar{\pi}_{\mathrm{B}}$ since the update rule is ${\pi_{\mathrm{F}}}^{(i)} \leftarrow \bar{\pi}_{\mathrm{B}}$ under the setting of Theorem 4.4. By definition,

$$\bar{\pi}_{\mathrm{B}}^{(i-1)}(x) = \mathbb{E}_{Q(y)}\left[\frac{e^{k(x,y)}}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}(x')}\left[e^{k(x',y)}\right]}\right]{\pi_{\mathrm{F}}}^{(i-1)}(x) \tag{4.19}$$

$$\propto \mathbb{E}_{Q(y)}\left[e^{k(x,y)}\right]{\pi_{\mathrm{F}}}^{(i-1)}(x). \tag{4.20}$$

The equation is a transformation of the probability measure from ${\pi_{\mathrm{F}}}^{(i-1)}$ to $\bar{\pi}_{\mathrm{B}}^{(i-1)}$ by multiplicative factors $\{\mathbb{E}[e^{k(x,y)}]\}_{x\in\mathcal{X}}$. We examine these multiplicative factors. We regard the multiplicative factors $\{\mathbb{E}[e^{k(x,y)}]\}_{x\in\mathcal{X}}$ as a vector in $\mathbb{R}^{\mathcal{X}}$. Then, the multiplicative factors are the average of $\boldsymbol{F}_y = (F_y(x))_{x\in\mathcal{X}} := (e^{k(x,y)})_{x\in\mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$ defined for each $y \in \mathcal{Y}$. We regard $\boldsymbol{F}_y$ as an embedding of the environmental state y into $\mathbb{R}^{\mathcal{X}}$. By using this embedding, we can measure the similarity of the environmental states y and y' by log-$\mathrm{Cov}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[\boldsymbol{k}_y, \boldsymbol{k}_{y'}\right]$, where

$$\boldsymbol{k}_y = (\log F_y(x))_{x\in\mathcal{X}} = (k(x,y))_{x\in\mathcal{X}} \in \mathbb{R}^{\mathcal{X}}. \tag{4.21}$$

By (3.7), we know that

$$\text{log-Cov}\left[\boldsymbol{k}_y, \boldsymbol{k}_{y'}\right] = \log\left(1 + \frac{\mathrm{Cov}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[F_y(x), F_{y'}(x)\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}[F_y(x)]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}[F_{y'}(x)]}\right), \tag{4.22}$$

and this quantity indeed measures the similarity between the embedding $\boldsymbol{F}_y$ and $\boldsymbol{F}_{y'}$ since it is monotonically increasing with respect to the covariance of the embedding. By considering the normalization factor in (4.19), we in addition define the scaled version $\boldsymbol{f}_y$ of $\boldsymbol{F}_y$ by

$$f_y(x) := \frac{F_y(x)}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}(x')}\left[e^{k(x',y)}\right]}, \tag{4.23}$$

which depends on the current strategy ${\pi_{\mathrm{F}}}^{(i-1)}$. By using $f_y(x)$, we can rewrite (4.19) as

$$\bar{\pi}_{\mathrm{B}}^{(i-1)}(x) = \mathbb{E}_{Q(y)}\left[f_y(x)\right]{\pi_{\mathrm{F}}}^{(i-1)}(x). \tag{4.24}$$

The updated strategy is more specialist if the environmental states are similar since, if $\boldsymbol{f}_y$ $(y \in \mathcal{Y})$ have similar peaks, then so does their average $\bar{\pi}_{\mathrm{B}}^{(i-1)}$ (Figure 4.4 (e)). Iteration of such updates leads to concentration on the types where the peaks lie. We will see in the next paragraph that the variance term (4.17) measures the similarity between environmental states and thus quantifies the fitness gain by acquiring a specialist strategy. On the other hand, the updated strategy is more generalist if the environmental states are dissimilar since, if $\boldsymbol{f}_y$ $(y \in \mathcal{Y})$ have different peaks, then their average $\bar{\pi}_{\mathrm{B}}^{(i-1)}$ becomes flat (Figure 4.4 (h)). Iteration of such update leads to a generalist since concentration does not occur and the probability $\pi_{\mathrm{F}}(x)$ is adjusted to maximize the population fitness during the process of ancestral learning. We will see that the KL term (4.18) measures the dissimilarity of environmental states and thus quantifies the fitness gain by becoming a generalist.

We rewrite (4.12) to see that the variance term $\tilde{\Sigma}^{(i)}$ and the KL term $\mathrm{KL}^{(i)}$ measures the similarity and the dissimilarity of the environmental state, respectively. We first treat the variance term. By definition, the variance term equals

$$\tilde{\Sigma}^{(i)} = \mathbb{E}_{Q(y)Q(y')}\left[\text{log-Cov}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[k_y(x), k_{y'}(x)\right]\right]. \tag{4.25}$$

Since the log-covaricene in the right hand side measures the similarity of environmental states, so does the variance term. We next treat the KL term. We can rewrite the KL term as follows.

Lemma 4.5.

$$\mathrm{KL}^{(i)} = \mathbb{E}_{Q(y)Q(y')}\left[-\log \frac{\mathbb{E}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[F_y(x)F_{y'}(x)\right]}{\mathbb{E}_{\pi_{\mathrm{F}}{}^{(i)}}\left[F_y(x)\right]\mathbb{E}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[F_{y'}(x)\right]}\right]. \tag{4.26}$$

By this lemma, we can see that the KL term is in principle greater when environmental states are dissimilar since the second moment of $\boldsymbol{F}_y$ appears in the numerator. However, the dependency on $\pi_{\mathrm{F}}{}^{(i)}$ in the denominator may change the relationship. Therefore, the KL term measures the dissimilarity of environmental states in principle.

4.7 Numerical Validation of FF-thm of Ancestral Learning

We numerically verified FF-thm. To check the interpretation in Section 4.6, we simulated four models whose environments $Q(y)$ are different. In each model, we checked that FF-thm held, i.e., $\Delta_{\mathrm{ac}}\lambda^{(i)} = \Delta_{\mathrm{ex}}\lambda^{(i)}$. In the following, we set $\alpha = 1.0$ unless otherwise specified. Also, we set $\tau_{\mathrm{est}} = 1000$ to suppress the stochastic fluctuation of j_{emp}.

We first validated FF-thm when the environment was constant. We simulated the model shown in Figure 4.4 (a) and call it a constant environment model. We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ along the lineage of an agent whose initial strategy was $\pi_{\mathrm{F}}{}^{(0)} = (0.5, 0.5)$ (Figure 4.4 (b)). To validate FF-thm beyond one initial strategy, we compared $\Delta_{\mathrm{ac}}\lambda^{(1)}$ and $\Delta_{\mathrm{ex}}\lambda^{(1)}$ of an agent which had a randomly generated initial strategy after the first update. We observed that $\Delta_{\mathrm{ac}}\lambda^{(1)} \approx \Delta_{\mathrm{ex}}\lambda^{(1)}$ for most of the initial strategies (Figure 4.4 (c)).

We next validated FF-thm when the environment was not constant by using three models. We first simulated the model shown in Figure 4.4 (d) whose environmental states were similar. We call the model the similar environment model. In this model, the optimal strategy is specialized $\{0\} \subseteq \mathcal{X}$ (Figure 4.4 (e)) and we expect that the variance term dominates. We compared the actual gain $\Delta_{\mathrm{ac}}\lambda^{(i)}$, the expected gain $\Delta_{\mathrm{ex}}\lambda^{(i)}$, the variance term $\tilde{\Sigma}^{(i)}$, and the KL term $\mathrm{KL}^{(i)}$ along the lineage of an agent whose initial strategy was ${\pi_{\mathrm{F}}}^{(0)} = (0.5, 0.5)$ in Figure 4.4 (f). We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ and that $\tilde{\Sigma}^{(i)}$ dominated as expected.

We next simulated the model shown in Figure 4.4 (g) whose environmental states are dissimilar. We call the model a different environment model. In this model, the optimal strategy is generalist and we expect that the KL term is not negligible (Figure 4.4 (h)). We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ and that $\mathrm{KL}^{(i)}$ is not negligible as expected along the lineage of an agent whose initial strategy is ${\pi_{\mathrm{F}}}^{(0)} = (0.9, 0.1)$ (Figure 4.4 (i)).

We finally simulated the model shown in Figure 4.4 (j). Since the red and yellow environmental states are similar while they are dissimilar from the blue state, we call the model a combined model. In this model, the optimal strategy is approximately $\pi^* \approx (0.72, 0, 0.28)$ and the combination of specialized to and generalized in $\mathcal{X}' = \{0, 2\}$. We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ along the lineage of an agent whose initial strategy was ${\pi_{\mathrm{F}}}^{(0)} = (1/3, 1/3, 1/3)$. We in addition observed that $\tilde{\Sigma}^{(i)}$ dropped faster than the KL term. This result indicated that an agent acquired a specialist strategy first and then the strategy became a generalist in the specialized types. This interpretation was supported by the strategy ${\pi_{\mathrm{F}}}^{(5)} = (0.31, 0.04, 0.65)$ just before the variance term became negative for the first time. The strategy was almost specialized $\mathcal{X}'$ while it was not generalized in $\mathcal{X}'$ since ${\pi_{\mathrm{F}}}^{(5)}(x)$ for $x \in \mathcal{X}'$ was still far from the optimizer $\pi^*(x)$. To validate FF-thm beyond one initial strategy, we compared $\Delta_{\mathrm{ac}}\lambda^{(1)}$ and $\Delta_{\mathrm{ex}}\lambda^{(1)}$ of an agent which had randomly generated initial strategy after the first update. We observed that $\Delta_{\mathrm{ac}}\lambda^{(1)} \approx \Delta_{\mathrm{ex}}\lambda^{(1)}$ for most of the initial strategies (Figure 4.4 (l)).

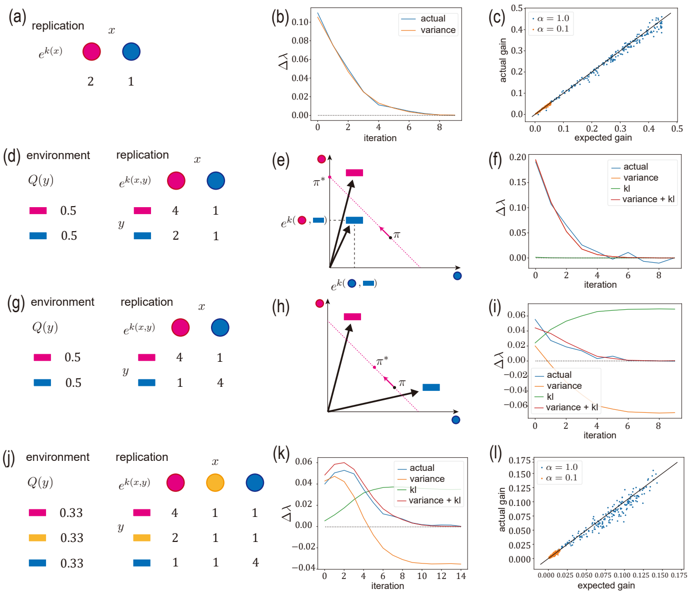


Figure 4.4: The numerical verification of FF-thm of ancestral learning. We adopted this figure from [75]. We set $\alpha = 1.0$ unless otherwise specified. (a–c) The constant environment model. (a) The parameters of the model. (b) The trajectories of the actual gain $\Delta_{\mathrm{ac}}\lambda^{(i)}$ (4.15) and the expected gain $\Delta_{\mathrm{ex}}\lambda^{(i)}$ (4.16) along the lineage of an agent. Since $\Delta_{\mathrm{ex}}\lambda^{(i)} = \tilde{\Sigma}^{(i)}$, we plot $\tilde{\Sigma}^{(i)}$. We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ and FF-thm held. (c) The comparison between $\Delta_{\mathrm{ac}}\lambda^{(1)}$ and $\Delta_{\mathrm{ex}}\lambda^{(1)}$ when an agent has a randomly generated strategy. We observed that $\Delta_{\mathrm{ac}}\lambda^{(1)} \approx \Delta_{\mathrm{ex}}\lambda^{(1)}$ for most of the initial strategies when $\alpha = 1.0$ and $\alpha = 0.1$. Therefore, FF-thm held beyond one lineage. (d–f) The similar environment model. (d) The parameters of the model. (e) An illustration of the embedding $\boldsymbol{F}_y$ of the environmental state y into $\mathbb{R}^{\mathcal{X}}$. The environmental states are similar in this model since the angle between two embedded states are small. Therefore, the optimal strategy is a specialist. The red line represents the constraint $\pi_{\mathrm{F}} \in \mathcal{P}(\mathcal{X})$. Ancestral learning updates the strategy towards the axis corresponding to the optimal type, which is the red type in the figure. (f) The trajectories of the actual gain $\Delta_{\mathrm{ac}}\lambda^{(i)}$, the expected gain $\Delta_{\mathrm{ex}}\lambda^{(i)}$, the variance term $\tilde{\Sigma}^{(i)}$, and the KL term $\mathrm{KL}^{(i)}$ along the lineage of an agent. We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$. In addition, $\tilde{\Sigma}^{(i)}$ dominates. (g–i) The different environment model. (g) The parameters of the model. (h) The same illustration as (e). We can see that the environmental states are dissimilar in this model. Therefore, the optimal strategy is a generalist and ancestral learning updates the strategy towards the intermediate point of the red line. (i) The same plot as (f). We can see that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$ and $\mathrm{KL}^{(i)}$ is not negligible. (j–l) The combined model. (j) The parameters of the model. (k) The same plot as (f). We observed that $\Delta_{\mathrm{ac}}\lambda^{(i)} \approx \Delta_{\mathrm{ex}}\lambda^{(i)}$. In this case, the variance term dropped faster than the KL term. This result indicated that an agent first acquired a specialist strategy and then acquired the generalist strategy. (i) The same plot as (c). We observed that FF-thm held for most of the initial strategies.

4.8 Trade-off Between Learning Rate and Update Interval

We extended FF-thm in the case where $\alpha = 1.0$ and $\tau_{\mathrm{est}} \approx \infty$. To address the third problem in other cases, we further extend FF-thm of ancestral learning to the case where $\alpha < 1.0$ and τ_{est} is finite. By this FF-thm, we can see a trade-off between α and τ_{est}.

As a preparation for further extension, we first introduce an α-log-covariance by generalizing (3.7):

$$\text{log-Cov}_p^{\alpha}\left[f(x), g(x)\right] := \log\left(1 + \alpha\frac{\mathrm{Cov}_p\left[e^{f(x)}, e^{g(x)}\right]}{\mathbb{E}_p\left[e^{f(x)}\right]\mathbb{E}_p\left[e^{g(x)}\right]}\right). \tag{4.27}$$

By using this quantity, we extend FF-thm of ancestral learning.

Theorem 4.6.

$$\begin{aligned}\Delta\lambda^{(i)} = \mathbb{E}_{Q(y)Q(y')}&\left[\text{log-Cov}^{\alpha}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[k(x,y), k(x,y')\right]\right]\\ &+ \mathcal{D}\left[Q(y)Q(y')\middle\|\bar{Q}^{(i)}_{\alpha}(y' \mid y)Q(y)\right],\end{aligned} \tag{4.28}$$

where

$$\bar{Q}^{(i)}_{\alpha}(y' \mid y) :\propto \sum_{x\in\mathcal{X}} e^{k(x,y)}\pi_{\alpha}{}^{(i-1)}(x \mid y')Q(y'), \tag{4.29}$$

and

$$\pi_{\alpha}{}^{(i-1)}(x \mid y') = \alpha\pi_{\mathrm{B}}{}^{(i-1)}(x \mid y') + (1-\alpha)\pi_{\mathrm{F}}{}^{(i-1)}(x). \tag{4.30}$$

We redefine the actual and the expected gain as the left and right hand side of (4.28):

$$\Delta_{\mathrm{ac}}\lambda^{(i)} := \lambda(\pi_{\mathrm{F}}{}^{(i)}) - \lambda(\pi_{\mathrm{F}}{}^{(i-1)}), \tag{4.31}$$

and

$$\begin{aligned}\Delta_{\mathrm{ex}}\lambda^{(i)} :=&\mathbb{E}_{Q(y)Q(y')}\left[\text{log-Cov}^{\alpha}_{\pi_{\mathrm{F}}{}^{(i-1)}}\left[k(x,y), k(x,y')\right]\right]\\ &+ \mathcal{D}\left[Q(y)Q(y')\middle\|\bar{Q}^{(i)}_{\alpha}(y' \mid y)Q(y)\right].\end{aligned} \tag{4.32}$$

To validate this FF-thm (Theorem 4.6), we simulated the constant and the combined model when the learning rate was $\alpha = 0.1$. We compared $\Delta_{\mathrm{ac}}\lambda^{(1)}$ and $\Delta_{\mathrm{ex}}\lambda^{(1)}$ of an agent that had a randomly generated initial strategy (Figure 4.4 (c,l)). We observed that $\Delta_{\mathrm{ac}}\lambda^{(1)} \approx \Delta_{\mathrm{ex}}\lambda^{(1)}$ for most of the initial strategies.

When τ_{est} is finite, FF-thm (Theorem 4.6) does not hold since the ancestral information $\jmath_{\mathrm{emp}}$ does not converge to $\bar{\pi}_{\mathrm{B}}$. Since $\jmath_{\mathrm{emp}}$ stochastically fluctuates around its expectation $\bar{\pi}_{\mathrm{B}}$, so does the updated strategy $\pi_{\mathrm{F}}{}^{(i)} = \alpha\jmath_{\mathrm{est}}^{\pi_{\mathrm{F}}{}^{(i-1)}} + (1-\alpha)\pi_{\mathrm{F}}{}^{(i-1)}$ around $\bar{\pi}_{\alpha}^{(i-1)} = \alpha\bar{\pi}_{\mathrm{B}}^{(i-1)} + (1-\alpha)\pi_{\mathrm{F}}{}^{(i-1)}$. Since the population fitness is concave, Jensen's inequality implies that $\Delta_{\mathrm{ac}}\lambda^{(i)} < \Delta_{\mathrm{ex}}\lambda^{(i)}$ in this case. We evaluate the deviation. When τ_{est} is sufficiently large (but finite), we can approximate $\jmath_{\mathrm{est}}$ by the central limit theorem [107] as

$$\jmath_{\mathrm{est}} \sim \mathcal{N}\left(\bar{\pi}_{\mathrm{B}}^{(i-1)}, \boldsymbol{V}\right), \tag{4.33}$$

where

$$V(x,x') = \mathbb{E}\left[\jmath_{\mathrm{est}}(x)\jmath_{\mathrm{est}}(x')\right] - \bar{\pi}_{\mathrm{B}}(x)\bar{\pi}_{\mathrm{B}}(x'). \tag{4.34}$$

By assuming this approximation, we can evaluate the deviation.

Proposition 4.7. Suppose that $j_{\text{est}} \sim \mathcal{N}\left(\bar{\pi}_{\text{B}}^{(i-1)}, \boldsymbol{V}\right)$. Then,

$$\Delta_{\text{ac}}\lambda^{(i)} = \Delta_{\text{ex}}\lambda^{(i)} + \frac{\alpha^2}{2}\text{Tr}\left(\boldsymbol{I}_\lambda \boldsymbol{V}\right) + O(\alpha^3) \tag{4.35}$$

$$\approx \Delta_{\text{ex}}\lambda^{(i)} + \frac{\alpha^2}{2}\text{Tr}\left(\boldsymbol{I}_\lambda \boldsymbol{V}\right), \tag{4.36}$$

where

$$I_\lambda(x, x') = \frac{\partial^2 \lambda(\bar{\pi}_{\text{B}})}{\partial\pi(x)\partial\pi(x')}. \tag{4.37}$$

We note that the second term of (4.35) is not positive since the concavity of λ implies that I_λ is negative semi-definite. Since $V = O(1/\tau_{\text{est}})$, the deviation $\frac{\alpha^2}{2}\text{Tr}\left(\boldsymbol{I}_\lambda \boldsymbol{V}\right)$ is negligible when the learning rate is sufficiently small. Precisely, we can neglect the deviation if $\alpha^2/\tau_{\text{est}} \ll 1$. We can regard this inequality as a trade-off between α and τ_{est} in relation to the efficiency of ancestral learning.

In Section 4.5, we focused on the case where $\tau_{\text{est}} \approx \infty$. However, a short τ_{est} is realistic and might be beneficial in biological systems. The advantageous point of a short τ_{est} is that an agent has more opportunities to increase the population fitness by the update of ancestral learning. On the other hand, the drawback is the decrease in the fitness gain $\Delta_{\text{ac}}\lambda^{(i)}$ compared to the case where $\tau_{\text{est}} \approx \infty$ due to the stochastic fluctuation of j_{emp}. Proposition 4.7 indicates that the decrease is $O(\alpha^2/\tau_{\text{est}})$. Therefore, an agent can keep the decrease small by adopting a small learning rate α, although such a small learning rate makes the learning speed slow. In other words, the decrease in memory size τ_{est} can be compensated by the decrease in the learning speed α. Since the decrease depends on the second order of α while it does on the first order of τ_{est}, an agent might prefer small α to large τ_{est}. Indeed, we showed that ancestral learning with small $\alpha = 0.01$ can acquire the optimal strategy even for $\tau_{\text{est}} = 1$ in Section 5.3. In such a situation, FF-thm (Theorem 4.6) is useful since the decrease of the fitness gain is negligible.

4.9 Connection with Sequential Monte-Carlo Methods

Before concluding this chapter, we point out the relationship between population dynamics and HMM. This point of view deepens our understanding of the population dynamics with learning and has potential for future applications.

Population dynamics (3.8) without learning is equivalent to the joint distribution of a hidden Markov model (3.44) up to the normalization factor if $\mathbb{T}_{\text{F}}(x^{(t)} \mid x^{(t-1)}) = \pi_{\text{F}}(x)$ and $K(x^{(t)} \mid y^{(t)}) = e^{k(x^{(t)}, y^{(t)})}$. This equivalence is known as an example of Feynman-Kac formula [68]. In addition to the model, the cumulative population fitness (3.12) is equivalent to log-likelihood (3.47). Specifically, the objective of learning by agents and that of the maximum likelihood estimation of $\mathbb{T}_{\text{F}}$ is equivalent.

This equivalence gives another explanation why ancestral information is sufficient to estimate the gradient of the population fitness. From the equivalence of (3.11) and (3.46) up to the normalization factor, we know that the ancestral information $j_{\text{emp}}^{\pi_{\text{F}}^{(i)}}$ (4.3) is equivalent to the average posterior distribution $1/\tau_{\text{est}} \cdot \sum_{t'=(i-1)\tau_{\text{est}}-1}^{i\tau_{\text{est}}-1} \mathbb{P}[x^{(t')} \mid \mathbb{Y}^{(t)}]$. If τ_{est} is sufficiently long,

$$\bar{\pi}_{\text{B}}(x) \approx j_{\text{emp}}^{\pi_{\text{F}}^{(i)}}(x) = 1/\tau_{\text{est}} \cdot \sum_{t'=(i-1)\tau_{\text{est}}-1}^{i\tau_{\text{est}}-1} \mathbb{P}[x^{(t')} \mid \mathbb{Y}^{(t)}], \tag{4.38}$$

since $y^{(t)}$ is i.i.d. Therefore, Proposition 3.2 is equivalent to Fisher's identity (Proposition 3.6). This is an proof of Theorem 4.1 from the view point of Feynman-Kac formula.

The equivalence also explains why ancestral learning can attain the optimal strategy. The update rule ${\pi_{\mathrm{F}}}^{(i)} \leftarrow \bar{\pi}_{\mathrm{B}}$ of ancestral learning when $\alpha = 1$ and $\tau_{\mathrm{est}} \approx \infty$ is equivalent to the update of π_{F} by EM-algorithm (3.49). Therefore, the populational fitness, which corresponds to log-likelihood, always increases by the update of ancestral learning since EM-algorithm always increases log-likelihood [14]. Since the populational fitness is concave, this is an alternative proof why ancestral learning acquires the optimal strategy.

4.10 Discussion

In this chapter, we investigated the acceleration of evolutionary process by individual learning from experience. We first introduced ancestral learning and numerically showed that ancestral learning accelerates the acquisition of the optimal strategy. We next prove that ancestral information j_{emp} is sufficient to estimate the gradient of the population fitness. Agents can estimate the gradient without communication among agents. We then quantified the acceleration of evolutionary process by ancestral learning via extending FF-thm of natural selection. Extended FF-thm decomposes the fitness gain into two terms, each of which corresponds to the gain by acquiring one of two representative strategies. We finally showed the trade-off between a learning rate and an update interval. All in all, we established a theoretical framework to discuss the acceleration of evolutionary process by individual learning from experience.

Since our framework is general, we may apply it to various biological problems. One example is the adaptation (micro-evolution) of microbes. Recent development in experimental techniques enables us to measure the phenotypic traits and their inheritance over generations at a single cell level. Examples are division times, sizes [113, 41], and sensitivity to chemical substances in chemotaxis [66]. Our framework may become a basis to understand such inheritance of phenotypic traits in the form of learning by cells. Moreover, our framework may be useful to design new experiments to measure and to understand the impact of learning. To analyze such data, modeling with continuous-time age-structured setting (3.27) and characterization of the convergence of the population level statistics in such model 3.39 might be useful [104, 76].

Our theory still has room for further improvement since there are some factors that might be useful to learn but we have not considered. An example is the type of a parent. Although individual agents use the type of parent in ancestral learning via the ancestral information j_{emp}, the type of parent can be used more directly. We can consider the situation where an agent expresses its type depending on the strategy and the type of the parent. The dependency might be useful when the environmental state is strongly correlated with the previous state. In such a situation, the strategy should be modeled as a Markov chain $\mathbb{T}_{\mathrm{F}}(x \mid x')$, where x' is the type of the parent, instead of the distribution $\pi_{\mathrm{F}}(x)$. To consider ancestral learning of $\mathbb{T}_{\mathrm{F}}$, the promising techniques are the variational principle, which we used in Section 4.4, for Markov chains in random environments [100, 52]. In addition, Feynman-Kac formula discussed in Section 4.9 is promising.

Another example is communication among agents. Although we showed that ancestral learning can accelerate the evolutionary process without communication, learning with communication might further accelerate the evolutionary process. In Section 4.8, we showed that the acceleration is small when τ_{est} is small due to the stochastic fluctuation of ancestral information j_{emp}. The communication might be beneficial to suppress the fluctuation.

The last example is the sensing of environmental states. In the context of population dynamics, researchers have considered the situation where an agent can receive a signal $z^{(t)}$ of environmental state $y^{(t)}$ and then expresses the type by using signal-dependent strategy $\pi_{\mathrm{F}}(x \mid z^{(t)})$. Since sensing is invidual learning from signals (See Section 1.1), the integration of sensing and learning from experience is important to fully understand the information processing of organisms. In such a situation, an agent might acquire the optimal signal-dependent strategy by extending ancestral learning. In addition, the signal of the environmental states may be useful to improve ancestral learning. To achieve such generalization, we should construct a framework that can treat the retrospective and prospective information processing of organisms by ancestral learning and sensing, respectively.

4.11 Derivations

In this section, we give proofs that are omitted from the main text.

4.11.1 Proofs in Section 4.4

Proof of Proposition 4.1. We prove Eq. (4.8) via the method of Lagrange multiplier. For sufficiently small ϵ, we need to solve the following linearized optimization:

$$\max_{\delta\pi}. \sum_{x\in\mathcal{X}} \frac{\bar{\pi}_{\mathrm{B}}(x)}{\pi_{\mathrm{F}}(x)} \delta\pi(x) \tag{4.39}$$

under the constraints $\pi_{\mathrm{F}} + \delta\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{D}[\pi_{\mathrm{F}} \| \pi_{\mathrm{F}} + \delta\pi] = \epsilon$. Since π_{F} is an interior point of $\mathcal{P}(\mathcal{X})$, the former constraint $\pi_{\mathrm{F}} + \delta\pi \in \mathcal{P}(\mathcal{X})$ is equivalent to $\sum_{x\in\mathcal{X}} \delta\pi(x) = 0$. For a sufficiently small ϵ, we can approximate $\mathcal{D}[\pi_{\mathrm{F}} \| \pi_{\mathrm{F}} + \delta\pi]$ by using the *Fisher information matrix* [4] as

$$\mathcal{D}[\pi_{\mathrm{F}} \| \pi_{\mathrm{F}} + \delta\pi] = \frac{1}{2} \sum_{x,x'\in\mathcal{X}} \delta\pi(x) \delta_{x,x'} \frac{1}{\pi_{\mathrm{F}}(x)} \delta\pi(x') = \frac{1}{2} \sum_{x\in\mathcal{X}} \frac{\delta\pi^2(x)}{\pi_{\mathrm{F}}(x)}. \tag{4.40}$$

Here, the Fisher information matrix is a $|\mathcal{X}| \times |\mathcal{X}|$ diagonal matrix with diagonal entries $\{1/\pi_{\mathrm{F}}(x)\}_{x\in\mathcal{X}}$. By using this approximation, the Lagrangian function is

$$L(\delta\pi; \lambda, \lambda') = \sum_{x\in\mathcal{X}} \frac{\bar{\pi}_{\mathrm{B}}(x)}{\pi_{\mathrm{F}}(x)} \delta\pi(x) + \lambda \left(\frac{1}{2} \sum_{x\in\mathcal{X}} \frac{\delta\pi^2(x)}{\pi_{\mathrm{F}}(x)} - \epsilon \right) + \lambda' \left(\sum_{x\in\mathcal{X}} \delta\pi(x) \right). \tag{4.41}$$

By differentiating L with respect to $\delta\pi(x)$, we have the stationary condition:

$$\frac{\partial L}{\partial \delta\pi(x)} = \frac{\bar{\pi}_{\mathrm{B}}(x)}{\pi_{\mathrm{F}}(x)} + \frac{\lambda \delta\pi(x)}{\pi_{\mathrm{F}}(x)} + \lambda' = 0, \tag{4.42}$$

for all $x \in \mathcal{X}$. By multiplying $\pi_{\mathrm{F}}(x)$ and taking sum $\sum_{x\in\mathcal{X}}$ of both sides of Eq. (4.42), we have

$$1 + \lambda' = 0. \tag{4.43}$$

We here used $\sum_{x\in\mathcal{X}} \delta\pi(x) = 0$. By rearranging Eq. (4.42) and substituting $\lambda' = -1$, we have

$$\delta\pi(x) = \frac{\pi_{\mathrm{F}}(x) - \bar{\pi}_{\mathrm{B}}(x)}{\lambda} \propto \bar{\pi}_{\mathrm{B}}(x) - \pi_{\mathrm{F}}(x). \tag{4.44}$$

□

4.11.2 Proofs in Section 4.5

Proof of Lemma 4.2. By a direct calculation,

$$\Delta\lambda^{(t)} = \log \sum_{x\in\mathcal{X}} e^{k(x)} p^{(t)}(x) - \log \sum_{x\in\mathcal{X}} e^{k(x)} p^{(t-1)}(x) \tag{4.45}$$

$$= \log \sum_{x\in\mathcal{X}} e^{k(x)} \frac{e^{k(x)} p^{(t-1)}(x)}{\sum_{x'\in\mathcal{X}} e^{k(x')} p^{(t-1)}(x')} - \log \sum_{x\in\mathcal{X}} e^{k(x)} p^{(t-1)}(x) \tag{4.46}$$

$$= \log \sum_{x\in\mathcal{X}} \left(e^{k(x)}\right)^2 p^{(t-1)}(x) - 2\log \sum_{x\in\mathcal{X}} e^{k(x)} p^{(t-1)}(x) \tag{4.47}$$

$$= \log \frac{\mathbb{E}_{p^{(t-1)}}\left[\left(e^{k(x)}\right)^2\right]}{\mathbb{E}_{p^{(t-1)}}\left[e^{k(x)}\right]^2} \tag{4.48}$$

$$= \log\text{-}\mathbb{V}_{p^{(t-1)}}[k(x)]. \tag{4.49}$$

□

Proof of Theorem 4.4. By a direct calculation,

$$\lambda({\pi_{\mathrm{F}}}^{(i)}) \tag{4.50}$$

$$= \mathbb{E}_{Q(y)}\left[\log \mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]\right] \tag{4.51}$$

$$= \mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]\right] \tag{4.52}$$

$$= \mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]\right] + \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{\mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]}{\mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}\right]. \tag{4.53}$$

We first treat the first term. By a similar argument to Eq. (4.10), the term inside the expectation satisfies the following relationship.

$$\log \mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right] - \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.54}$$

$$= \log \sum_{x\in\mathcal{X}} e^{k(x,y)} \frac{e^{k(x,y')}{\pi_{\mathrm{F}}}^{(i-1)}(x)}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}(x')}\left[e^{k(x',y')}\right]} - \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.55}$$

$$= \log \frac{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)+k(x,y')}\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y')}\right]} \tag{4.56}$$

$$= \log\text{-Cov}\left[k(x,y), k(x,y')\right]. \tag{4.57}$$

By taking average with respect to $Q(y)Q(y')$, we have

$$\mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]\right] - \lambda({\pi_{\mathrm{F}}}^{(i-1)}) \tag{4.58}$$

$$= \mathbb{E}_{Q(y)Q(y')}\left[\log\text{-Cov}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}, e^{k(x,y')}\right]\right]. \tag{4.59}$$

We next treat the second term of Eq. (4.53). By definition,

$$\frac{\bar{Q}^{(i)}(y' \mid y)}{Q(y')} = \frac{\sum_{x\in\mathcal{X}} e^{k(x,y)}{\pi_{\mathrm{B}}}^{(i-1)}(x \mid y')}{\sum_{x\in\mathcal{X}, y'\in\mathcal{Y}} e^{k(x,y)}{\pi_{\mathrm{B}}}^{(i-1)}(x \mid y')Q(y')} \tag{4.60}$$

$$= \frac{\mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}{\mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]}. \tag{4.61}$$

We thus have

$$\begin{aligned}
\mathbb{E}_{Q(y)Q(y')}\left[\log \frac{\mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]}{\mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}\right] &= \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{Q(y')}{\bar{Q}^{(i)}(y' \mid y)}\right] && (4.62)\\
&= \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{Q(y)Q(y')}{\bar{Q}^{(i)}(y' \mid y)Q(y)}\right] && (4.63)\\
&= \mathcal{D}\left[Q(y)Q(y')\middle\|\bar{Q}^{(i)}(y' \mid y)Q(y)\right]. && (4.64)
\end{aligned}$$

In conclusion, we proved Eq. (4.12). □

4.11.3 Proofs in Section 4.6

Proof of Lemma 4.5. By Eq. (4.60),

$$\begin{aligned}
\log \frac{Q(y)Q(y')}{\bar{Q}^{(i)}(y' \mid y)Q(y)} &= -\log \frac{\mathbb{E}_{{\pi_{\mathrm{B}}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}{\mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]} && (4.65)\\
&= -\log \frac{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)+k(x,y')}\right]}{\mathbb{E}_{\bar{\pi}_{\mathrm{B}}^{(i-1)}}\left[e^{k(x,y)}\right]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y')}\right]} && (4.66)\\
&= -\log \frac{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[F_y(x)F_{y'}(x)\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i)}}\left[F_y(x)\right]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[F_{y'}(x)\right]}. && (4.67)
\end{aligned}$$

By averaging with respect to $Q(y)Q(y')$, we have Eq. (4.26). □

4.11.4 Proofs in Section 4.8

Proof of Theorem 4.6. We can prove the theorem by almost the same argument as Theorem 4.4. Let $\bar{\pi}_{\alpha}^{(i-1)}(x) := \alpha\bar{\pi}_{\mathrm{B}}^{(i-1)}(x) + (1-\alpha){\pi_{\mathrm{F}}}^{(i-1)}(x)$. By a direct calculation, we have

$$\begin{aligned}
&\lambda({\pi_{\mathrm{F}}}^{(i)}) && (4.68)\\
&= \mathbb{E}_{Q(y)}\left[\log \mathbb{E}_{\bar{\pi}_{\alpha}^{(i-1)}}\left[e^{k(x,y)}\right]\right] && (4.69)\\
&= \mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{\bar{\pi}_{\alpha}^{(i-1)}}\left[e^{k(x,y)}\right]\right] && (4.70)\\
&= \mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{{\pi_{\alpha}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]\right] + \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{\mathbb{E}_{\bar{\pi}_{\alpha}^{(i-1)}}\left[e^{k(x,y)}\right]}{\mathbb{E}_{{\pi_{\alpha}}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}\right]. && (4.71)
\end{aligned}$$

We first treat the first term. By a similar argument to Eq. (4.10), the term inside the expectation satisfies

$$\log \mathbb{E}_{{\pi_\alpha}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right] - \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.72}$$

$$= \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}(x)}\left[e^{k(x,y)}\left(\alpha \frac{e^{k(x,y')}}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}(x')}\left[e^{k(x',y')}\right]} + 1 - \alpha\right)\right]$$
$$- \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.73}$$

$$= \log\left(\alpha \frac{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)+k(x,y')}\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y')}\right]} + (1-\alpha)\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right]\right) \tag{4.74}$$

$$- \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.75}$$

$$= \log\left(\alpha \frac{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)+k(x,y')}\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y')}\right]} + 1 - \alpha\right). \tag{4.76}$$

By (3.6),

$$\log \mathbb{E}_{{\pi_\alpha}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right] - \log \mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right] \tag{4.77}$$

$$= \log\left(1 + \alpha \frac{\mathrm{Cov}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}, e^{k(x,y')}\right]}{\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y)}\right]\mathbb{E}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[e^{k(x,y')}\right]}\right) \tag{4.78}$$

$$= \text{log-Cov}^{\alpha}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[k(x,y), k(x,y')\right]. \tag{4.79}$$

By taking average with respect to $Q(y)Q(y')$, we have

$$\mathbb{E}_{Q(y)Q(y')}\left[\log \mathbb{E}_{{\pi_\alpha}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]\right] - \lambda({\pi_{\mathrm{F}}}^{(i-1)}) \tag{4.80}$$

$$= \mathbb{E}_{Q(y)Q(y')}\left[\text{log-Cov}^{\alpha}_{{\pi_{\mathrm{F}}}^{(i-1)}}\left[k(x,y), k(x,y')\right]\right]. \tag{4.81}$$

We next treat the second term of Eq. (4.71). By definition,

$$\frac{\bar{Q}^{(i)}_{\alpha}(y' \mid y)}{Q(y')} = \frac{\sum_{x\in\mathcal{X}} e^{k(x,y)}{\pi_\alpha}^{(i-1)}(x \mid y')}{\sum_{x\in\mathcal{X},y'\in\mathcal{Y}} e^{k(x,y)}{\pi_\alpha}^{(i-1)}(x \mid y')Q(y')} \tag{4.82}$$

$$= \frac{\mathbb{E}_{{\pi_\alpha}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}{\mathbb{E}_{\bar{\pi}^{(i-1)}_{\alpha}}\left[e^{k(x,y)}\right]}. \tag{4.83}$$

Thus,

$$\mathbb{E}_{Q(y)Q(y')}\left[\log \frac{\mathbb{E}_{\bar{\pi}^{(i-1)}_{\alpha}}\left[e^{k(x,y)}\right]}{\mathbb{E}_{{\pi_\alpha}^{(i-1)}(x|y')}\left[e^{k(x,y)}\right]}\right] = \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{Q(y')}{\bar{Q}^{(i)}_{\alpha}(y' \mid y)}\right] \tag{4.84}$$

$$= \mathbb{E}_{Q(y)Q(y')}\left[\log \frac{Q(y)Q(y')}{\bar{Q}^{(i)}_{\alpha}(y' \mid y)Q(y)}\right] \tag{4.85}$$

$$= \mathcal{D}\left[Q(y)Q(y') \middle\| \bar{Q}^{(i)}_{\alpha}(y' \mid y)Q(y)\right]. \tag{4.86}$$

In conclusion, we proved Eq. (4.28). □

Proof of Proposition 4.7. The updated strategy ${\pi_{\rm F}}^{(i)} = \alpha j_{\rm est} + (1-\alpha){\pi_{\rm F}}^{(i-1)}$ satisfies

$$ {\pi_{\rm F}}^{(i)} \sim \mathcal{N}\left(\bar{\pi}_\alpha, \alpha^2 \boldsymbol{V}\right), \tag{4.87} $$

where we omitte the superscript of $\bar{\pi}_\alpha^{(i-1)}$ to avoid the complication. The growth rate is approximated as

$$ \lambda(\pi_\alpha + \delta\pi) = \lambda(\pi_\alpha) + \sum_{x\in\mathcal{X}} \frac{\partial\lambda}{\partial\pi(x)}\delta\pi(x) + \frac{1}{2}\sum_{x,x'\in\mathcal{X}} \delta\pi(x)\frac{\partial^2\lambda}{\partial\pi(x)\pi(x')}\delta\pi(x') + O(\delta\pi^3) \tag{4.88} $$

$$ = \lambda(\pi_\alpha) + \sum_{x\in\mathcal{X}} \frac{\partial\lambda}{\partial\pi(x)}\delta\pi(x) + \frac{1}{2}\sum_{x,x'\in\mathcal{X}} \delta\pi(x) I_\lambda(x,x')\delta\pi(x') + O(\delta\pi^3). \tag{4.89} $$

We note that $O(\delta\pi^3) = O(\alpha^3)$ by the update rule of ancestral learning. By this approximation,

$$ \Delta_{\rm ac}\lambda^{(i)} = \mathbb{E}\left[\lambda(j_{\rm est})\right] - \lambda^{(i-1)} \tag{4.90} $$

$$ = \lambda(\pi_\alpha) - \lambda^{(i-1)} + \mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\sum_{x\in\mathcal{X}} \frac{\partial\lambda}{\partial\pi(x)}\delta\pi(x)\right] + \frac{1}{2}\mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\sum_{x,x'\in\mathcal{X}} \delta\pi(x) I_\lambda(x,x')\delta\pi(x')\right] + O(\alpha^3) \tag{4.91} $$

$$ = \Delta_{\rm ex}\lambda^{(i)} + \frac{1}{2}\mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\sum_{x,x'\in\mathcal{X}} \delta\pi(x) I_\lambda(x,x')\delta\pi(x')\right] + O(\alpha^3). \tag{4.92} $$

In the last equation, the third term vanishes because

$$ \mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\sum_{x\in\mathcal{X}} \frac{\partial\lambda}{\partial\pi(x)}\delta\pi(x)\right] \tag{4.93} $$

$$ = \sum_{x\in\mathcal{X}} \frac{\partial\lambda}{\partial\pi(x)}\mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\delta\pi(x)\right] \tag{4.94} $$

$$ = 0. \tag{4.95} $$

By the usual matrix calculation [83],

$$ \mathbb{E}_{\mathcal{N}(0,\alpha^2\boldsymbol{V})}\left[\sum_{x,x'} \delta\pi(x) I_\lambda(x,x')\delta\pi(x')\right] \tag{4.96} $$

$$ = \alpha^2 {\rm Tr}\left(\boldsymbol{I}_\lambda \boldsymbol{V}\right). \tag{4.97} $$

In all, we proved (4.35). □

Chapter 5

Theoretical Analysis of Evolutionary Algorithms via Techniques from Population Dynamics

In this chapter, we aim to solve the problems about the theoretical guarantees of the evolutionary algorithms (See Section 1.4.2 and Section 1.5.2). See Section 5.6 for the proofs that we omit from the main text.

We consider the following minimization problem in this chapter:

$$\begin{array}{ll} \text{minimize} & f(X), \\ \text{subject to} & x \in \mathcal{X} \subseteq \mathbb{R}^d. \end{array} \tag{5.1}$$

We do not assume any property of f at this moment. We consider general f in Section 5.3 and possibly not strongly convex f in Section 5.4. When we assume that f is a convex function, let x^* be a minimizer.

5.1 Parallel Algorithm and Branching Algorithm

We address the two issues raised in Section 1.4.2 by considering the relative evaluation in Section 1.5.2: We will compare the BA with individual learning $\mathcal{L}$ to parallel execution of $\mathcal{L}$. For this comparison, we clarify the individual learning rule $\mathcal{L}$ in this context. Individual learning in the context is update rules of the iterative optimization algorithms (See Section 2.1.2). We use stochastic iterative optimization algorithms since individual learning in this context is a generalization of mutation, which should generate a variety of solutions. We also focus on the iterative optimization algorithm $\mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)})$ that can be represented as a time-dependent Markov chain. Examples are the update rule of GD (3.58) and SGD (3.60). Another example is random update used as mutation in conventional evolutionary algorithms. The inclusion of random update enables us to analyze the conventional evolutionary algorithm in our framework.

We first introduce the parallel execution of iterative optimization algorithm $\mathcal{L}$ since it is simpler than the BA. A *Parallel Algorithm (PA)* (Algorithm 5.1) execute a given iterative optimization algorithm $\mathcal{L}^{(t)}$ on N_{size}-computational nodes independently. The algorithm first samples a set of initial solutions $\{x_i^{(0)}\}_{i \in [N_{\mathrm{size}}]}$ from a certain initial distribution $\nu(x)$. After that, each solution $x_i^{(t)}$ is recursively updated by $\mathcal{L}^{(t)}$. The solution $x_i^{(t)}$ at time $t+1$ is sampled from the distribution $\mathcal{L}^{(t)}(x_i^{(t+1)} \mid x_i^{(t)})$.

We next introduce the BA (Algorithm 5.2; Figure 5.1). The BA has hyperparameters $\mathbb{\beta}_g{}^{(T)} = \{\beta^{(t)}\}_{t=1,2,\ldots,T}$ with $\beta^{(t)} \geq 0$ for all t. Here, the subscript g stands for "growth". The BA first samples the population of the tentative solutions $\{{x'}_i^{(0)}\}_{i\in[N_{\mathrm{size}}]}$ from $\nu(x)$. After that, the set of initial solutions is constructed by sampling (Step 2 in Algorithm 5.2): For each $i \in [N_{\mathrm{size}}]$, the next solution $x_i^{(0)}$ is independently sampled from the population $\{{x'}_i^{(0)}\}_{i\in[N_{\mathrm{size}}]}$ with weight [1] $e^{-{\beta_g}^{(0)} f({x'}_i^{(0)})}$. After that, BA iterates the loop of learning and sampling: At learning (Step 4 in Algorithm 5.2), the BA updates the solution $x_i^{(t)}$ by $\mathcal{L}^{(t)}$: A tentative solution ${x'}_i^{(t+1)}$ is sampled from $\mathcal{L}^{(t)}({x'}_i^{(t+1)} \mid x^{(t)})$. After that, the population at the next step is constructed by sampling (Step 5 in Algorithm 5.2), which is the same procedure as Step 2 in Algorithm 5.2. We call $x_i^{(t+1)}$ is a *daughter* of $x_j^{(t)}$ if $x_i^{(t+1)} = {x'}_j^{(t+1)}$. We call $x_j^{(t)}$ is a *parent* of $x_i^{(t+1)}$ if $x_i^{(t+1)}$ is the daughter of $x_j^{(t)}$.

We note that we can implement the sampling step by using copies of the solutions. In this implementation, we first make Z_i-copies of ${x'}_i^{(t+1)}$, where $Z_i^{(t+1)}$ be any random variable in $\mathbb{R}_{\geq 0}$ with $\mathbb{E}[Z_i^{(t+1)}] = e^{-{\beta_g}^{(t+1)} f({x'}_i^{(t+1)})}$. After that, we select N_{size}-solutions uniformly at random from the copied solutions. This procedure implements the sampling step. In this sense, the factor $e^{-{\beta_g}^{(t+1)} f(x^{(t+1)})}$ can be interpreted as populational evolution.

We also note that the BA satisfies the Markov property: The time evolution of the solution $x^{(t)}$ is not affected by events at time t' ($t' < t$) if the population of the solutions at time t is given.

We explain the reason that we use the growth factor of the form $e^{-\beta f(x)}$ instead of the conventional choice like $f(x)$. It is because the BA satisfies the invariance property with respect to the affine transformation of the objective function $f(x)$, which is desirable for evolutionary algorithms [81]. The scalar addition $f_{+c}(x) := f(x) + c$ for some constant $c \in \mathbb{R}$ does not change the behavior of the algorithm since the weight $e^{-{\beta_g}^{(t+1)} f_{+c}({x'}_i^{(t+1)})}$ is the same as the original weight up to the constant factor e^c. Also, the scalar multiplication $f_{\times a}(x) := af(x)$ for $a > 0$ does not change the behavior of the BA substantially: the behavior of the BA for f with ${\beta_g}^{(t+1)}$ is the same as that for $f_{\times a}$ with ${\beta_g}^{(t+1)}/a$.

Before proceeding, we discuss when we use the BA with SGD since we adopt it as examples in Section 5.3.4 and Section 5.4.3. Since the BA requires the exact value of $f(x)$, we can compute $\nabla f(x)$ by numerical differentiation. Although using SGD seems to be irrational when we have exact $\nabla f(x)$, we have the following example where we use the BA with SGD. The situation is where the dimension d of x is large and we want to save the time of numerical differentiation. Suppose that $f(x)$ is of the form (2.2) and the time to compute $f(x)$ is $O(nh(d))$ for some function h. We also assume that we do not have an analytical form of $\nabla f(x)$. In this situation, the numerical differentiation requires $O(ndh(d))$-computational time. When both n and d is large, we might want to reduce the time to compute $\nabla f(x)$ by using a mini-batch (2.3). Indeed, if the size of a mini-batch is $O(1)$, we can compute the estimator g_S of the gradient in $O(dh(d))$-time.

[1] If the weight is too small or too large, we can use $e^{-{\beta_g}^{(0)} f({x'}_i^{(0)}) + c^{(0)}}$ as an equivalent weight, where $c^{(0)} \in \mathbb{R}$ is a certain constant.

Algorithm 5.1 Simple Parallel Algorithm

```
Sample initial solutions $\{x_i^{(0)}\}_{i \in [N_{\text{size}}]}$ from distribution $\nu(x)$.
for $t = 0, 1, \ldots, T-1$ do
    Update each solution by $\mathcal{L}^{(t)}$: $x_i^{(t+1)} \sim \mathcal{L}^{(t)}(x \mid {x'}_i^{(t)})$ for each $i \in [N_{\text{size}}]$.
end for
```

Algorithm 5.2 Branching Algorithm

```
Sample $\{{x'}_i^{(0)}\}_{i \in [N_{\text{size}}]}$ from distribution $\nu(x)$.
For each $i \in [N_{\text{size}}]$, independently sample the initial solution $x_i^{(0)}$ from the population $\{{x'}_i^{(0)}\}_{i \in [N_{\text{size}}]}$ with weight $e^{-\beta^{(0)} f({x'}_i^{(0)})}$.
for $t = 0, 1, \ldots, T-1$ do
    Update each solution by learning: ${x'}_i^{(t+1)} \sim \mathcal{L}^{(t)}({x'}_i^{(t)} \mid x_i^{(t)})$ for each $i \in [N_{\text{size}}]$.
    For each $i \in [N_{\text{size}}]$, independently sample the solution $x_i^{(t+1)}$ from the population $\{{x'}_i^{(t+1)}\}_{i \in [N_{\text{size}}]}$ with weight $e^{-\beta^{(t+1)} f({x'}_i^{(t+1)})}$.
end for
```

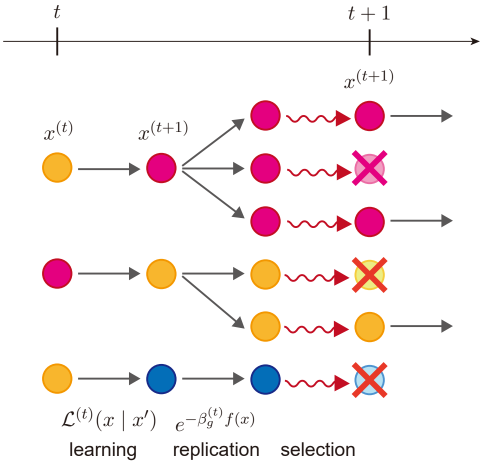


Figure 5.1: Schematic representation of the BA. In this figure, each circle represents a solution and its color does the value of $x^{(t)}$. Each solution in the population is updated by learning rule $\mathcal{L}^{(t)}$. Then, a solution x reproduces $e^{-{\beta_g}^{(t)} f(x)}$-daughters on average. Finally, N_{size} solutions are selected uniformly at random.

5.2 Infinite Population Approximation and Pathwise Formulation

Following the approach of [45, 110], we analyze the performance of the PA and the BA in the situation where the size N_{size} of the population is sufficiently large. See Section 5.5 for a further discussion about this assumption. Under this assumption, we can approximate the BA as a dynamical system. This approximation admits a pathwise formulation of the BA, which we use for further analysis throughout the chapter.

We first introduce the pathwise formulation of PA as a basis of that of BA (Figure 5.2 (a)). After the execution of PA, we have N_{size}-sequences $\{x_i^{(t)}\}_{t=0,1,\ldots,T}$ ($i \in [N_{\text{size}}]$) of solutions. We call each of the sequence $\mathbb{X}_i^{(T)} = \{x_i^{(t)}\}_{t=0,1,\ldots,T}$ a *path* of the solutions. By the definition of the algorithm, the solutions $\{x_i^{(t)}\}_{t=0,1,\ldots,T}$ on the i-th computational node is a Markov chain generated by $\mathcal{L}^{(t)}$. Therefore, the probability $\mathbb{P}_{\text{F}}[\mathbb{X}^{(t)}]$ that a realization of the i-th path $\{x_i^{(t)}\}_{t=0,1,\ldots,T}$ equals $\mathbb{X}^{(T)} = \{x^{(t)}\}_{t=0,1,\ldots,T}$ is given by

$$\mathbb{P}_{\text{F}}[\mathbb{X}^{(t)}] = \nu(x^{(0)}) \prod_{t=0}^{T-1} \mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)}) \tag{5.2}$$

$$= \prod_{t=-1}^{T-1} \mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)}), \tag{5.3}$$

where we specially define $\mathcal{L}^{(-1)}(x^{(0)} \mid x^{(-1)}) := \nu(x^{(0)})$ for notational simplicity.

We next consider the infinite population approximation of the BA. We will approximate the empirical distribution $j(x)$ of the solution at the end of the t-th loop of the BA. The empirical distribution is defined by

$$j^{(t)}(x) := \frac{1}{N_{\text{size}}} \sum_{i \in [N_{\text{size}}]} \delta_{x_i^{(t)}, x}, \tag{5.4}$$

where $\delta_{x,x'}$ is the delta function. In the limit $N_{\text{size}} \to \infty$, the empirical distribution converges to its expectation $\mathbb{P}_{\text{B}}^{(t)}$ that evolves as follows:

$$\mathbb{P}_{\text{B}}^{(t+1)}(x) = \frac{\sum_{x' \in \mathcal{X}} e^{-\beta_g{}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x') \mathbb{P}_{\text{B}}^{(t)}(x')}{\sum_{x', x'' \in \mathcal{X}} e^{-\beta_g{}^{(t+1)} f(x'')} \mathcal{L}^{(t)}(x'' \mid x') \mathbb{P}_{\text{B}}^{(t)}(x')}, \tag{5.5}$$

where we specially define

$$\mathbb{P}_{\text{B}}^{(0)}(x) := \frac{e^{-\beta_g{}^{(0)} f(x)} \nu(x)}{\sum_{x'' \in \mathcal{X}} e^{-\beta_g{}^{(0)} f(x'')} \nu(x'')}. \tag{5.6}$$

We call $\mathbb{P}_{\text{B}}$ the infinite population approximation of the BA. See Section 5.6 for the derivation of Equations (5.5) and (5.6).

We simplify the infinite population approximation (5.5) by considering the pathwise formulation of BA. To introduce a pathwise formulation, we first define the path of the BA (Figure 5.2 (b)). The path of the BA is a sequence of the solutions followed backwardly from a solution $x_i^{(T)}$ at time T. Precisely, let us take any solution $x_{i(T)}^{(T)}$ at the end of the algorithm. We recursively define that $x_{i(t)}^{(t)}$ is a parent of $x_{i(t+1)}^{(t+1)}$ for $t = T-1, T-2, \ldots, 0$. Then, a *path* of the BA with respect to the solution $x_{i(T)}^{(T)}$ is the sequence $\{x_{i(t)}^{(t)}\}_{t=0,1,\ldots,T}$. We note that the paths of the BA might join while those of the PA are independent (Figure 5.2).

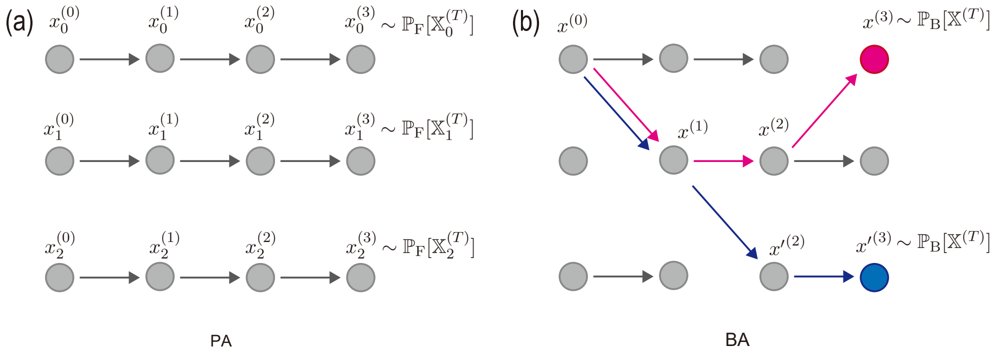

(a)
$x_0^{(0)}$
$x_0^{(1)}$
$x_0^{(2)}$
$x_0^{(3)} \sim \mathbb{P}_{\mathrm{F}}[\mathbb{X}_0^{(T)}]$
$x_1^{(0)}$
$x_1^{(1)}$
$x_1^{(2)}$
$x_1^{(3)} \sim \mathbb{P}_{\mathrm{F}}[\mathbb{X}_1^{(T)}]$
$x_2^{(0)}$
$x_2^{(1)}$
$x_2^{(2)}$
$x_2^{(3)} \sim \mathbb{P}_{\mathrm{F}}[\mathbb{X}_2^{(T)}]$
PA
(b)
$x^{(0)}$
$x^{(3)} \sim \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)}]$
$x^{(1)}$
$x^{(2)}$
$x'^{(2)}$
$x'^{(3)} \sim \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)}]$
BA

Figure 5.2: Schemetic representation of the pathwise-formulation of the PA (a) and the BA (b). In the PA, each path $\mathbb{X}_i^{(T)} = \{x_i^{(t)}\}_{t=0,1,\ldots,T}$ $(i \in [N_{\mathrm{size}}])$ independently follows $\mathbb{P}_{\mathrm{F}}$ defined by (5.2). In (b), the head of the arrow indicates a daughter and the tail does the parent. In the BA, we choose a solution at time T and consider the path $\mathbb{X}^{(T)}$ followed backwardly from the chosen solution. For example, the solutions connected by the red arrows are the path followed backwardly from the red solution $x^{(3)}$. In BA, the paths are not independent in general: For different solutions $x^{(T)}$ and $x'^{(T)}$ at time T, the paths may join. For example, the path with respect to the red solution $x^{(3)}$ and that with respect to the blue solution $x'^{(3)}$ join. The path $\mathbb{X}^{(T)}$ follows the distribution $\mathbb{P}_{\mathrm{B}}$ biased from $\mathbb{P}_{\mathrm{F}}$. The relationship between $\mathbb{P}_{\mathrm{F}}$ and $\mathbb{P}_{\mathrm{B}}$ is given by (5.7).

We consider the probability $\mathbb{P}_{\mathrm{R}}[\mathbb{X}^{(T)}]$ that a realization of a path of the BA is $\mathbb{X}^{(T)} = \{x^{(t)}\}_{t=0,1,\ldots,T}$. Let $f[\mathbb{X}^{(T)}] = \{f(x^{(t)})\}_{t=0,1,\ldots,T}$ be the path of the value of the objective function evaluated along $\mathbb{X}^{(T)}$. By recursively applying (5.5), we have the following.

Proposition 5.1 (Pathwise Formulation of Branching Algorithm)**.**

$$\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)}] = \frac{e^{-\langle {\beta_g}^{(T)}, f[\mathbb{X}^{(T)}]\rangle}\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}]}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\langle {\beta_g}^{(T)}, f[\mathbb{X}'^{(T)}]\rangle}\mathbb{P}_{\mathrm{F}}[\mathbb{X}'^{(T)}]}. \tag{5.7}$$

The probability $\mathbb{P}_{\mathrm{B}}$ is biased from $\mathbb{P}_{\mathrm{F}}$ due to the effect of populational evolution. In other words, the probability $\mathbb{P}_{\mathrm{B}}$ incorporates the effect of populational evolution. Therefore, the pathwise formulation reduces the problem of the comparison of the PA and the BA to the comparison of the properties of the probability measures $\mathbb{P}_{\mathrm{F}}$ and $\mathbb{P}_{\mathrm{B}}$. The pathwise formulation is simpler than (5.5) since the effect $\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}]$ of individual learning and that $e^{-\langle {\beta_g}^{(T)}, f[\mathbb{X}^{(T)}]\rangle}$ of populational evolution are separated.

We discuss the connection between these results and population dynamics. This difference equation (5.5) is the same as the following population dynamics of $N^{(t)}(x)$ up to the normalization factor.

$$N^{(t+1)}(x) = \sum_{x' \in \mathcal{X}} e^{-{\beta_g}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x') N^{(t)}(x'). \tag{5.8}$$

Since the population dynamics is insightful, especially in Section 5.4.1, we explain its interpretation in detail. The quantity $N^{(t)}(x)$ can be interpreted as the expected number of the solution x at time t when we execute the BA without caring capacity as follows. Suppose that we implement the sampling step by using copies of the solutions (See section 5.1). We consider the modified implementation that the population at the next step is all copied solutions instead of the randomly selected N_{size}-solutions. We call this modification the BA without caring capacity. Let us consider the number of the solution $j'^{(t)}(x) := \sum_i \delta_{x_i^{(t)},x}$ in the population $\{x_i^{(t)}\}_i$ at the t-th step of the BA without caring capacity. By a similar argument to (5.5), we know that $j'^{(t)}(x)$ converges to its expectation $N^{(t)}(x)$ satisfying (5.8) as $N_{\text{size}} \to \infty$. The correspondence between (5.5) and (5.8) shows that we can neglect the effect of selection in the limit $N_{\text{size}} \to \infty$. Also, the correspondence enables us to calculate $\mathbb{P}_{\text{B}}^{(t)}(x)$ by normalizing $N^{(t)}(x)$. This calculation is useful because $N^{(t)}(x)$ is more intuitive than $\mathbb{P}_{\text{B}}^{(t)}(x)$. This correspondence is extended to the path level ((3.11) and (5.7)).

5.3 Populational Evolution can Accelerate Stochastic Optimization

In this section, we prove that the BA always performs better than the PA by using the pathwise formulation (5.7). In particular, we extend FF-thm for natural selection to the BA.

5.3.1 Criterion to Compare Parallel and Branching Algorithms

To compare the performance of the PA and the BA quantitatively, we need a criterion to measure the performance of the PA and the BA. Therefore, we first introduce such a criterion. Since both the PA and the BA yields N_{size}-paths $\{\mathbb{X}_i^{(T)}\}_{i\in[N_{\text{size}}]}$ of the solutions, the criterion must (1) measure the quality of a path $\mathbb{X}^{(T)}$ and (2) integrate the qualities of each path measured by (1). For the first point (1), a trivial measure of the quality is $f(x^{(T)})$. However, the intermediate progress $f(x^{(t)})$ ($t < T$) of the algorithm might be important in addition to $f(x^{(T)})$. For example, such intermediate progress is important when we use SGD with the averaging over time (Section 2.1.2). We therefore need a more sophisticated measure $c\colon \mathcal{X}^{(T+1)} \to \mathbb{R}$ of the quality that converts the whole path $\mathbb{X}^{(T)}$ to a scalar. For the second part (2), the integration is difficult because applying the measure of the quality for each path yields $\{c[\mathbb{X}^{(T)}]\}_{i\in[N_{\text{size}}]} \in \mathbb{R}^{N_{\text{size}}}$, which is not comparable in general since $\mathbb{R}^{N_{\text{size}}}$ is not totally ordered. We therefore need a function $\pi\colon \mathbb{R}^k \to \mathbb{R}$ to compare the measured qualities of the paths. Typical choices are the average and the minimum functions. The former is suitable for parallelized SGD with the averaging over paths (Section 2.1.2). On the other hand, the minimum function is another natural candidate especially for non-convex functions. To make the criterion general, the function π should be an unification of the average and the minimum functions. In addition, since the minimum function behaves singularly, the function π should be a smoothed version of the minimum function. For example, the minimum of $(f(x_0^{(T)}), f(x_1^{(T)}), \ldots, f(x_{N_{\text{size}}-1}^{(T)}))$ converges to $f(x^*)$ as $N_{\text{size}} \to \infty$ under a mild assumption on $\mathcal{L}^{(t)}$.

Let us consider the first part (1). To measure the quality of the solutions over the whole path, we use a *weighted path-level average*

$$\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle, \tag{5.9}$$

where $\mathbb{\beta}_w{}^{(T)} := \{\beta_w{}^{(0)}, \beta_w{}^{(1)}, \ldots, \beta_w{}^{(T)}\}$ is a weight with $\beta_w{}^{(t)} \geq 0$ for all t. Here, the subscript w stands for “weight”. A typical choice of the weight is $\mathbb{1}^{(T)} :=$

$\{1, 1, \ldots, 1\} \in \mathbb{R}^{(T+1)}$. If we are only interested in $x^{(T)}$, we can use the weight defined by ${\beta_w}^{(T)} = 1$ and ${\beta_w}^{(t)} = 0$ otherwise.

We next consider the second part (2). To compare the qualities of the multiple paths measured by the weighted path-level average, we introduce a smooth minimum that unifies the average and the minimum functions. For $z_i \in \mathbb{R}$ ($i \in [N_{\text{size}}]$), we define a *smooth minimum* by

$$\text{smin}^{\beta_c}(z_0, z_1, \ldots, z_{N_{\text{size}}-1}) := -\frac{1}{\beta_c} \log \left[\frac{1}{N_{\text{size}}} \sum_{i \in [N_{\text{size}}]} e^{-\beta_c z_i} \right]. \tag{5.10}$$

Here, $\beta_c > 0$ is a hyperaparameter and c stands for “criteion”.

The smooth minimum is indeed an unification of the average and the minimum functions:

Proposition 5.2.

$$\lim_{\beta_c \to 0+} \text{smin}^{\beta_c}(z_0, z_1, \ldots, z_{N_{\text{size}}-1}) = \frac{1}{N_{\text{size}}} \sum_{i \in [N_{\text{size}}]} z_i, \tag{5.11}$$

$$\lim_{\beta_c \to \infty} \text{smin}^{\beta_c}(z_0, z_1, \ldots, z_{N_{\text{size}}-1}) = \min_{i \in [N_{\text{size}}]} z_i. \tag{5.12}$$

To connect the smooth minimum and the infinite population approximation, we introduce a smooth minimum for $N_{\text{size}} \to \infty$. For a real random variable Z with distribution $p(z)$, we define the smooth minimum of Z by

$$\text{smin}_p^{\beta_c}[Z] := -\frac{1}{\beta_c} \log \mathbb{E}_p \left[e^{-\beta_c Z} \right]. \tag{5.13}$$

The smooth minimum for $N_{\text{size}} \to \infty$ satisfies the same property as Proposition 5.2.

Proposition 5.3. If $\mathbb{E}[Z]$ and $\mathbb{E}[Z^2 e^{-\beta_c Z}]$ exist for all sufficiently small $\beta_c > 0$, then

$$\lim_{\beta_c \to 0+} \text{smin}_p^{\beta_c}[Z] = \mathbb{E}_p[Z]. \tag{5.14}$$

Also,

$$\lim_{\beta_c \to \infty} \text{smin}_p^{\beta_c}[Z] = \inf\{z \mid F_Z(z) > 0\}, \tag{5.15}$$

where F_Z is the cumulative distribution function of Z.

In conclusion, we measure the performance of the PA and the BA by combining the weighted path-level average and the smooth minimum. The performance of the PA is

$$\text{PA}^f_{\mathcal{L}^{(t)}} := \text{smin}^{\beta_c}_{\mathbb{P}_\text{F}} \left[\left\langle \mathbb{B}_w^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle \right], \tag{5.16}$$

and the performance of BA is

$$\text{BA}^f_{\mathcal{L}^{(t)}} := \text{smin}^{\beta_c}_{\mathbb{P}_\text{B}} \left[\left\langle \mathbb{B}_w^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle \right]. \tag{5.17}$$

We note that these quantities are similar to the cumulative fitness 3.12 if we interpret $e^{-{\beta_g}^{(t)} f(x)}$ as the individual fitness.

Before proceeding, we note that the weighted path-level average and the smooth minimum are natural choices although there are other alternatives. The weighted path-level average is a natural measure since it has a connection to averaging over time. If we have a bound

$$\frac{\left\langle {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle}{\sum_{t=0}^{T} {\beta_w}^{(t)}} - f(x^*) \leq C, \tag{5.18}$$

and f is convex, then the Jensen inequality implies that

$$f(\bar{x}^{(T)}) - f(x^*) \leq C, \tag{5.19}$$

where

$$\bar{x}^{(T)} := \frac{\sum_{t=0}^{T} {\beta_w}^{(t)} x^{(t)}}{\sum_{t=0}^{T} {\beta_w}^{(t)}}. \tag{5.20}$$

Therefore, the bound for the weighted path-level average implies the bound for the output $\bar{x}^{(T)}$ averaged over time. This technique is used in many previous studies to show the bound for SGD with averaging over time [124, 102, 56, 7, 117]. It means that we might be able to utilize the proof of the previous studies to bound the weighted path-level average.

The smooth minimum is a natural choice for the following reasons. The smooth minimum is similar to the LogSumExp function and an example of Kolmogorov's mean [26]. Also, the smooth minimum is conjugate to the negative Shannon entropy [17] and its derivative is the SoftMax function. Moreover, the smooth minimum is known as the Helmholtz free energy in statistical mechanics [123].

5.3.2 Fisher's Fundamental Theorem of Branching Algorithm

We compare the performance of the PA and the BA measured by the criterion in the previous section. To address the second issue raised in Section 1.4.2, we evaluate the difference $\mathrm{PA}^f_{\mathcal{L}^{(t)}} - \mathrm{BA}^f_{\mathcal{L}^{(t)}}$ of two performances by using the pathwise formulation (5.7).

Theorem 5.4 (Fisher's Fundamental Theorem of Branching Algorithm)**.**

$$\mathrm{PA}^f_{\mathcal{L}^{(t)}} - \mathrm{BA}^f_{\mathcal{L}^{(t)}} = \frac{1}{\beta_c} \operatorname{log-Cov}_{\mathbb{P}_\mathrm{F}} \left[-\left\langle {\mathbb{\beta}_g}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle, -\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle \right]. \tag{5.21}$$

The theorem implies that BA always performs better than the PA. Indeed, since ${\mathbb{\beta}_g}^{(T)}$ is a hyperparameter of the BA, we can choose it arbitrary. In particular, we can choose ${\mathbb{\beta}_g}^{(T)} = \beta_g {\mathbb{\beta}_w}^{(T)}$ for a given ${\mathbb{\beta}_w}^{(T)}$, where $\beta_g > 0$ is some constant. For this choice of ${\mathbb{\beta}_g}^{(T)}$, we have the following bound.

Corollary 5.5. If ${\mathbb{\beta}_g}^{(T)} = \beta_g {\mathbb{\beta}_w}^{(T)}$, then

$$\mathrm{PA}^f_{\mathcal{L}^{(t)}} - \mathrm{BA}^f_{\mathcal{L}^{(t)}} = \frac{1}{\beta_c} \operatorname{log-Cov}_{\mathbb{P}_\mathrm{F}} \left[-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle, -\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle \right] \geq 0. \tag{5.22}$$

5.3.3 Simplified Calculations When Performance Satisfies Central Limit Theorem

The performance of the PA and the BA has a simple form when we can assume the normality of $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ by using the CLT. To demonstrate it, we first prove the following formula for the smooth minimum and FF-thm of the BA when $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ is a normal random variable.

Proposition 5.6. Suppose that $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then,

$$\mathrm{smin}^{\beta_c}[Z] = \mu - \frac{\beta_c \sigma^2}{2}. \tag{5.23}$$

Corollary 5.7 (Corollary of Theorem 5.4)**.** Suppose that $\mathbb{\beta}_g{}^{(T)} = \beta_g \mathbb{\beta}_w{}^{(T)}$ and $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ follows $\mathcal{N}\left(\mu, \sigma^2\right)$ under the probability measure $\mathbb{P}_{\mathrm{F}}$. Then,

$$\mathrm{PA}^f_{\mathcal{L}^{(t)}} - \mathrm{BA}^f_{\mathcal{L}^{(t)}} = \beta_c \beta_g \sigma^2. \tag{5.24}$$

Intuitively, Corollary 5.7 indicates that the performance of the BA is better when the variance of the performance of the PA is larger, namely, the learning rule $\mathcal{L}$ is exploratory.

These results help us to prove the upper bound of the performance of the PA and the BA. Since $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ is a sum of random variables, we can use the CLT and assume the normality when the dependency between the random variables is weak (Section 3.6). Such an assumption holds when $\mathcal{L}^{(t)}$ is SGD and the objective function is strongly convex (See Section 5.3.4) or when $\mathcal{L}^{(t)}$ is a random update that satisfies a certain ergodicity [47]. Under this assumption, we can calculate the performance via Proposition 5.6. To bound the right hand side of (5.23) for $Z = \left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$, what we have to prove is the followings two bounds under $\mathbb{P}_{\mathrm{F}}$ or $\mathbb{P}_{\mathrm{B}}$: (1) the upper bound for the mean of $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$; and (2) the lower bound for the variance of $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$. Let us first consider the case of $\mathbb{P}_{\mathrm{F}}$. The former (1) is usually proven by the previous studies as we discussed in Section 5.3.1. Therefore, the remaining task is the latter (2). In other words, previous studies do half of the jobs to bound the smooth minimum. In the case of $\mathbb{P}_{\mathrm{B}}$, we can calculate in a similar way by Corollary 5.7. We will see this procedure by using an example in Section 5.3.4.

We note that the assumption of the normality is not justified for too large β_g and β_c even if $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ follows a normal distribution with a small error under $\mathbb{P}_{\mathrm{F}}$. In the proof of Corollary 5.7, we use the normality of $Z_1 := \left\langle \beta_g \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ and $Z_2 := \left\langle \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ that follow from the normality of $Z_3 := \left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$. The deviations of Z_1 and Z_2 from normal distributions are greater than that of Z_3 when β_c and β_g are large. Since Z_3 is not exactly a normal random variable, this enlargement of the deviation makes the assumption of normality unjustified when β_c and β_g are too large. Indeed, since $f(x)$ is bounded from below while the support of the normal distribution is $\mathbb{R}$, the normality of $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ under $\mathbb{P}_{\mathrm{F}}$ does not hold exactly. This argument resolves the following confusion about Corollary 5.7. At first sight, Corollary 5.7 seems to contradict to the fact that $f(x)$ is bounded from below by $f(x^*)$: It seems that we can make $\mathrm{BA}^f_{\mathcal{L}^{(t)}}$ arbitrary small by taking large β_g and β_c. This is not true since we cannot apply Corollary 5.7 to the case of too large β_c and β_g.

5.3.4 Application to Stochastic Gradient Descent

Let us consider a simple example to see the consequence of FF-thm. Let $f(x) = \|x\|^2/2$. We note that $x^* = \mathbf{0}$ and $f(x^*) = 0$. We assume that the noise $\xi^{(t)}$ independently follows $\mathcal{N}\left(0, a^2 I/d\right)$. We set $\mathcal{L}^{(t)}$ to SGD (Algorithm 3.2) with

$$\eta^{(t)} = \eta, \tag{5.25}$$

$$\beta_w{}^{(t)} = 1/(T+1), \tag{5.26}$$

where η is some constant.

We first check that $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$ satisfies CLT. Namely, we check that $Z^{(t)} = f(x^{(t)}) - \mathbb{E}[f(x^{(t)})]$ is a mixingale (See Section 3.6). Let $\mathcal{F}^{(t)}$ be the σ-algebra generated by $\{\xi^{(t')}\}_{t' \leq t}$. In addition, let $Z^{(t)} = f(x^{(t)}) - \mathbb{E}[f(x^{(t)})]$. We suppose that $\xi^{(t)}$ is bounded [2]. Under this setting, we have the following:

Lemma 5.8. If $\|\xi^{(t)}\| \leq C$, then

$$\mathbb{E}[Z^{(t)} \mid \mathcal{F}^{(t-k)}] = O((1-\eta)^{2k}). \tag{5.27}$$

Therefore, $Z^{(t)}$ is a mixingale with size $-1/2$ and $Z^{(t)}$ satisfies CLT.

To use Corollary 5.7, let us prove an upper bound of the mean μ and a lower bound of the variance σ^2 of $\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$. For this setting, the following upper bound of μ is know.

Lemma 5.9 (Theorem 4.8 in [15])**.** If $\eta \leq 1$, then

$$\mathbb{E}[f(x^{(t)})] \leq \frac{\eta a^2}{2} + (1-\eta)^t f(x^{(0)}). \tag{5.28}$$

Since the second term decays exponentially, it implies that

$$\mathrm{PA}^f_{\mathcal{L}^{(t)}} = \mathbb{E}\left[\left\langle \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right] \approx \frac{\eta a^2}{2}. \tag{5.29}$$

For the lower bound of σ, we have the following lemma:

Lemma 5.10.

$$\sigma^2 \geq 2\eta^4 a^4. \tag{5.30}$$

Therefore, we know that

$$\mathrm{PA}^f_{\mathcal{L}^{(t)}} \leq \frac{\eta a^2}{2} - \beta_c \eta^4 a^4 \tag{5.31}$$

$$\mathrm{BA}^f_{\mathcal{L}^{(t)}} \leq \frac{\eta a^2}{2} - (1 + 2\beta_g)\beta_c \eta^4 a^4. \tag{5.32}$$

Since $\mathrm{PA}^f_{\mathcal{L}^{(t)}} = O(\eta)$, taking $\beta_g = \Omega(\eta^{-3}\beta_c^{-1})$ substantially improves the smooth minimum of $\mathrm{BA}^f_{\mathcal{L}^{(t)}}$.

We summarize implications of the example. The example illustrates that FF-thm is useful to tune the hyperparemters. Also, since we can use the previous result (ex. [124, 102, 56, 7, 117]) to bound μ, Proposition 5.6 and Corollary 5.7

[2] This assumption is not problematic because the probability of the exceptional event $\|\xi^{(t)}\| > C$ decays exponentially with respect to $C \in \mathbb{R}$. We can therefore neglect the exceptional event if we take sufficiently large C like $C = \Omega(\log T)$.

indeed help us to apply FF-thm to concrete examples. On the other hand, the calculation in Lemma 5.10 is heavy and might have a room to tighten the bound. One reason is that the calculation of the performance of the PA and the BA requires the smooth mean or the variance of $\left\langle \mathbb{B}_w{}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle$, whose state space is $\mathcal{X}^{T+1}$. The huge state space $\mathcal{X}^{(T+1)}$ makes the calculation difficult. Although FF-thm is useful to prove general results like Corollary 5.5, application of FF-thm to concrete examples is difficult. We resolve this problem in the next section.

5.4 Retrospective Process of Branching Algorithm and its Application to Stochastic Gradient Descent

We next introduce an easier technique to evaluate the performance of the BA than FF-thm. Recall that the difficulty in the application of FF-thm is the huge state space $\mathcal{X}^{T+1}$ of $\mathbb{P}_\mathrm{F}[\mathbb{X}^{(T)}]$ and $\mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}]$. A simpler representation of $\mathbb{P}_\mathrm{B}$ than the path-level probability is preferable. In this section, we represent $\mathbb{P}_\mathrm{B}$ as a Markov chain, which will turn out to be the retrospective process. In this section, we consider the case where $\beta_c \to 0+$ for simplicity. In other words, we evaluate the performance of the BA by $\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[\left\langle \mathbb{B}_w{}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle\right]$.

5.4.1 Retrospective Process of Branching Algorithm

Let us represent $\mathbb{P}_\mathrm{B}[\mathbb{X}^{(t)}]$ as a Markov chain with time-dependent transition matrix $\mathcal{L}^{(t')}_{T,\mathrm{B}}$:

$$\mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}] = \mathbb{P}^{(0)}_\mathrm{B}(x^{(0)}) \prod_{t'=0}^{T-1} \mathcal{L}^{(t')}_{T,\mathrm{B}}(x^{(t'+1)} \mid x^{(t')}), \tag{5.33}$$

where $\mathcal{L}^{(t)}_{T,\mathrm{B}}$ is defined as follows. Let us assume that $x^{(t)} = x'$ and consider the probability $\mathcal{L}^{(t)}_{T,\mathrm{B}}(x \mid x')$ that $x^{(t+1)} = x$ under the conditional probability $\mathbb{P}_\mathrm{B}[\mathbb{X}^{(t:T)} \mid x^{(t)} = x']$. By the Markov property of the BA, we know that $\mathcal{L}^{(t')}_{T,\mathrm{B}}$ can decompose $\mathbb{P}_\mathrm{B}$ as (5.33). We call $\mathcal{L}^{(t)}_{x,\mathrm{B}}$ the transition matrix of the *retrospective process.*

We explicitly calculate the retrospective process $\mathcal{L}^{(t)}_{T,\mathrm{B}}$. As a preparation, we define a *lineage fitness* $u^{(t+1:T)}(x^{(t+1)})$ by

$$u^{(t+1:T)}(x^{(t+1)}) := \sum_{x^{(t+2)},x^{(t+3)},\dots,x^{(T)}\in\mathcal{X}} \prod_{t'=t+1}^{T-1} e^{-\beta_g{}^{(t'+1)} f(x^{(t'+1)})} \mathcal{L}^{(t')}(x^{(t'+1)} \mid x^{(t')}). \tag{5.34}$$

We specially define $u^{(T:T)}(x) = 1$. By using the BA without caring capacity (5.8), we can interpret the lineage fitness as the expected number of the descendants of the solution $x^{(t+1)}$ at time T. We can calculate $\mathcal{L}^{(t)}_{T,\mathrm{B}}$ by using lineage fitness.

Theorem 5.11.

$$\mathcal{L}^{(t)}_{T,\mathrm{B}}(x \mid x') = \frac{u^{(t+1:T)}(x) e^{-\beta_g{}^{(t)} f(x)} \mathcal{L}^{(t)}(x \mid x')}{u^{(t:T)}(x')}. \tag{5.35}$$

We intuitively derive the theorem via the BA without caring capacity (5.8) (Figure 5.3). We first give an operational interpretation of $\mathbb{P}_\mathrm{B}[\mathbb{X}^{(t:T)} \mid x^{(t)} = x']$.

Let us consider the subpopulation that consists of the descendants of the solution x' at time t. Let us take a path $\mathbb{X}^{(t:T)} = \{x^{(t')}\}_{t'=t,t+1,\dots,T}$ of the BA without caring capacity in the subpopulation. Precisely, we choose an descendant at time T in the sub-population uniformly at random and then take the path $\mathbb{X}^{(t:T)}$ with respect to the chosen solution. Under this setting, $\mathbb{X}^{(t:T)}$ follows $\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(t:T)} \mid x^{(t)} = x']$. Therefore, $\mathcal{L}^{(t)}_{T,\mathrm{B}}(x \mid x')$ is the probability that $x^{(t+1)} = x$ on the path $\mathbb{X}^{(t:T)}$. We evaluate this probability. By (5.8), the expected number of the solution x at time $t+1$ in the sub-population is $e^{-\beta_g{}^{(t+1)}f(x)}\mathcal{L}^{(t)}(x \mid x')$. Each solution x at time $t+1$ yields $u^{(t+1:T)}(x)$ descendants at time T. Therefore, the number of paths that satisfies $x^{(t+1)} = x$ is $u^{(t+1:T)}(x)e^{-\beta_g{}^{(t+1)}f(x)}\mathcal{L}^{(t)}(x \mid x')$. Also, the total number of paths (total number of solutions at time T) is $u^{(t:T)}(x')$. Therefore, the probability is given by (5.35).

We give some remarks on the retrospective process. Since $\mathcal{L}^{(t)}_{T,\mathrm{B}}$ has the same information as $\mathbb{P}_{\mathrm{B}}$, the retrospective process is a simpler representation of $\mathbb{P}_{\mathrm{B}}$. Also, since $\mathcal{L}^{(t)}_{x,\mathrm{B}}$ defines the one-step time evolution of the solution, we can regard it as an iterative optimization algorithm biased from $\mathcal{L}^{(t)}$ due to populational evolution. By using the convention $\mathcal{L}^{(-1)}(x^{(0)} \mid x^{(-1)}) = \nu(x)$, we can simplify (5.33) as

$$\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)}] = \prod_{t'=-1}^{T-1} \mathcal{L}^{(t')}_{T,\mathrm{B}}(x^{(t'+1)} \mid x^{(t')}). \tag{5.36}$$

We note that $\mathcal{L}^{(-1)}_{T,\mathrm{B}}(x^{(0)} \mid x^{(-1)})$ does not depends on $x^{(-1)}$.

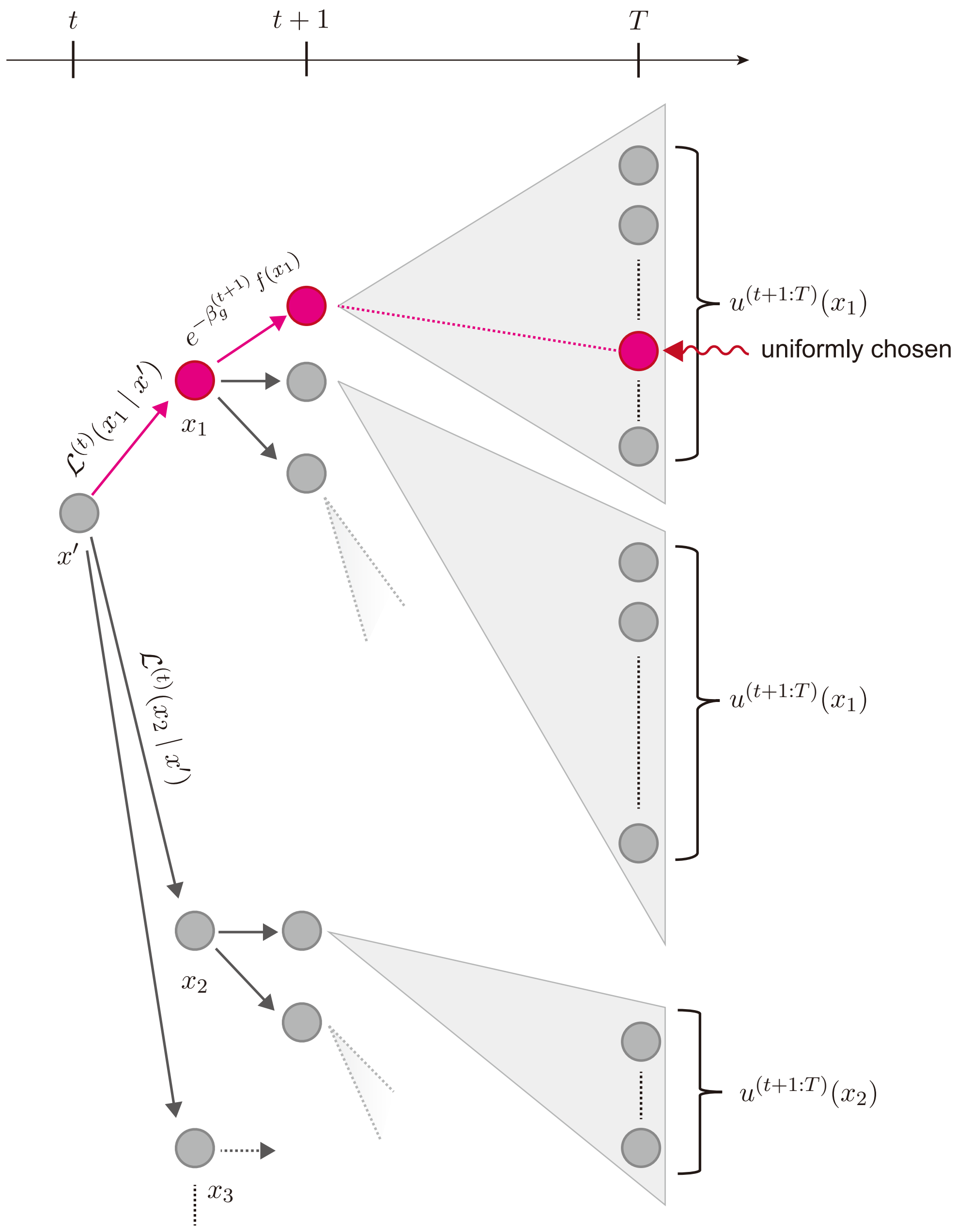


Figure 5.3: Schematic representation of the intuitive proof of Theorem 5.11. The retrospective process $\mathcal{L}_{T,\mathrm{B}}^{(t)}(x_1 \mid x')$ can be interpreted as the following probability: We consider the BA without caring capacity (5.8). We focus on the sub-population that consists of the descendants of the solution x' at time t. We choose a descendant at time T in the sub-population uniformly at random. In the figure, the red solution is chosen. Under this setting, $\mathcal{L}_{T,\mathrm{B}}^{(t)}(x_1 \mid x')$ is the probability that $x^{(t+1)} = x_1$ on the path followed backwardly from the chosen solution connected by red allows. We evaluate the probability. The expected number of the solution x_1 at time $t+1$ in the sub-population is $e^{-\beta_g{}^{(t+1)} f(x_1)} \mathcal{L}^{(t)}(x_1 \mid x')$. Also, the solution x_1 at time $t+1$ yields $u^{(t+1:T)}(x_1)$-descendants at time T. Therefore, the probability is proportional to $u^{(t+1:T)}(x_1) e^{-\beta_g{}^{(t+1)} f(x_1)} \mathcal{L}^{(t)}(x_1 \mid x')$, which proves Theorem 5.11.

5.4.2 One-Step Approximation of the Retrospective Process

Although the retrospective process is simpler than the path probability $\mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}]$, it is still difficult to calculate in concrete examples. The reason is that the lineage fitness $u^{(T:t)}(x)$ in (5.35) is difficult to calculate. The calculation of $u^{(t:T)}(x)$ is similar to the calculation of the normalization factor and thus difficult in general [14].

To avoid these problems, we use an *one-step approximation of retrospective process*. In this approximation, we substitute $u^{(t+1:t)}(x)$ with unity in the numerator of (5.35) and define a new transition matrix $\mathcal{L}^{(t)}_{+1,\mathrm{B}}$ by

$$\mathcal{L}^{(t)}_{+1,\mathrm{B}}(x \mid x') := \frac{e^{-\beta_g{}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x')}{u'^{(t)}_{+1}(x')}, \tag{5.37}$$

where u'_{+1} is an *one-step lineage fitness* defined by

$$u'^{(t)}_{+1}(x') := \sum_{x \in \mathcal{X}} e^{-\beta_g{}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x'). \tag{5.38}$$

In addition, we define the path probability of the one-step approximation by

$$\mathbb{P}_\mathrm{R}[\mathbb{X}^{(t)}] := \mathbb{P}^{(0)}_\mathrm{B}(x^{(0)}) \prod_{t'=0}^{T-1} \mathcal{L}^{(t')}_{+1,\mathrm{B}}(x^{(t'+1)} \mid x^{(t')}) \tag{5.39}$$

$$= \prod_{t'=-1}^{T-1} \mathcal{L}^{(t')}_{+1,\mathrm{B}}(x^{(t'+1)} \mid x^{(t')}). \tag{5.40}$$

The numerator of (5.39) incorporates the effect $e^{-\beta_g{}^{(t+1)} f(x)}$ of replication at present time. On the other hand, the numerator of (5.35) incorporates the effect $u^{(t+1:T)}(x)$ of replication from the next step to time T in addition to $e^{-\beta_g{}^{(t+1)} f(x)}$. We therefore call $\mathcal{L}^{(t)}_{+1,\mathrm{B}}$ the one-step approximation of the retrospective process. Since (5.39) does not contain the lineage fitness, the one-step approximation $\mathcal{L}^{(t)}_{+1,\mathrm{B}}$ is easier to calculate than $\mathcal{L}^{(t)}_{T,\mathrm{B}}$.

Although the one-step approximation is different from the original retrospective process, it is useful to prove an upper bound of $\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[\left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right]$. We aim to prove the following conjecture in the rest of the section:

Conjecture 5.12.

$$\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[\left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right] \leq \mathbb{E}_{\mathbb{P}_\mathrm{R}}\left[\left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right]. \tag{5.41}$$

If the conjecture holds, then an upper bound $\mathbb{E}_{\mathbb{P}_\mathrm{R}}\left[\left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right] \leq C$ for the one-step approximation with a constant C implies that the upper bound $\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[\left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle\right] \leq C$ for the BA with the same constant C.

We explain an intuition why we expect that the conjecture holds, although we need additional assumptions to rigorously prove the conjecture. The discussion will reveal the additional assumptions. To make the discussion concise, let $g[\mathbb{X}^{(T)}] := \left\langle \beta_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle$. Recall that the one-step approximation incorporates the replication $e^{-\beta_g{}^{(t+1)} f(x)}$ at present time while it does not the replication $u^{(t+1:T)}(x)$ until time T. Since FF-thm indicates that

$$\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[g[\mathbb{X}^{(T)}]\right] \leq \mathbb{E}_{\mathbb{P}_\mathrm{F}}\left[g[\mathbb{X}^{(T)}]\right], \tag{5.42}$$

and the one-step approximation partly incorporates the acceleration by populational evolution, we expect that

$$\mathbb{E}_{\mathbb{P}_{\mathrm{B}}}\left[g[\mathbb{X}^{(T)}]\right] \leq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[g[\mathbb{X}^{(T)}]\right] \leq \mathbb{E}_{\mathbb{P}_{\mathrm{F}}}\left[g[\mathbb{X}^{(T)}]\right]. \tag{5.43}$$

However, the inequality does not hold in general because the averaged future values $g[\mathbb{X}^{(t+1:T)}]$ of objective function might not be monotonically decreasing with respect to the replication $u^{(t+1:T)}(x^{(t+1)})$ until time T. Since $u^{(t+1:T)}(x^{(t+1)})$ depends on the hyperparameter ${\beta_g}^{(T)}$ while $g[\mathbb{X}^{(t+1:T)}]$ does on the different hyperparameter ${\beta_w}^{(T)}$, such a monotonicity does not hold in general. In such a situation, $\mathcal{L}^{(t)}_{T,\mathrm{B}}$ tends to update $x^{(t)}$ to a solution $x^{(t+1)}$ with higher $\mathbb{E}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)}]$ than other alternatives due to the factor $u^{(t+1:T)}(x^{(t+1)})$ in the numerator. In short, $u^{(t+1:T)}(x^{(t+1)})$ may increase $\mathbb{E}_{\mathbb{P}_{\mathrm{B}}}\left[g[\mathbb{X}^{(T)}]\right]$. Since $\mathbb{P}_{\mathrm{R}}$ does not have the factor $u^{(t+1:T)}(x^{(t+1)})$, the expectation $\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(T)}]]$ may become smaller than $\mathbb{E}_{\mathbb{P}_{\mathrm{B}}}[g[\mathbb{X}^{(T)}]]$. The discussion reveals that the monotonicity of $u^{(t+1:T)}(x^{(t+1)})$ with respect to $\mathbb{E}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)}]$ is necessary to prove Conjecture 5.12.

We precisely state the assumption of the monotonicity. For a technical reason, we consider the following form of the monotonicity.

Assumption 5.13 (Local Monotonicity)**.** For any $t = -1, 0, \dots, T-1$ and $x^{(t)} \in \mathcal{X}$, let $\mathcal{X}^{(t+1)}$ be the set of the solution $x^{(t+1)}$ at the next step sampled from $\mathcal{L}^{(t)}_{+1,\mathrm{B}}(x^{(t+1)} \mid x^{(t)})$. Then, for all $x, x' \in \mathcal{X}^{(t+1)}$ with $f(x) \geq f(x')$, we have the following:

$$u^{(t+1:T)}(x) \leq u^{(t+1:T)}(x'), \tag{5.44}$$

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)} = x] \geq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)} = x']. \tag{5.45}$$

The assumption is sufficient to prove Conjecture 5.12.

Theorem 5.14. Under Assumption 5.13, Conjecture 5.12 holds.

Before proceeding, we discuss that we can expect that the assumption holds for a broad class of iterative optimization algorithm $\mathcal{L}^{(t)}$. We can expect that the monotonicity holds if the update by $\mathcal{L}^{(t)}$ is local, continuous, and mixing: The locality means that $x^{(t+1)}$ and $x^{(t)}$ are sufficiently close under $\mathbb{P}_{\mathrm{F}}$; The continuity means that, if $x^{(t+1)}$ and ${x'}^{(t+1)}$ are sampled from $\mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)})$ and $\mathcal{L}^{(t)}({x'}^{(t+1)} \mid {x'}^{(t)})$ respectively and $x^{(t)}$ is close to ${x'}^{(t)}$, then the distribution of $x^{(t+1)} - x^{(t)}$ and that of ${x'}^{(t+1)} - {x'}^{(t)}$ is close; The mixing means that, if $x^{(t)}$ and ${x'}^{(t)}$ are close, then the distribution of $x^{(t+k)}$ updated k-times from $x^{(t)}$ and that of ${x'}^{(t+k)}$ updated from ${x'}^{(t)}$ are almost identical for sufficiently large k. Since these assumptions are mild, we can expect that they hold for a broad class of $\mathcal{L}^{(t)}$.

An intuitive explanation why these conditions imply the monotonicity is as follows (Figure 5.4). We first consider the monotonicity of $u^{(t+1:T)}(x^{(t+1)})$ with respect to $f(x^{(t+1)})$. Let us take two updated solutions $x^{(t+1)}$ and ${x'}^{(t+1)}$ sampled from $\mathcal{L}^{(t)}(\cdot \mid x^{(t)})$. To compare $u^{(t+1:T)}(x^{(t+1)})$ and $u^{(t+1:T)}({x'}^{(t+1)})$, we focus on the path $\mathbb{X}^{(t+1:T)}$ of the solutions after $x^{(t+1)}$ and the path ${\mathbb{X}'}^{(t+1:T)} := \{{x'}^{(t')}\}_{t'}$ after ${x'}^{(t+1)}$. By locality, the solutions $x^{(t+1)}$ and ${x'}^{(t+1)}$ are also close. It implies that the difference between the distribution of $x^{(t')}$ and ${x'}^{(t')}$ disappears in a few steps due to the mixing property. Therefore, the difference between $u^{(t+1:T)}(x)$ and $u^{(t+1:T)}(x')$ is determined in the next few steps. Precisely, we expect that

$$u^{(t+1:T)}(x^{(t+1)}) \propto u^{(t+1:t+k)}(x^{(t+1)}), \tag{5.46}$$

for some $t + k \ll T$. In the next few steps $i = 1, 2, \ldots, k$, we can expect that the distribution of $f(x^{(t+i)})$ is that of $f({x'}^{(t+i)})$ shifted by $f(x^{(t+1)}) - f({x'}^{(t+1)})$ since k is small and $\mathcal{L}^{(t+i)}$ is continuous (Figure 5.4). It implies the monotonicity of $u^{(t:T)}(x^{(t+1)})$ with respect to $f(x^{(t+1)})$ due to (5.46). By a similar argument, we can expect that $\mathbb{E}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)}]$ is monotonically non-decreasing with respect to $f(x^{(t+1)})$. Therefore, we expect that Assumption 5.13 holds for a broad class of learning rules.

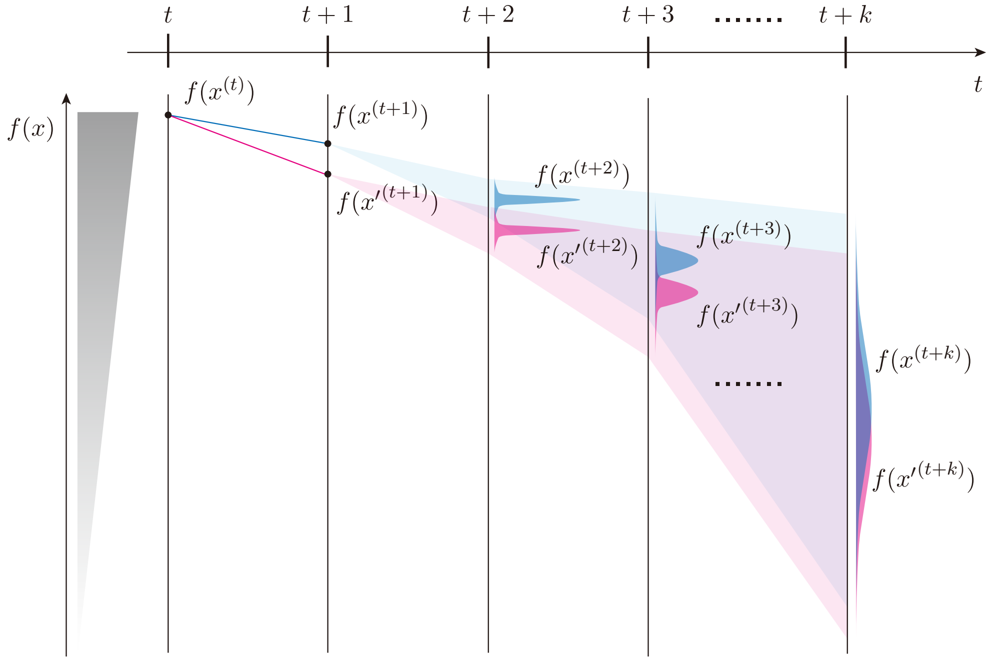


Figure 5.4: Schematic representation of the intuitive proof why the monotonicity holds. We first intuitively prove that $u^{(t+1:T)}(x)$ is monotonically decreasing with respect to $f(x)$ when the update by the iterative optimization algorithm $\mathcal{L}^{(t)}(\cdot \mid x')$ that is local, continuous, and mixing. We consider two updated solutions $x^{(t+1)}$ and $x'^{(t+1)}$ sampled from $\mathcal{L}^{(t)}(\cdot \mid x^{(t)})$. We suppose that $f(x^{(t+1)}) \geq f(x'^{(t+1)})$. We consider the trajectory $\{x^{(t+i)}\}_{i=2,3,\dots}$ updated from $x^{(t+1)}$ and that $\{x'^{(t+i)}\}_i$ updated from $x'^{(t+1)}$. In the figure, the vertical axis represent the value of $f(x)$ and the horizontal axis does the distribution of $f(x^{(t+i)})$ and $f(x'^{(t+i)})$. The blue bell-type distribution shows the distributions of $f(x^{(t+i)})$ and the red distribution does for $f(x'^{(t+i)})$. Since $x^{(t+i)}$ and $x'^{(t+i)}$ $(i = 2, 3, \dots)$ are stochastically determined, we represent the parent-daughter relationship as a blue or a red trapezoid between two successive times. When i is small $(i \ll k)$, we can expect that the distribution of $f(x^{(t+i)})$ is that of $f(x'^{(t+i)})$ shifted by $f(x^{(t+1)}) - f(x'^{(t+1)})$ due to the continuity of $\mathcal{L}^{(t+i)}$. When i is large $(i \geq k)$, we expect that the two distributions of $f(x^{(t+i)})$ and $f(x'^{(t+i)})$ are almost identical since $x^{(t+1)}$ and $x'^{(t+1)}$ are close and $\mathcal{L}^{(t)}$ is mixing. Therefore, we can expect that $u^{(t+1:T)}(x)$ is mainly determined by $f(x^{(t+i)})$ for $i = 1, 2, \dots, k$. These two arguments imply the monotonicity of the lineage fitness $u^{(t+1:T)}(x)$. A similar argument holds for the monotonicity of $\mathbb{E}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)}]$.

5.4.3 Retrospective Behavior of Stochastic Gradient Descent

We will see an application of the one-step approximation of the retrospective process (5.37) and Theorem 5.14 to SGD. Since the one-step approximation is easy to calculate, we can derive an explicit formula of $\mathcal{L}^{(t)}_{+1,\mathrm{B}}$ in this example by using a linear approximation of f.

To make the calculation easier, we assume that $x^{(t)}$ is interior of the constraint $\mathcal{X}$. In this setting, we can omit the projection operator in (3.62) and the update becomes

$$x^{(t+1)} = x^{(t)} - \eta^{(t+1)}(\nabla f(x^{(t)}) + \xi^{(t+1)}). \tag{5.47}$$

We also assume that the noise $\xi^{(t)}$ follows a normal distribution.

Assumption 5.15 (Normality of the noise)**.** For all t and $x^{(t)} \in \mathcal{X}$

$$\xi^{(t+1)} \sim_{\mathbb{P}_\mathrm{F}} \mathcal{N}\left(\mathbf{0}, \Sigma(x^{(t)}))\right), \tag{5.48}$$

where $\Sigma(x^{(t)})$ is some positive definite matrix.

These assumptions imply that $x^{(t+1)} \sim \mathcal{N}\left({\mu_\mathrm{F}}^{(t)}, (\eta^{(t+1)})^2\Sigma(x^{(t)})\right)$ where

$${\mu_\mathrm{F}}^{(t)} = x^{(t)} - \eta^{(t+1)}\nabla f(x^{(t)}). \tag{5.49}$$

In other words,

$$\begin{aligned}
&\mathcal{L}^{(t)}(x \mid x') && (5.50)\\
&= \frac{\exp\left(-\frac{1}{2}(\eta^{(t+1)})^{-2}(x - x' + \eta^{(t+1)}\nabla f(x'))^\top \Sigma^{-1}(x')(x - x' + \eta^{(t+1)}\nabla f(x'))\right)}{(2\pi)^{-d/2}(\eta^{(t+1)})^d|\Sigma(x')|^{1/2}}. && (5.51)
\end{aligned}$$

For notational simplicity, we define $\Delta = x - x' + \eta^{(t+1)}\nabla f(x')$ and rewrite the above equation as

$$\mathcal{L}^{(t)}(x \mid x') = \frac{\exp\left(-\frac{1}{2}(\eta^{(t+1)})^{-2}\Delta^\top\Sigma^{-1}(x')\Delta\right)}{(2\pi)^{-d/2}(\eta^{(t+1)})^d|\Sigma(x')|^{1/2}}. \tag{5.52}$$

The normality of $\xi^{(t)}$ and a linear approximation of f enable us to calculate $\mathcal{L}^{(t)}_{+1,\mathrm{B}}(x \mid x')$. If x is sampled from $\mathcal{L}^{(t)}(x \mid x')$, then the distance $\|x - x'\| = O(\eta^{(t+1)})$. When $\eta^{(t+1)}$ is small, we can approximate $f(x)$ as

$$\begin{aligned}
f(x) &= f(x') + \left\langle \nabla f(x'), x - x' \right\rangle + O((\eta^{(t+1)})^2) && (5.53)\\
&\approx f(x') + \left\langle \nabla f(x'), x - x' \right\rangle && (5.54)\\
&= f(x') + \left\langle \nabla f(x'), \Delta - \eta^{(t+1)}\nabla f(x') \right\rangle. && (5.55)
\end{aligned}$$

By this approximation and (5.52), we have [3]

$$(2\pi)^{-d/2}(\eta^{(t+1)})^d|\Sigma(x')|^{1/2}{u'}^{(t)}_{+1}(x')\mathcal{L}^{(t)}_{+1,\mathrm{B}}(x \mid x') \tag{5.56}$$

$$= e^{-{\beta_g}^{(t+1)}f(x)}\mathcal{L}^{(t)}(x \mid x') \tag{5.57}$$

$$\approx \exp\left(-{\beta_g}^{(t+1)}\left(f(x') + \left\langle \nabla f(x'), \Delta - \eta^{(t+1)}\nabla f(x')\right\rangle\right)\right) \tag{5.58}$$

$$\times \exp\left(-\frac{1}{2}(\eta^{(t+1)})^{-2}\Delta^\top\Sigma^{-1}(x')\Delta\right) \tag{5.59}$$

$$= \exp\left(-{\beta_g}^{(t+1)}\left(f(x') - \eta^{(t+1)}\|\nabla f(x')\|^2\right)\right) \tag{5.60}$$

$$\times \exp\left(-{\beta_g}^{(t+1)}\left\langle \nabla f(x'), \Delta\right\rangle - \frac{1}{2}(\eta^{(t+1)})^{-2}\Delta^\top\Sigma^{-1}(x')\Delta\right) \tag{5.61}$$

$$= \exp\left(-{\beta_g}^{(t+1)}\left(f(x') - \eta^{(t+1)}\|\nabla f(x')\|^2\right) + \frac{(\beta^{(t+1)}\eta^{(t+1)})^2}{2}\|\nabla f(x)\|^2_{\Sigma(x)}\right) \tag{5.62}$$

$$\times \exp\left(-\frac{1}{2}(\eta^{(t+1)})^{-2}\Delta_\mathrm{R}^\top\Sigma^{-1}(x')\Delta_\mathrm{R}\right), \tag{5.63}$$

where

$$\Delta_\mathrm{R} := \Delta + {\beta_g}^{(t+1)}(\eta^{(t+1)})^2\Sigma(x')\nabla f(x'). \tag{5.64}$$

By comparing (5.50) and (5.62), we know that $x^{(t+1)}$ approximately follows distribution $\mathcal{N}\left({\mu_\mathrm{R}}^{(t+1)}, (\eta^{(t+1)})^2\Sigma(x^{(t)})\right)$ under $\mathbb{P}_\mathrm{R}$, where

$${\mu_\mathrm{R}}^{(t+1)} = x^{(t)} - \eta^{(t+1)}\left(I + {\beta_g}^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\right)\nabla f(x^{(t)}). \tag{5.65}$$

We can also see that

$${u'}^{(t)}_{+1}(x^{(t)}) \approx \exp\left(-{\beta_g}^{(t+1)}\left(f(x^{(t)}) - \eta^{(t+1)}\|\nabla f(x^{(t)})\|^2\right)\right. \tag{5.66}$$

$$\left. + \frac{({\beta_g}^{(t+1)}\eta^{(t+1)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}\right). \tag{5.67}$$

We calculate $\mathbb{E}_{\mathbb{P}_\mathrm{R}}[f(x^{(t+1)}) \mid x^{(t)}]$ for the later purpose. By a direct calculation,

$$\mathbb{E}_{\mathbb{P}_\mathrm{R}}[f(x^{(t+1)}) \mid x^{(t)}] \tag{5.68}$$

$$= \mathbb{E}_{\mathcal{N}\left({\mu_\mathrm{R}}^{(t+1)}, (\eta^{(t+1)})^2\Sigma(x^{(t)})\right)}[f(x^{(t+1)})] \tag{5.69}$$

$$\approx \mathbb{E}_{\mathcal{N}\left({\mu_\mathrm{R}}^{(t+1)}, (\eta^{(t+1)})^2\Sigma(x^{(t)})\right)}\left[f(x^{(t)}) + \left\langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)}\right\rangle\right] \tag{5.70}$$

$$= f(x^{(t)}) - \eta^{(t+1)}\|\nabla f(x^{(t)})\|^2 - {\beta_g}^{(t+1)}(\eta^{(t+1)})^2\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}. \tag{5.71}$$

We can generalize the argument to the solution $x^{(t+k)}$ after k-steps when we can assume $k\eta^{(t+1)} = O(\eta^{(t+1)})$. When $\nabla f(x)$ and $\Sigma(x)$ is sufficiently smooth, we have

$$\nabla f(x^{(t+i+1)}) = \nabla f(x^{(t)}) + O(\eta^{(t+1)}), \tag{5.72}$$

$$\Sigma(x^{(t+i)}) = \Sigma(x^{(t)}) + O(\eta^{(t+1)}), \tag{5.73}$$

[3] Here, we left $O((\eta^{(t+1)})^2)$ terms to keep the equations exact for linear f.

for any $i \in [k]$. By using these linear approximations, we can calculate $\mathcal{L}_{+1,\mathrm{B}}^{(t+i)}$ and other quantities as before. For $\mathcal{L}_{+1,\mathrm{B}}^{(t+i)}$, we know that [4]

$$x^{(t+i+1)} \sim_{\mathbb{P}_\mathrm{R}} \mathcal{N}\left({\mu_\mathrm{R}}^{(t+i+1)}, (\eta^{(t+i+1)})^2\Sigma(x^{(t+i)})\right), \tag{5.74}$$

where

$$\begin{align}
&{\mu_\mathrm{R}}^{(t+i+1)} \tag{5.75}\\
&\approx x^{(t+i)} - \eta^{(t+i+1)}\left(I + {\beta_g}^{(t+i+1)}\eta^{(t+i+1)}\Sigma(x^{(t+i)})\right)\nabla f(x^{(t+i)}) \tag{5.76}\\
&= x^{(t+i)} - \eta^{(t+i+1)}\left(I + {\beta_g}^{(t+i+1)}\eta^{(t+i+1)}\Sigma(x^{(t)})\right)\nabla f(x^{(t)}) + O((\eta^{(t+1)})^2) \tag{5.77}\\
&\approx x^{(t+i)} - \eta^{(t+i+1)}\left(I + {\beta_g}^{(t+i+1)}\eta^{(t+i+1)}\Sigma(x^{(t)})\right)\nabla f(x^{(t)}). \tag{5.78}
\end{align}$$

We also have

$$\begin{align}
{u'}_{+1}^{(t+i)}(x^{(t+i)}) &\approx \exp\left(-{\beta_g}^{(t+i+1)}\left(f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t+i)})\|^2\right)\right. \tag{5.79}\\
&\qquad\left.+\frac{({\beta_g}^{(t+i+1)}\eta^{(t+i+1)})^2}{2}\|\nabla f(x^{(t+i)})\|^2_{\Sigma(x^{(t+i)})}\right) \tag{5.80}\\
&= \exp\left(-{\beta_g}^{(t+i+1)}\left(f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t)})\|^2\right)\right. \tag{5.81}\\
&\qquad\left.+\frac{({\beta_g}^{(t+i+1)}\eta^{(t+i+1)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} + O((\eta^{(t+1)})^2)\right) \tag{5.82}\\
&\approx \exp\left(-{\beta_g}^{(t+i+1)}\left(f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t)})\|^2\right)\right. \tag{5.83}\\
&\qquad\left.+\frac{({\beta_g}^{(t+i+1)}\eta^{(t+i+1)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}\right). \tag{5.84}
\end{align}$$

In addition,

$$\begin{align}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[f(x^{(t+i+1)}) \mid x^{(t+i)}] \tag{5.85}\\
&\approx f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t+i)})\|^2 - {\beta_g}^{(t+i+1)}(\eta^{(t+i+1)})^2\|\nabla f(x^{(t+i)})\|^2_{\Sigma(x^{(t+i)})} \tag{5.86}\\
&= f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t)})\|^2 \tag{5.87}\\
&\qquad {\beta_g}^{(t+i+1)}(\eta^{(t+i+1)})^2\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} + O((\eta^{(t+1)})^2) \tag{5.88}\\
&\approx f(x^{(t+i)}) - \eta^{(t+i+1)}\|\nabla f(x^{(t)})\|^2 - {\beta_g}^{(t+i+1)}(\eta^{(t+i+1)})^2\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}. \tag{5.89}
\end{align}$$

We state the above results in the form of an assumption to make further discussion clear.

Assumption 5.16 (Local Linearity of f and $\Sigma(x)$)**.** Let $k \in N$ such that we can still regard $k\eta^{(t+1)}$ as $O(\eta^{(t+1)})$. For any $i \in [k]$, Equations (5.74), (5.79), and (5.85) hold.

We will see the consequences of the local linearity of f. The first consequence is the measure transformation of the noise $\xi^{(t+1)}$.

[4] We note that these equations are exact if f is linear and $\Sigma(x)$ is constant.

Lemma 5.17. Under the Assumption 5.16,

$$\xi^{(t+1)} \sim_{\mathbb{P}_\mathrm{R}} \mathcal{N}\left({\beta_g}^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\nabla f(x^{(t)}), \Sigma(x^{(t)})\right). \tag{5.90}$$

When $\Sigma(x^{(t)}) = I$, then the above lemma implies that $\xi^{(t)}$ is biased toward the negative gradient under $\mathbb{P}_\mathrm{R}$. The bias accelerates SGD under $\mathbb{P}_\mathrm{R}$ compared to $\mathbb{P}_\mathrm{F}$. Indeed, we can see the acceleration from the third term of (5.85).

The other consequence is the local monotonicity (Assumption 5.13) and Theorem 5.14. To show the local monotonicity of SGD, we need an additional assumption about the mixing property. Since the solution $x^{(t)}$ is perturbed by $\xi^{(t+1)}$, a similar argument to (5.46) implies that we can expect

$$u^{(t+1:T)}(x^{(t+1)}) \propto u^{(t+1:t+k)}(x^{(t+1)}), \tag{5.91}$$

$$\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t+1:T)}] \mid x^{(t+1)}] \propto \mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t+1:t+k)}] \mid x^{(t+1)}], \tag{5.92}$$

for some $t + k \ll T$ and any $x^{(t+1)}$ sampled from $\mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)})$. We state this property in the form of the following assumption to make the further discussion clear.

Assumption 5.18 (Local Mixing)**.** There exists $k = O(1)$ such that (5.91) and (5.92) hold for any $x^{(t+1)}$ sampled from $\mathcal{L}^{(t)}(x^{(t+1)} \mid x^{(t)})$.

Lemma 5.19. Under Assumption 5.16 and Assumption 5.18, Assumption 5.13 holds.

In particular, Theorem 5.14 is applicable to SGD. One-step approximation $\mathcal{L}^{(t)}_{+1,\mathrm{B}}$ calculated in this section is therefore a useful tool to bound $\mathbb{E}_{\mathbb{P}_\mathrm{B}}\left[\left\langle {\beta_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle\right]$.

5.4.4 Theoretical Guarantee of Branching Algorithm with Stochastic Gradient Descent

In this section, we will see the application of the calculated distribution of the noise $\xi^{(t)}$ under $\mathbb{P}_\mathrm{R}$ (Lemma 5.17).

We first show that the BA accelerates SGD with Nesterov's acceleration by effectively reducing the variance of the noise $\xi^{(t)}$. Suppose the assumptions of Assumption 3.10. Also, we assume that $\Sigma(x) = \sigma^2 I/d$ for simplicity. We set the hyperparameters including $\eta^{(t)}$ of Nesterov's acceleration as in Theorem 3.10. Let us take

$$\beta^{(t)} = \frac{d(A-1)}{\sigma^2\eta^{(t)}}, \tag{5.93}$$

for some constant $A > 1$. Then, Lemma 5.17 implies that

$$g^{(t)} \sim_{\mathbb{P}_\mathrm{R}^f} \mathcal{N}\left(A\nabla f(x^{(t)}), \sigma^2 I/d\right). \tag{5.94}$$

We add the superscript of $\mathbb{P}_\mathrm{R}$ to clarify the dependency on the objective function The distribution is the same as that of $g^{(t)}$ under $\mathbb{P}_\mathrm{F}$ when the objective function is $f_A(x) = Af(x)$:

$$g^{(t)} \sim_{\mathbb{P}_\mathrm{F}^{f_A}} \mathcal{N}\left(A\nabla f(x^{(t)}), \sigma^2 I/d\right). \tag{5.95}$$

Intuitively, optimizing $f(x)$ by the BA with SGD is equivalent to optimizing $f_A(x)$ by conventional SGD. In particular, by applying Theorem 3.10 for $f_A(x)$, we can prove the following acceleration.

Theorem 5.20.

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}^{f}}[f(x^{(T)})] - f(x^*) \le C\left(\frac{\sigma^2}{A^2 T\alpha} + \frac{D^2(\alpha+\gamma)}{T^2}\right). \tag{5.96}$$

By taking A sufficiently large, we can neglect the leading $O(1/T)$ term and achieve $O(1/T^2)$-convergence. Although the assumption $\Sigma(x) = \sigma^2 I/d$ is restrictive, we can assume it by adding sufficiently large isotropic noise artificially on $g^{(t)}$.

We next prove that the BA with SGD has an $O(1/T)$-convergence rate beyond strongly convex functions. Before stating a rigorous theory, we explain the intuition why the BA can achieve an $O(1/T)$-convergence rate. For simplicity, we assume that $\Sigma(x) = I$ in this paragraph. In the case of conventional SGD, the key property to prove an $O(1/T)$-convergence rate is the following [124, 88, 56]:

$$\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle\right] \ge \alpha\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}[\|x^{(t)} - x^*\|^2], \tag{5.97}$$

which follows from the strong convexity since $\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}[g^{(t+1)} \mid x^{(t)}] = \nabla f(x^{(t)})$. The corresponding condition for the BA is

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle\right] \ge \alpha\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2]. \tag{5.98}$$

If this condition holds, we can prove $O(1/T)$-convergence under $\mathbb{P}_{\mathrm{R}}$, i.e., of the BA with SGD by following the previous proofs [124, 88, 56]. However, since we consider not strongly convex functions, we need to adopt a different approach to prove (5.98). By Lemma 5.17, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle\right] = \left(1 + {\beta_g}^{(t+1)}\eta^{(t+1)}\right)\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle \nabla f(x^{(t)}), x^{(t)} - x^*\right\rangle\right]. \tag{5.99}$$

Therefore, it suffices to prove

$$\left(1 + {\beta_g}^{(t+1)}\eta^{(t+1)}\right)\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}\left[\left\langle \nabla f(x^{(t)}), x^{(t)} - x^*\right\rangle\right] \ge \alpha\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2]. \tag{5.100}$$

This condition holds even for not strongly convex functions by taking sufficiently large ${\beta_g}^{(t+1)}$. Indeed, if we know that

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle \nabla f(x^{(t)}), x^{(t)} - x^*\right\rangle\right] \approx h(t)\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}[\|x^{(t)} - x^*\|^2], \tag{5.101}$$

for some positive function h, then taking

$$\beta^{(t+1)} = \frac{\alpha}{\eta^{(t+1)}h(t)}, \tag{5.102}$$

seems to be sufficient to prove (5.98).

To formalize the above idea, we introduce the following concepts about the objective functions. We first introduce a weaker concept of the α-strong convexity by generalizing (3.54). Since $\nabla f(x^*) = 0$ from the first-order stationary condition, (3.54) is simplified as

$$\langle \nabla f(x), x - x^*\rangle \ge \alpha\|x - x^*\|^2, \tag{5.103}$$

for all $x \in \mathcal{X}$. By using $\theta_1 \ge 0$, we have a weaker concept: For all $x \in \mathcal{X}$,

$$\langle \nabla f(x), x - x^*\rangle \ge \alpha\|x - x^*\|^{2(1+\theta_1)}. \tag{5.104}$$

To incorporate the noise of the SGD, we consider the following modification: For all $x \in \mathcal{X}$,

$$\left\langle \tilde{\Sigma}(x)\nabla f(x), x - x^* \right\rangle \geq \alpha \|x - x^*\|^{2(1+\theta_1)}, \tag{5.105}$$

where $\tilde{\Sigma}(x)$ is a positive semi-definite matrix. We call the above condition $(\alpha, \theta_1, \tilde{\Sigma})$-*strong convexity.* We next introduce a modified version of the γ-smoothness by generalizing (3.55). Fix $0 < \theta_2 \leq 1$. A function is said to be $(\gamma, \theta_2, \tilde{\Sigma})$-smooth if

$$\|\tilde{\Sigma}(x)\nabla f(x)\| \leq \gamma \|x - x^*\|^{\theta_2}, \tag{5.106}$$

for all $x \in \mathcal{X}$.

By using Lemma 5.17, we can prove that the convergence rate of $(\alpha, \theta_1, \Sigma)$-strongly convex and $(\gamma, \theta_2, \Sigma)$-smooth function f is $O(1/T)$. To prove the convergence rate, we introduce the following usual assumptions [124].

Assumption 5.21. The second moment $E_{\mathbb{P}_{\mathrm{F}}}[\|g^{(t)}\|^2] \leq G^2$ of $g^{(t)}$ is bounded for all t.

We also assume the following to avoid the divergence of $\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t)}\|^2]$ (See the proof of Lemma 5.23).

Assumption 5.22.

$$0 \leq \theta_2 - 2\theta_1 < 1. \tag{5.107}$$

Let us take $\eta^{(t)}$ and ${\beta_g}^{(t)}$. Let $\kappa > 0$ be an arbitrary constant. By following [124, 56], we use

$$\eta^{(t)} = \frac{\kappa}{t+1}. \tag{5.108}$$

We next take $\beta^{(t)}$ as follows:

$$\beta^{(t)} = \frac{(t+1)^{\theta_1}}{\alpha\kappa\eta^{(t)}(G_{\mathrm{B}}\kappa)^{2\theta_1}}, \tag{5.109}$$

where G_{B} is a constant defined later.

Since the usual proof of $O(1/T)$-convergence requires the bound for $\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}[\|g^{(t)}\|^2]$, we need to bound $\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t)}\|^2]$. To bound $\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t)}\|^2]$, we introduce a constant G_{B}. Let G_{B} be any constant satisfying the following inequalities.

$$G_{\mathrm{B}}^2 \geq 4G^2 + 4\frac{3^{\theta_2}\gamma^2}{\alpha^2\kappa^2}(G_{\mathrm{B}}\kappa)^{2\theta_2 - 4\theta_1}, \tag{5.110}$$

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(0)} - x^*\|^2] \leq 6G_{\mathrm{B}}^2\kappa^2. \tag{5.111}$$

Sufficiently large G_{B} satisfies the above inequality because the order of G_{B} in the left hand side is greater than that in the right hand side due to the fact $\theta_2 - 2\theta_1 < 1$. By this preparation, we have the following bound:

Lemma 5.23. For $t = 1, 2, \ldots,,$

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t)}\|^2] \leq G_{\mathrm{B}}^2. \tag{5.112}$$

Under this setting, we can prove the $O(1/T)$-convergence rate.

Theorem 5.24. For $t = 0, 1, \ldots,$

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\|x^{(t)} - x^*\|^2\right] \leq \frac{3G_{\mathrm{B}}^2\kappa^2}{t+2}, \tag{5.113}$$

for $t \geq 1$.

Theorem 5.25. Let

$$s^{(t)} := \frac{2t}{(T+1)(T+2)}, \tag{5.114}$$

$$\bar{x}^{(T)} := \sum_{t=0}^{T} s^{(t+1)} x^{(t)}. \tag{5.115}$$

Then,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[f(\bar{x}^{(T)}) - f(x^*)] \leq \frac{7G_{\mathrm{B}}^2\kappa}{T+1}. \tag{5.116}$$

5.4.5 Examples

We will see an example of a not strongly convex function that achieves $O(1/T)$-convergence by Theorem 5.25. Let us consider $f(x) = \|x - x^*\|_p^p$ for some $2 < p \leq 3$ and some constant vector $x^* \in \mathcal{X}$. We assume that the covariance of the noise $\xi^{(t)}$ is $\Sigma = I$. By the equivalence of the norm [49], we know that there exist constants $C_{2,p}$ and $C_{p-1,2}$ such that,

$$C_{2,p}\|x - x^*\|_2 \leq \|x - x^*\|_p, \tag{5.117}$$

$$C_{p-1,2}\|x - x^*\|_{p-1} \leq \|x - x^*\|_2, \tag{5.118}$$

for all $x \in \mathcal{X}$. Since

$$\langle x - x^*, \nabla f(x)\rangle = p\|x - x^*\|_p^p \geq pC_{2,p}^p\|x - x^*\|_2^{2\left(1+\frac{p-2}{2}\right)}, \tag{5.119}$$

f is $(pC_{2,p}^p, \frac{p-2}{2}, I)$-strongly convex. Also, we have

$$\|\nabla f(x)\| = p\|x - x^*\|_{p-1}^{p-1} \leq pC_{p-1,2}^{1-p}\|x - x^*\|_2^{p-1}. \tag{5.120}$$

Therefore, by taking sufficiently large R like the diameter of $\mathcal{X}$, we have

$$\|\nabla f(x)\| \leq R\|x - x^*\|, \tag{5.121}$$

and f is $(R, 1, I)$-smooth. Therefore, we can apply Theorem 5.25 and we have an $O(1/T)$-convergence.

5.5 Discussion

In this chapter, we proposed two methods to theoretically analyze the performance of the evolutionary algorithms. To focus on the effect of populational evolution separately from the recombination, we introduced the BA that omits the recombination step. The BA with $\mathcal{L}$ can incorporate the iterative optimization algorithm $\mathcal{L}$ as a generalization of the mutation step. We then extended FF-thm and proved that the BA always performs better than the parallel execution of $\mathcal{L}$. We next introduced the retrospective process and its one-step approximation as a convenient way to evaluate the performance of the BA. To demonstrate the power of the retrospective process, we explicitly calculated the one-step approximation for SGD and showed that the BA accelerated the SGD. In particular, we showed $O(1/T)$-convergence for not strongly-convex functions, which improves the existing $O(1/\sqrt{T})$-convergence. The analysis showed that both FF-thm and the retrospective process are useful and have potential for further analysis of the evolutionary algorithms.

Although the BA uses an iterative optimization algorithm $\mathcal{L}$ as mutation, our work has theoretical implications for the conventional evolutionary algorithm. Indeed, the BA coincides with the conventional evolutionary algorithm without recombination if $\mathcal{L}$ is a random update. In this setting, FF-thm implies that the conventional evolutionary algorithm always performs better than the parallel execution of conventional mutation. This acceleration can be regarded as the effect of populational evolution in conventional evolutionary algorithms. Since the acceleration (Corollary 5.7) is larger when the variance is larger, the evolutionary algorithm might perform better when the mutation is more exploratory. Also, the quantification of the acceleration by FF-thm might be useful to tune hyperparameters like ${\beta_g}^{(T)}$.

We have room for theoretical extensions and further analysis of the BA. The first direction is an application of our methods to other examples. Although we focused on SGD in the examples, both FF-thm and the retrospective process are applicable to all iterative optimization algorithm $\mathcal{L}$. Since we demonstrated how to use FF-thm and the retrospective process in examples, further analysis by a similar approach might reveal accelerations of the BA for other iterative optimization algorithms $\mathcal{L}$.

Another direction is to include the environmental state y (See Chapter 4). In Chapter 4, we considered the situation where the replication depends on a random variable y as $e^{k(x,y)}$, while the BA uses the deterministic factor $e^{-{\beta_g}^{(t)}f(x)}$ in the sampling step. The inclusion of the environmental state y might be useful when the estimation of $f(x)$ and the gradient $\nabla f(x)$ at each computational nodes correlate. For this generalization, the techniques developed for a multitype branching process in random environments in Chapter 4 might be useful. Another promising approach is the retrospective process extended to population dynamics in random environments [8, 20, 48, 13, 5, 108]. Other techniques used in population dynamics in random environments might also be useful. The examples are the product of random matrices [34, 108] and linear random dynamical systems [42, 52, 100].

Other direction is the non-synchronization of the replication of the solutions. In BA, the replication of the solutions at the same generation synchronizes. However, this synchronization requires communication cost among computational nodes. By employing continuous-time age-structured model (3.27), we may extend our theory to the case where the replication does not synchronize.

The last direction is the consideration of the finiteness of the population. We assumed that the size of the population is infinite in this thesis. In practice, the size of the population is finite and the problem called degeneration may occur: The daughters of a solution $x^{(t)}$ with small $f(x^{(t)})$ may dominate for most of the population at the next step and the variety of solutions is lost. The variety of the solution is important for the acceleration by populational evolution as FF-thm indicates. To avoid degeneration, we need to use a moderate ${\beta_g}^{(t)}$ with respect to the size of the population. Since the number of offspring is $e^{-{\beta_g}^{(t)}f(x)}$, we can roughly estimate that ${\beta_g}^{(t)} = O(\log N_{\mathrm{size}})$ is necessary to completely avoid the degeneration. However, in practice, we need not to avoid the degeneration completely. To choose the hyperparameter ${\beta_g}^{(t)}$ that maximizes the performance of the BA, we should quantify how much the degeneration impairs the performance. For such analysis, promising mathematical tools are stochastic population dynamics with finite populations like branching processes [40] and coalescent processes [112].

By using the rough estimation ${\beta_g}^{(t)} = O(\log N_{\text{size}})$, we can compare the performance of the BA with other parallel algorithms than the PA. For example, let us compare the noise reduction in the Nesterov's acceleration by a large mini-batch and by the BA. By using N_{size}-computational nodes, we can use a N_{size}-times larger mini-batch to estimate the gradient than the BA. In this situation, the variance of the noise $\xi^{(t)}$ is reduced by $1/N_{\text{size}}$. On the other hand, the rough estimation implies that the BA can reduce the variance by $1/(\log N_{\text{size}})^2$. Although the BA is less efficient, it might be superior to the large mini-batch when the communication cost between computational nodes to process the large mini-batch is large. Another example is the comparison of the PA with averaging over paths on the BA when $\mathcal{L}^{(t)}$ is SGD (Section 2.1.2). The PA with averaging over paths has an $O(1/TN_{\text{size}})$-convergence rate when $N_{\text{size}} < \sqrt{T}$ for strongly convex functions. On the other hand, the BA with Nesterov's acceleration has an $O(1/T(\log N_{\text{size}})^2)$-convergence rate by the rough estimation. Therefore, the PA is better than the BA when $N_{\text{size}} \leq \sqrt{T}$, while the converse holds when N_{size} is large. In this sense, the BA might become more important when we invent a device that enables huge N_{size}. Although the above discussion does not show that the BA is superior to other parallel algorithms, it demonstrates that our approach is powerful for theoretically analyzing the evolutionary algorithms.

5.6 proofs

5.6.1 Proofs in Section 5.2

Derivation of Equation (5.5) and Equation (5.6). We use the same notation as Algorithm 5.2. Equation (5.6) trivially follows from the sampling of initial solutions of the BA.

We next prove (5.5). Let us consider a population P that consists of the solutions $\{{x'}_i^{(t+1)}\}_{i\in N_{\text{size}}}$ updated by $\mathcal{L}$. The empirical distribution of the solution x in P is

$$j'^{(t+1)}(x) = 1/N_{\text{size}} \cdot \sum_{i\in[N_{\text{size}}]} \delta_{x,{x'}_i^{(t+i)}}. \tag{5.122}$$

By the learning step of the BA, we know that $\mathbb{E}[j'^{(t+1)}(x)] = \mathcal{L}^{(t)}(x \mid x')\mathbb{P}_{\text{B}}^{(t)}(x')$. By the sampling step of the BA, we have

$$\mathbb{E}[j^{(t+1)}(x) \mid j'^{(t+1)}] = \frac{e^{-{\beta_g}^{(t+1)}f(x)}j'^{(t+1)}(x)}{\sum_{x''\in\mathcal{X}} e^{-{\beta_g}^{(t+1)}f(x'')}j'^{(t+1)}(x'')}. \tag{5.123}$$

Since the learning and the sampling steps are independent and $j'^{(t+1)}(x)$ converges to its expectation as $N_{\text{size}} \to \infty$, we have

$$\mathbb{P}_{\text{B}}^{(t+1)}(x) = \mathbb{E}[j^{(t+1)}(x)] = \frac{e^{-{\beta_g}^{(t+1)}f(x)}\mathcal{L}^{(t)}(x \mid x')\mathbb{P}_{\text{B}}^{(t)}(x')}{\sum_{x'\in\mathcal{X}} e^{-{\beta_g}^{(t+1)}f(x'')}\mathcal{L}^{(t)}(x'' \mid x')\mathbb{P}_{\text{B}}^{(t)}(x')}. \tag{5.124}$$

□

Proof of Proposition 5.1. We prove the proposition by induction on T. The base case $T = 0$ holds by definition (5.6). Let us consider the step case $T > 0$.

We fix a path $\mathbb{X}^{(T)} = \{x^{(0)}, x^{(1)}, \dots, x^{(T)}\}$ and let $\mathbb{X}^{(T-1)}$ be its truncation $\{x^{(0)}, x^{(1)}, \dots, x^{(T-1)}\}$. By the Markov property of the BA and (5.5), we have

$$\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)}] = \frac{e^{-\beta_g{}^{(T)} f(x^{(T)})} \mathcal{L}^{(T-1)}(x^{(T)} \mid x^{(T-1)}) \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T-1)}]}{\sum_{x \in \mathcal{X}, \mathbb{X}'^{(T-1)} \in \mathcal{X}^T} e^{-\beta_g{}^{(T)} f(x)} \mathcal{L}^{(T-1)}(x \mid x'^{(T-1)}) \mathbb{P}_{\mathrm{B}}[\mathbb{X}'^{(T-1)}]}, \tag{5.125}$$

where $\mathbb{X}'^{(T-1)} = \{x'^{(t)}\}_t$. By applying the induction hypothesis on $\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T-1)}]$ and rewriting the equation, we have (5.7) for T. □

5.6.2 Proofs in Section 5.3

Proof of Proposition 5.2. We first prove (5.11). By the Taylor's theorem,

$$1 + x \le e^x \le 1 + x + \frac{x^2 e^x}{2}. \tag{5.126}$$

Therefore,

$$\begin{align}
& -\frac{1}{\beta_c} \log \left[\frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \left(1 - \beta_c z_i + \frac{\beta_c^2 z_i^2 e^{-\beta_c z_i}}{2} \right) \right] \tag{5.127} \\
& \le \mathrm{smin}^{\beta_c}(z_0, z_1, \dots, z_{N_{\mathrm{size}}-1}) \tag{5.128} \\
& \le -\frac{1}{\beta_c} \log \left[\frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} (1 - \beta_c z_i) \right]. \tag{5.129} \\
& \tag{5.130}
\end{align}$$

Equivalently, we have

$$\begin{align}
& -\frac{1}{\beta_c} \log \left[1 - \frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \left(\beta_c z_i - \frac{\beta_c^2 z_i^2 e^{-\beta_c z_i}}{2} \right) \right] \tag{5.131} \\
& \le \mathrm{smin}^{\beta_c}(z_0, z_1, \dots, z_{N_{\mathrm{size}}-1}) \tag{5.132} \\
& \le -\frac{1}{\beta_c} \log \left[1 - \frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \beta_c z_i \right]. \tag{5.133} \\
& \tag{5.134}
\end{align}$$

By using the Taylor's theorem again, we have

$$x \le -\log(1 - x) \le x + \frac{x^2}{2(1 - x)^2}, \tag{5.135}$$

for $0 \le x \le 1$. Since we consider the limit $\beta_c \to 0+$, we can assume that $0 \le \frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \left(\beta_c z_i - \frac{\beta_c^2 z_i^2 e^{-\beta_c z_i}}{2} \right) \le 1$ and $0 \le \frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \beta_c z_i \le 1$. Therefore, we have

$$\begin{align}
& \frac{1}{\beta_c} \left(\frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \left(\beta_c z_i - \frac{\beta_c^2 z_i^2 e^{-\beta_c z_i}}{2} \right) \right) \tag{5.136} \\
& \le \mathrm{smin}^{\beta_c}(z_0, z_1, \dots, z_{N_{\mathrm{size}}-1}) \tag{5.137} \\
& \le \frac{1}{\beta_c} \left(\frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \beta_c z_i + \frac{\left(\frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \beta_c z_i \right)^2}{2(1 - \frac{1}{N_{\mathrm{size}}} \sum_{i \in [N_{\mathrm{size}}]} \beta_c z_i)^2} \right). \tag{5.138}
\end{align}$$

By rewriting it, we have

$$\frac{1}{N_{\text{size}}}\sum_{i\in[N_{\text{size}}]} z_i + O(\beta_c) \tag{5.139}$$

$$\leq \operatorname{smin}^{\beta_c}(z_0, z_1, \dots, z_{N_{\text{size}}-1}) \tag{5.140}$$

$$\leq \frac{1}{N_{\text{size}}}\sum_{i\in[N_{\text{size}}]} z_i + O(\beta_c). \tag{5.141}$$

Since both hand sides converges to $\frac{1}{N_{\text{size}}}\sum_{i\in[N_{\text{size}}]} z_i$ as $\beta_c \to 0+$, we have (5.11) by the squeeze theorem.

We next prove (5.12). Since

$$e^{-\beta_c \min_{i\in[N_{\text{size}}]} z_i} \leq \sum_{i\in[N_{\text{size}}]} e^{-\beta_c z_i} \leq N_{\text{size}} e^{-\beta_c \min_{i\in[N_{\text{size}}]} z_i}, \tag{5.142}$$

we have

$$\min_{i\in[N_{\text{size}}]} z_i - \frac{1}{\beta_c}\log N_{\text{size}} \leq \operatorname{smin}^{\beta_c}(z_0, z_1, \dots, z_{N_{\text{size}}-1}) \leq \min_{i\in[N_{\text{size}}]} z_i. \tag{5.143}$$

Since the left hand side converges to the right hand side as $\beta_c \to \infty$, we have (5.12) by the squeeze theorem. □

Proof of Proposition 5.3. We can prove (5.14) by a similar argument to Proposition 5.2. Indeed, by replacing $\frac{1}{N_{\text{size}}}\sum_{i\in[N_{\text{size}}]}$ with $\mathbb{E}_p$, we can prove (5.14).

We next prove (5.15). Let $m = \inf\{z \mid F_Z(z) > 0\}$. We can easily prove that

$$\lim_{\beta_c\to\infty}\left[-\frac{1}{\beta_c}\log E_p\left[e^{-\beta_c Z}\right]\right] \geq m, \tag{5.144}$$

since $\mathbb{E}[e^{-\beta_c Z}] \leq e^{-\beta_c m}$. We show the converse in the rest of the proof. Let us take sufficiently small $\epsilon > 0$. By the definition of $\inf Z$, there exist $\delta(\epsilon) > 0$ such that $F_Z(\delta(\epsilon)) = \epsilon$ and $\delta(\epsilon) \to 0$ as $\epsilon \to 0+$. Since $\mathbb{P}[Z \leq m + \delta(\epsilon)] = \epsilon$, we have

$$\mathbb{E}_p\left[e^{-\beta_c Z}\right] = \mathbb{P}[Z \leq m+\delta(\epsilon)]\mathbb{E}_p\left[e^{-\beta_c Z} \mid Z \leq m+\delta(\epsilon)\right] \tag{5.145}$$

$$+ \mathbb{P}[Z \geq m+\delta(\epsilon)]\mathbb{E}_p\left[e^{-\beta_c Z} \mid Z \geq m+\delta(\epsilon)\right] \tag{5.146}$$

$$\geq \mathbb{P}[Z \leq m+\delta(\epsilon)]\mathbb{E}_p\left[e^{-\beta_c Z} \mid Z \leq m+\delta(\epsilon)\right] \tag{5.147}$$

$$\geq \mathbb{P}[Z \leq m+\delta(\epsilon)]e^{-\beta_c(m+2\delta(\epsilon))} \tag{5.148}$$

$$= \epsilon e^{-\beta_c(m+2\delta(\epsilon))}. \tag{5.149}$$

By taking the logarithm and dividing by $-\beta_c$, we have

$$-\frac{1}{\beta_c}\log\mathbb{E}_p\left[e^{-\beta_c Z}\right] \leq -\frac{1}{\beta_c}\left(-\beta_c(m+\delta(\epsilon)) + \log(\epsilon)\right) \tag{5.150}$$

$$\leq m + \delta(\epsilon) - \frac{1}{\beta_c}\log\epsilon. \tag{5.151}$$

Therefore, we have

$$\lim_{\beta_c\to\infty}\left[-\frac{1}{\beta_c}\log\mathbb{E}_p\left[e^{-\beta_c Z}\right]\right] \leq m + \delta(\epsilon). \tag{5.152}$$

Since the above inequality holds for all $\delta(\epsilon) > 0$ by taking sufficiently small ϵ, we have

$$\lim_{\beta_c \to \infty} \left[-\frac{1}{\beta_c} \log \mathbb{E}_p \left[e^{-\beta_c Z} \right] \right] \leq m. \tag{5.153}$$

It completes the proof. □

Proof of Proposition 5.6. Since $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$, the random variable $e^{-\beta_c Z}$ follows a log-normal distribution [14] with mean parameter $-\beta_c \mu$ and variance parameter $(\beta_c \sigma)^2$. By the formula of the mean of log-normal distributions, $E[e^{-\beta_c Z}] = e^{-\beta_c \mu + (\beta_c \sigma)^2/2}$. Therefore,

$$\operatorname{smin}^{\beta_c}[Z] = \mu - \frac{\beta_c \sigma^2}{2}. \tag{5.154}$$

□

Proof of Theorem 5.4. By definition,

$$\mathrm{BA}^f_{\mathcal{L}^{(t)}} = \operatorname{smin}^{\beta_c}_{\mathbb{P}_\mathrm{B}} \left[\left\langle \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle \right] = -\frac{1}{\beta_c} \log \mathbb{E}_{\mathbb{P}_\mathrm{B}} \left[\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle \right] \tag{5.155}$$

We evaluate the expectation inside the logarithm. By using (5.7), we have

$$\mathbb{E}_{\mathbb{P}_\mathrm{B}} \left[\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle \right] \tag{5.156}$$

$$= \sum_{\mathbb{X}^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle} \mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}] \tag{5.157}$$

$$= \sum_{\mathbb{X}^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle} \frac{e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \mathbb{P}_\mathrm{F}[\mathbb{X}^{(T)}]}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}'^{(T)}] \right\rangle} \mathbb{P}_\mathrm{F}[\mathbb{X}'^{(T)}]} \tag{5.158}$$

$$= \frac{\sum_{\mathbb{X}^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)} + \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \mathbb{P}_\mathrm{F}[\mathbb{X}^{(T)}]}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}'^{(T)}] \right\rangle} \mathbb{P}_\mathrm{F}[\mathbb{X}'^{(T)}]} \tag{5.159}$$

$$= \frac{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)} + \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right]}{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right]}. \tag{5.160}$$

By using this equality, we have

$$\beta_c \left(\mathrm{PA}^f_{\mathcal{L}^{(t)}} - \mathrm{BA}^f_{\mathcal{L}^{(t)}} \right) \tag{5.161}$$

$$= -\log \mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle} \right] + \log \frac{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)} + \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right]}{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right]} \tag{5.162}$$

$$= \log \frac{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)} + \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right]}{\mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle} \right] \mathbb{E}_{\mathbb{P}_\mathrm{F}} \left[e^{-\left\langle \beta_c \mathbb{\beta}^{(T)}_w, f[\mathbb{X}^{(T)}] \right\rangle} \right]} \tag{5.163}$$

$$= \text{log-}\mathrm{Cov}_{\mathbb{P}_\mathrm{F}} \left[-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle, -\left\langle \beta_c \mathbb{\beta}_w{}^{(T)}, f[\mathbb{X}^{(T)}] \right\rangle \right]. \tag{5.164}$$

This equality implies the statement of the theorem. □

Proof of Corollary 5.5. The equality directly follows from Theorem 5.4. The inequality follows from the property of the log-covariance (3.7). Since both $e^{-\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}$ and $e^{-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}$ is monotonically non-decreasing with respect to $e^{-\left\langle {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}$, their covariance is non-negative [95]:

$$\mathrm{Cov}_{\mathbb{P}_{\mathrm{F}}}\left[e^{-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}, {}^{-\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}\right] \geq 0. \tag{5.165}$$

This inequality and (3.7) implies that the inequality in the statement of this corollary. □

Proof of Corollary 5.7. We use a similar argument as Proposition 5.6. Since $e^{-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}$ follows a log-normal distribution with mean parameter $\beta_g \mu$ and variance parameter $(\beta_g \sigma)^2$, we have

$$\mathbb{E}\left[e^{-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}\right\rangle]}\right] = \exp\left(-\beta_g \mu + \frac{\beta_g^2 \sigma^2}{2}\right). \tag{5.166}$$

By a similar argument, we have

$$\mathbb{E}\left[e^{-\left\langle \beta_c {\mathbb{\beta}_g}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}\right] = \exp\left(-\beta_c \mu + \frac{\beta_c^2 \sigma^2}{2}\right), \tag{5.167}$$

and

$$E\left[e^{-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle} e^{-\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle}\right] = \exp\left(-(\beta_g + \beta_c)\mu + \frac{(\beta_g + \beta_c)^2 \sigma^2}{2}\right). \tag{5.168}$$

Therefore, by the definition of the log-covariance, we have

$$\text{log-}\mathrm{Cov}_{\mathbb{P}_{\mathrm{F}}}\left[-\left\langle \beta_g {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle, -\left\langle \beta_c {\mathbb{\beta}_w}^{(T)}, f[\mathbb{X}^{(T)}]\right\rangle\right] = \beta_g \beta_c \sigma^2. \tag{5.169}$$

□

Proof of Lemma 5.8. We explicitly calculate $x^{(t)}$. By definition, we know that

$$x^{(1)} = x^{(0)} - \eta^{(1)}\left(\nabla f(x^{(0)}) + \xi^{(1)}\right) = (1-\eta)x^{(0)} - \eta \xi^{(1)}. \tag{5.170}$$

For $x^{(2)}$, we have

$$x^{(2)} = (1-\eta^{(2)})x^{(1)} + \eta^{(2)}\xi^{(2)} \tag{5.171}$$

$$= (1-\eta)^2 x^{(0)} - \eta(1-\eta)\xi^{(1)} - \eta\xi^{(2)}. \tag{5.172}$$

By continuing this, we have

$$x^{(t)} = (1-\eta)^t x^{(0)} - \eta \sum_{t'=1}^{t} (1-\eta)^{t-t'} \xi^{(t')}. \tag{5.173}$$

From this equation, we have

$$2f(x^{(t)}) = (1-\eta)^{2t} f(x^{(0)}) + (1-\eta)^t \eta \sum_{s=1}^{t} (1-\eta)^{t-s} \left\langle x^{(0)}, \xi^{(u)} \right\rangle \tag{5.174}$$

$$+ \eta^2 \sum_{s=1}^{t} \sum_{u=1}^{t} (1-\eta)^{2t-s-u} \left\langle \xi^{(s)}, \xi^{(u)} \right\rangle. \tag{5.175}$$

In particular,

$$2\mathbb{E}[f(x^{(t)})] = (1-\eta)^{2t} f(x^{(0)}) + a^2\eta^2 \sum_{t'=1}^{t} (1-\eta)^{2(t-t')}, \tag{5.176}$$

since the expectation of the first order term of $\xi^{(u)}$ (i.e. $\left\langle \xi^{(u)}, x^{(0)} \right\rangle$ and $\left\langle \xi^{(s)}, \xi^{(u)} \right\rangle$ for $s \neq u$) is zero. By a similar calculation, we have

$$2\mathbb{E}[f(x^{(t)}) \mid \mathcal{F}^{(t-k)}] = (1-\eta)^{2t} f(x^{(0)}) + a^2\eta^2 \sum_{t'=t-k+1}^{t} (1-\eta)^{2(t-t')} \tag{5.177}$$

$$+ \eta^2 \sum_{t'=1}^{t-k} (1-\eta)^{2(t-t')} \|\xi^{(t')}\|^2. \tag{5.178}$$

Since we assumed that $\|\xi^{(t)}\|^2$ is bounded, we have the statement of the lemma from these two equations. □

Proof of Lemma 5.10. Let us calculate Cov $\left[f(x^{(t)}), f(x^{(t')})\right]$. We note that

$$\mathrm{Cov}\left[f(x^{(t)}), f(x^{(t')})\right] = \mathbb{E}[Z^{(t)} Z^{(t')}]. \tag{5.179}$$

By (5.173) and (5.176),

$$Z^{(t)} = \eta^2 \sum_{s=1}^{t} (1-\eta)^{2(t-s)} \left(\|\xi^{(s)}\|^2 - \sigma^2 \right) + \eta^2 \sum_{s \neq u} (1-\eta)^{2t-s-u} \left\langle \xi^{(s)}, \xi^{(u)} \right\rangle. \tag{5.180}$$

Since $\mathbb{E}[\xi^{(t)}] = 0$, only even terms in $Z^{(t)} Z^{(t')}$ are non-zeros after taking the expectation. Here, even terms are the terms of the following form:

$$\prod_{i=1}^{k} \|\xi^{(j_i)}\|^{2n_i}, \tag{5.181}$$

for some indices j_i and $n_i \in \mathbb{N}$. □

Therefore, we know that $\mathbb{E}[Z^{(t)} Z^{(t')}]$ consists of the expectation of the following terms up to positive coefficient; (1) For $s = 1, 2, \ldots, t$ and $u = 1, 2, \ldots, t'$

$$\left(\|\xi^{(s)}\|^2 - a^2 \right) \left(\|\xi^{(u)}\|^2 - a^2 \right), \tag{5.182}$$

and (2) for $s = 1, 2, \ldots, t$ and $u = 1, 2, \ldots, t'$,

$$\left\langle \xi^{(s)}, \xi^{(t)} \right\rangle^2. \tag{5.183}$$

All of these terms have non-negative expected values. Therefore, we know that

$$\mathrm{Cov}\left[f(x^{(t)}), f(x^{(t')})\right] \geq 0. \tag{5.184}$$

In particular, we have a bound

$$\sigma^2 \geq \frac{1}{T+1} \sum_{s=1}^{t} \mathbb{V}[f(x^{(t)})]. \tag{5.185}$$

Let us bound $\mathbb{V}[f(x^{(t)})] = \mathbb{E}\left[\left(Z^{(t)}\right)^2\right]$. As a special case of the discussion in the previous paragraph, $\mathbb{E}\left[\left(Z^{(t)}\right)^2\right]$ consists of the terms with non-negative expected values. Therefore, we can construct a lower bound by extracting a subset of the terms. In particular, we have

$$\mathbb{E}\left[\left(Z^{(t)}\right)^2\right] \geq \eta^4 \mathbb{E}\left[\left(\|\xi^{(t)}\|^2 - \sigma^2\right)^2\right]. \tag{5.186}$$

By the property of normal distribution [83],

$$\mathbb{E}\left[\left(Z^{(t)}\right)^2\right] \geq 2\eta^4\sigma^2. \tag{5.187}$$

By (5.185) and (5.187), we have the statement of the lemma.

5.6.3 Proofs in Section 5.4

Proof of Theorem 5.11. Let $\mathbb{X}^{(t)} = \{x^{(0)}, x^{(1)}, \dots, x^{(T)}\}$ be the random variable on $\mathcal{X}^{T+1}$ sampled from $\mathbb{P}_{\mathrm{B}}$. We use the following notation in this proof for simplicity:

$$\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x'} := \{x^{(0)}, x^{(1)}, \dots, x^{(t-1)}, x', x^{(t+1)}, x^{(t+2)}, \dots, x^{(T)}\}, \tag{5.188}$$

$$\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x', x^{(t+1)}=x} := \{x^{(0)}, x^{(1)}, \dots, x^{(t-1)}, x', x, x^{(t+2)}, x^{(t+3)}, \dots, x^{(T)}\}. \tag{5.189}$$

For a function $g\colon \mathcal{X}^{T+1} \to \mathbb{R}$, we in addition define

$$\sum_{\mathbb{X}^{(T)} \setminus x^{(t)}} g[\mathbb{X}^{(t)} \upharpoonright_{x^{(t)}=x'}] := \sum_{x^{(0)}, x^{(1)}, \dots, x^{(t-1)}, x^{(t+1)}, x^{(t+2)}, \dots, x^{(T)} \in \mathcal{X}} g[\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x'}], \tag{5.190}$$

$$\sum_{\mathbb{X}^{(T)} \setminus x^{(t)}, x^{(t+1)}} g[\mathbb{X}^{(t)} \upharpoonright_{x^{(t)}=x', x^{(t+1)}=x}] \tag{5.191}$$

$$:= \sum_{x^{(0)}, x^{(1)}, \dots, x^{(t-1)}, x^{(t+2)}, x^{(t+3)}, \dots, x^{(T)} \in \mathcal{X}} g[\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x', x^{(t+1)}=x}]. \tag{5.192}$$

By the pathwise formulation (5.7), the probability that $x^{(t)} = x$ is

$$\sum_{\mathbb{X}^{(T)} \setminus x^{(t)}} \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x'}] \tag{5.193}$$

$$= \sum_{\mathbb{X}^{(T)} \setminus x^{(t)}} \frac{e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x'}] \right\rangle} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)} \upharpoonright_{x^{(t)}=x'}]}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}'^{(T)}] \right\rangle} \mathbb{P}_{\mathrm{F}}[\mathbb{X}'^{(T)}]}. \tag{5.194}$$

By using the convention $\mathcal{L}^{(-1)}(x^{(0)} \mid x^{(-1)}) = \nu(x^{(0)})$, the numerator is factorized as

$$\sum_{\mathbb{X}^{(T)} \backslash x^{(t)}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x'}] \right\rangle} \mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x'}] \tag{5.195}$$

$$= \sum_{\mathbb{X}^{(T)} \backslash x^{(t)}} \left(\prod_{t'=t+1}^{T} e^{-\beta_g{}^{(t'+1)} f(x^{(t'+1)})} \mathcal{L}^{(t')}(x^{(t'+1)} \mid x^{(t')}) \right) \tag{5.196}$$

$$\times e^{-\beta_g{}^{(t+1)} f(x^{(t+1)})} \mathcal{L}^{(t)}(x^{(t+1)} \mid x') \tag{5.197}$$

$$\times e^{-\beta_g{}^{(t)} f(x')} \mathcal{L}^{(t-1)}(x' \mid x^{(t-1)}) \left(\prod_{t'=-1}^{t-2} e^{-\beta_g{}^{(t'+1)} f(x^{(t'+1)})} \mathcal{L}^{(t')}(x^{(t'+1)} \mid x^{(t')}) \right) \tag{5.198}$$

$$= \left[\sum_{x^{(t+1)}, x^{(t+2)}, \dots x^{(T)}} \left(\prod_{t'=t+1}^{T} e^{-\beta_g{}^{(t'+1)} f(x^{(t'+1)})} \mathcal{L}^{(t')}(x^{(t'+1)} \mid x^{(t')}) \right) \right. \tag{5.199}$$

$$\left. \times e^{-\beta_g{}^{(t+1)} f(x^{(t+1)})} \mathcal{L}^{(t)}(x^{(t+1)} \mid x') \right] \tag{5.200}$$

$$\times \left[\sum_{x^{(0)}, x^{(1)}, \dots, x^{(t-1)}} e^{-\beta_g{}^{(t)} f(x')} \mathcal{L}^{(t-1)}(x' \mid x^{(t-1)}) \right. \tag{5.201}$$

$$\left. \left(\prod_{t'=-1}^{t-2} e^{-\beta_g{}^{(t'+1)} f(x^{(t'+1)})} \mathcal{L}^{(t')}(x^{(t'+1)} \mid x^{(t')}) \right) \right] \tag{5.202}$$

$$= A u^{(t:T)}(x') \mathbb{P}_{\mathrm{B}}^{(t)}(x'), \tag{5.203}$$

where we used the pathwise formulation (5.7) until $T = t$ in the last transformation and

$$A := \sum_{\mathbb{X}^{(t)} \in \mathcal{X}^{t+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(t)}, \mathbb{X}^{(t)} \right\rangle}, \tag{5.204}$$

is the normalization factor of the pathwise formulation. Therefore,

$$\sum_{\mathbb{X}^{(T)} \backslash x^{(t)}} \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x'}] = \frac{A u^{(t:T)}(x') \mathbb{P}_{\mathrm{B}}^{(t)}(x')}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}'^{(T)}] \right\rangle} \mathbb{P}_{\mathrm{F}}[\mathbb{X}'^{(T)}]}. \tag{5.205}$$

By a similar calculation, we have

$$\sum_{\mathbb{X}^{(T)} \backslash x^{(t)}, x^{(t+1)}} \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x', x^{(t+1)}=} x] \tag{5.206}$$

$$= \frac{A u^{(t+1:T)}(x) e^{-\beta_g{}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x') \mathbb{P}_{\mathrm{B}}^{(t)}(x')}{\sum_{\mathbb{X}'^{(T)} \in \mathcal{X}^{T+1}} e^{-\left\langle \mathbb{\beta}_g{}^{(T)}, f[\mathbb{X}'^{(T)}] \right\rangle} \mathbb{P}_{\mathrm{F}}[\mathbb{X}'^{(T)}]}. \tag{5.207}$$

By using these two equalities, we have

$$\mathcal{L}_{T,\mathrm{B}}^{(t)}(x \mid x') = \frac{\sum_{\mathbb{X}^{(T)} \backslash x^{(t)}, x^{(t+1)}} \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x', x^{(t+1)}=} x]}{\sum_{\mathbb{X}^{(T)} \backslash x^{(t)}} \mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \restriction_{x^{(t)}=x'}]} \tag{5.208}$$

$$= \frac{u^{(t+1:T)}(x) e^{-\beta_g{}^{(t+1)} f(x)} \mathcal{L}^{(t)}(x \mid x')}{u^{(t:T)}(x')}. \tag{5.209}$$

□

To prove Theorem 5.14, we need the following definitions and lemmas. For $t \le t'$, let

$$u'_{+1}[\mathbb{X}^{(t)}] := \prod_{s=0}^{t} u'^{(s)}_{+1}(x^{(s)}), \tag{5.210}$$

$$u'_{+1}[\mathbb{X}^{(t:t')}] := \prod_{s=t}^{t'} u'^{(s)}_{+1}(x^{(s)}). \tag{5.211}$$

The convention $\mathcal{L}^{(-1)}(x^{(0)} \mid x^{(-1)}) := \nu(x^{(0)})$ implies that $u^{(-1:T)}(x)$ is the normalization factor of (5.7). To emphasize the fact that $u^{(-1:T)}(x)$ is independent of x, we denote it by $u^{(-1:T)}$. The convention also implies that $u'^{(-1)}_{+1}(x) = \sum_{x' \in \mathcal{X}} e^{-\beta_g{}^{(0)} f(x')} \nu(x')$, which is independent of x.

Lemma 5.26. For any path function $g \colon \mathcal{X}^{T+1} \to \mathbb{R}$,

$$\mathbb{E}_{\mathbb{P}_\mathrm{B}}[g[\mathbb{X}^{(T)}]] = \frac{\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}] g[\mathbb{X}^{(T)}]]}{\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]]}. \tag{5.212}$$

Proof. The transition matrices of the retrospective process (5.35) and its one-step approximation (5.39) satisfies

$$\frac{u^{(t:T)}(x')}{u^{(t+1:T)}(x)} \mathcal{L}^{(t)}_{T,\mathrm{B}}(x \mid x') = u'^{(t)}_{+1}(x') \mathcal{L}^{(t)}_{+1,\mathrm{B}}(x \mid x'). \tag{5.213}$$

Therefore, by comparing (5.36) to (5.40), we have

$$\left(\prod_{t'=-1}^{T-1} \frac{u^{(t':T)}(x^{(t')})}{u^{(t'+1:T)}(x^{(t'+1)})} \right) \mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}] = \left(\prod_{t'=-1}^{T-1} u'^{(t')}_{+1}(x^{(t)}) \right) \mathbb{P}_\mathrm{R}[\mathbb{X}^{(T)}]. \tag{5.214}$$

Equivalently,

$$u^{(-1:T)} \mathbb{P}_\mathrm{B}[\mathbb{X}^{(T)}] = u'_{+1}[\mathbb{X}^{(-1:T-1)}] \mathbb{P}_\mathrm{R}[\mathbb{X}^{(T)}]. \tag{5.215}$$

By taking the summation $\sum_{\mathbb{X} \in \mathcal{X}^{T+1}}$ of the both hand side of (5.215), we have

$$u^{(-1:T)} = \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]]. \tag{5.216}$$

Also, multiplying $g[\mathbb{X}^{(T)}]$ and taking the summation $\sum_{\mathbb{X} \in \mathcal{X}^{T+1}}$ of the both hand side of (5.215), we have

$$u^{(-1:T)} \mathbb{E}_{\mathbb{P}_\mathrm{B}}[g[\mathbb{X}^{(T)}]] = \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}] g[\mathbb{X}^{(T)}]]. \tag{5.217}$$

Using these two equations, we have the statement of the lemma. □

Lemma 5.27.

$$u^{(t:T)}(x) = \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t:T-1)}] \mid x^{(t)} = x]. \tag{5.218}$$

Proof. We can prove this lemma by a similar argument to the derivation of (5.216). □

Proof of Theorem 5.14. By Lemma 5.26, we have

$$\mathbb{E}_{\mathbb{P}_\mathrm{B}}[g[\mathbb{X}^{(T)}]] = \frac{\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]g[\mathbb{X}^{(T)}]]}{\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]]}. \tag{5.219}$$

Therefore, it suffices to prove

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]g[\mathbb{X}^{(T)}]] && (5.220)\\
&\le \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(-1:T-1)}]]\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(T)}]]. && (5.221)
\end{aligned}$$

In the following, we prove the following by induction on $t' = T, \dots, 0$: For any $x^{(t'-1)} \in \mathcal{X}$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{(t':T)}] \mid x^{(t'-1)}] && (5.222)\\
&\le \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}] \mid x^{(t'-1)}]\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t'-1)}]. && (5.223)
\end{aligned}$$

Here, we specially define that $\mathbb{E}_{\mathbb{P}_\mathrm{R}}[\cdot \mid x^{(-1)}] = \mathbb{E}_{\mathbb{P}_\mathrm{R}}[\cdot]$ and $u'_{+1}[\mathbb{X}^{(T:T-1)}] = 1$. We note that the above statement for $t' = 0$ implies the statement of the theorem. Indeed, since ${u'}_{+1}^{(-1)}(x^{(-1)})$ is independent of $x^{(-1)}$, the statement for $t' = 0$ implies that (5.219).

The base case $t' = T$ trivially holds. We consider the step case t'. For simplicity, we omit the conditioning $x^{(t'-1)}$ of the expectation in the rest of the proof. By the linearity of the expectation, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{(t':T)}]] && (5.224)\\
&= \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]{\beta_w}^{(t')}f(x^{(t')})] + \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{((t'+1):T)}]]. && (5.225)
\end{aligned}$$

For the second term in the right hand side of (5.224), we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{((t'+1):T)}]] && (5.226)\\
&= \mathbb{E}_{\mathbb{P}_\mathrm{R}^{(t')}(x^{(t')})}\left[{u'}_{+1}^{(t')}(x^{(t')})\mathbb{E}_{\mathbb{P}_\mathrm{B}[\mathbb{X}^{((t'+1):T)}]}\left[u'_{+1}[\mathbb{X}^{(t'+1:T-1)}]g[\mathbb{X}^{(t'+1:T)}] \mid x^{(t')}\right]\right]. && (5.227)
\end{aligned}$$

By using ${u'}_{+1}^{(t')}(x^{(t')}) > 0$ and the induction hypothesis for $t' + 1$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{((t'+1):T)}]] && (5.228)\\
&\le \mathbb{E}_{\mathbb{P}_\mathrm{R}^{(t')}(x^{(t')})}\left[{u'}_{+1}^{(t')}(x^{(t')})\mathbb{E}_{\mathbb{P}_\mathrm{B}[\mathbb{X}^{((t'+1):T)}]}\left[u'_{+1}[\mathbb{X}^{(t'+1:T-1)}] \mid x^{(t')}\right]\right. && (5.229)\\
&\qquad\qquad \left.\times\mathbb{E}_{\mathbb{P}_\mathrm{R}}\left[g[\mathbb{X}^{(t'+1:T)}] \mid x^{(t')}\right]\right] && (5.230)\\
&\le \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t'+1:T)}] \mid x^{(t')}]]. && (5.231)
\end{aligned}$$

By (5.224) and (5.228), we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{(t':T)}]] && (5.232)\\
&\le \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t')}]] && (5.233)\\
&= \mathbb{E}_{\mathbb{P}_\mathrm{R}^{(t')}}[u^{(t':T)}(x^{(t')})\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t')}]] && (5.234)\\
&= \mathbb{E}_{\mathbb{P}_\mathrm{R}^{(t')}}[u^{(t':T)}(x^{(t')})]\mathbb{E}_{\mathbb{P}_\mathrm{R}^{(t')}}\left[\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t')}]]\right] && (5.235)\\
&\qquad + \mathrm{Cov}_{\mathbb{P}_\mathrm{R}^{(t')}}\left[u^{(t':T)}(x^{(t')}), \mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t')}]\right] && (5.236)\\
&= \mathbb{E}_{\mathbb{P}_\mathrm{R}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]]\mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}]] && (5.237)\\
&\qquad + \mathrm{Cov}_{\mathbb{P}_\mathrm{R}^{(t')}}\left[u^{(t':T)}(x^{(t')}), \mathbb{E}_{\mathbb{P}_\mathrm{R}}[g[\mathbb{X}^{(t':T)}] \mid x^{(t')}]\right], && (5.238)
\end{aligned}$$

where we used Lemma 5.27. By Assumption 5.13, the covariance in the above inequality is not positive [95]. Therefore,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[u'_{+1}[\mathbb{X}^{(t':T-1)}]g[\mathbb{X}^{(t':T)}]] \leq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[u^{(t':T)}(x^{(t')})]\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(t':T)}]], \tag{5.239}$$

which is the induction hypothesis for t'. □

Proof of Lemma 5.17. By solving (5.47) for $\xi^{(t)}$ and applying Assumption 5.16, we have the statement. □

Proof of Lemma 5.19. Let us take $x^{(t)}$. We then sample $x^{(t+1)}$ and $x'^{(t+1)}$ from $\mathcal{L}^{(t)}_{+1,\mathrm{B}}(\cdot \mid x^{(t)})$. Suppose that $f(x^{(t+1)}) \geq f(x'^{(t+1)})$.

We first prove (5.44). By Assumption 5.18, it suffices to prove that

$$u^{(t+1:t+k)}(x^{(t+1)}) \leq u^{(t+1:t+k)}(x'^{(t+1)}). \tag{5.240}$$

We prove by induction on $i = 1, 2, 3, \ldots, k$ that

$$u^{(t+1:t+i)}_{\mathbb{\beta}_g{}^{(T)}}(x^{(t+1)}) \leq u^{(t+1:t+i)}_{\mathbb{\beta}_g{}^{(T)}}(x'^{(t+1)}), \tag{5.241}$$

for any of $\mathbb{\beta}_g{}^{(T)}$. We add the subscript of $u^{(t+1:t+i)}$ to clarify the dependency on $\mathbb{\beta}_g{}^{(T)}$. The base case $i = 1$ trivially holds because

$$u^{(t+1:t+1)}(x^{(t+1)}) = u^{(t+1:t+1)}(x'^{(t+1)}) = 1. \tag{5.242}$$

We consider the step case $i > 1$. By definition,

$$u^{(t+1:t+i)}_{\mathbb{\beta}_g{}^{(T)}}(x^{(t+1)}) \tag{5.243}$$

$$= \sum_{\mathbb{X}^{(t+2:t+i)} \in \mathcal{X}^{i-1}} \prod_{s=t+1}^{t+i-1} e^{-\beta_g{}^{(s+1)} f(x^{(s+1)})} \mathcal{L}^{(s)}(x^{(s+1)} \mid x^{(s)}) \tag{5.244}$$

$$= \sum_{\mathbb{X}^{(t+2:t+i-1)} \in \mathcal{X}^{i-2}} \prod_{s=t+1}^{t+i-2} e^{-\beta_g{}^{(s+1)} f(x^{(s+1)})} \mathcal{L}^{(s)}(x^{(s+1)} \mid x^{(s)}) \tag{5.245}$$

$$\times \sum_{x^{(t+i)} \in \mathcal{X}} e^{-\beta_g{}^{(t+i)} f(x^{(t+i)})} \mathcal{L}^{(s)}(x^{(t+i)} \mid x^{(t+i-1)}) \tag{5.246}$$

$$= \sum_{\mathbb{X}^{(t+2:t+i-1)} \in \mathcal{X}^{i-2}} u'^{(t+i-1)}_{+1}(x^{(t+i-1)}) \prod_{s=t+1}^{t+i-2} e^{-\beta_g{}^{(s+1)} f(x^{(s+1)})} \mathcal{L}^{(s)}(x^{(s+1)} \mid x^{(s)}). \tag{5.247}$$

By Assumption 5.16,

$$
\begin{aligned}
&u^{(t+1:t+i)}(x^{(t+1)}) \qquad\qquad \text{(5.248)}\\
&= \sum_{\mathbb{X}^{(t+2:t+i-1)}\in\mathcal{X}^{i-2}} e^{-\beta_g{}^{(t+i)}\left(f(x^{(t+i-1)})-\eta^{(t+i)}\|\nabla f(x^{(t)})\|^2\right)} \qquad \text{(5.249)}\\
&\quad\times e^{\frac{(\beta_g{}^{(t+i)}\eta^{(t+i)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}} \qquad \text{(5.250)}\\
&\quad\times \prod_{s=t+1}^{t+i-2} e^{-\beta_g{}^{(s+1)}f(x^{(s+1)})}\mathcal{L}^{(s)}(x^{(s+1)}\mid x^{(s)}) \qquad \text{(5.251)}\\
&= e^{\beta_g{}^{(t+i)}\eta^{(t+i)}\|\nabla f(x^{(t)})\|^2+\frac{(\beta_g{}^{(t+i)}\eta^{(t+i)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}} \qquad \text{(5.252)}\\
&\quad\times \sum_{\mathbb{X}^{(t+2:t+i-1)}\in\mathcal{X}^{i-2}}\prod_{s=t+1}^{t+i-2} e^{-\beta_g'{}^{(s+1)}f(x^{(s+1)})}\mathcal{L}^{(s)}(x^{(s+1)}\mid x^{(s)}) \qquad \text{(5.253)}\\
&= e^{\beta_g{}^{(t+i)}\eta^{(t+i)}\|\nabla f(x^{(t)})\|^2+\frac{(\beta_g{}^{(t+i)}\eta^{(t+i)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}}\times u^{(t+1:t+i-1)}_{\boldsymbol{\beta}_g'{}^{(T)}}(x^{(t+1)}), \qquad \text{(5.254)}
\end{aligned}
$$

where $\boldsymbol{\beta}_g'{}^{(T)}=\{\beta_g'{}^{(s)}\}_s$ satisfies that $\beta_g'{}^{(t+i-1)}=\beta_g{}^{(t+i-1)}+\beta_g{}^{(t+i)}$ and $\beta_g'{}^{(s)}=\beta_g{}^{(s)}$ otherwise. The induction hypothesis for $\boldsymbol{\beta}_g'{}^{(T)}$ implies that

$$
u^{(t+1:t+i-1)}_{\boldsymbol{\beta}_g'{}^{(T)}}(x^{(t+1)})\le u^{(t+1:t+i-1)}_{\boldsymbol{\beta}_g'{}^{(T)}}(x'^{(t+1)}). \tag{5.255}
$$

By this inequality and (5.248) for $x^{(t+1)}$ and $x'^{(t+1)}$, we have

$$
\begin{aligned}
&u^{(t+1:t+i)}_{\boldsymbol{\beta}_g{}^{(T)}}(x^{(t+1)}) \qquad\qquad \text{(5.256)}\\
&= e^{\beta_g{}^{(t+i)}\eta^{(t+i)}\|\nabla f(x^{(t)})\|^2+\frac{(\beta_g{}^{(t+i)}\eta^{(t+i)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}}\times u^{(t+1:t+i-1)}_{\boldsymbol{\beta}_g'{}^{(T)}}(x^{(t+1)}) \qquad \text{(5.257)}\\
&\le e^{\beta_g{}^{(t+i)}\eta^{(t+i)}\|\nabla f(x^{(t)})\|^2+\frac{(\beta_g{}^{(t+i)}\eta^{(t+i)})^2}{2}\|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}}\times u^{(t+1:t+i-1)}_{\boldsymbol{\beta}_g'{}^{(T)}}(x'^{(t+1)}) \qquad \text{(5.258)}\\
&\le u^{(t+1:t+i)}_{\boldsymbol{\beta}_g{}^{(T)}}(x'^{(t+1)}), \qquad \text{(5.259)}
\end{aligned}
$$

which is the induction hypothesis for i.

We next prove (5.45). By Assumption 5.18, it suffices to prove

$$
\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(t+1:t+k)}]\mid x^{(t+1)}]\ge\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g[\mathbb{X}^{(t+1:t+k)}]\mid x'^{(t+1)}]. \tag{5.260}
$$

We prove by induction on $i=1,2,\dots,k$ that

$$
\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\boldsymbol{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}]\mid x^{(t+1)}]\ge\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\boldsymbol{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}]\mid x'^{(t+1)}], \tag{5.261}
$$

for any $\boldsymbol{\beta}_w{}^{(T)}$. We added the subscript of g to clarify the dependency on $\boldsymbol{\beta}_w{}^{(T)}$. The base case $i=1$ trivially holds. We consider the step case $i>1$. By a direct calculation,

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\boldsymbol{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}]\mid x^{(t+1)}] \qquad \text{(5.262)}\\
&=\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\boldsymbol{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}]+\beta_w{}^{(t+i)}f(x^{(t+i)})\mid x^{(t+1)}] \qquad \text{(5.263)}\\
&=\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\boldsymbol{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}]\mid x^{(t+1)}]+\mathbb{E}_{\mathbb{P}_{\mathrm{B}}}[\beta_w{}^{(t+i)}f(x^{(t+i)})\mid x^{(t+1)}]. \qquad \text{(5.264)}
\end{aligned}
$$

For the second term, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[{\beta_w}^{(t+i)} f(x^{(t+i)}) \mid x^{(t+1)}] \tag{5.265}$$

$$= \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[{\beta_w}^{(t+i)} f(x^{(t+i)}) \mid x^{(t+i-1)}] \mid x^{(t+1)}]. \tag{5.266}$$

By Assumption 5.16,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[f(x^{(t+i)}) \mid x^{(t+i-1)}] \tag{5.267}$$

$$= f(x^{(t+i-1)}) - \eta^{(t+i)} \|\nabla f(x^{(t)})\|^2 - {\beta_g}^{(t+i)} (\eta^{(t+i)})^2 \|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})}. \tag{5.268}$$

Therefore, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}] \mid x^{(t+1)}] \tag{5.269}$$

$$= \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x^{(t+1)}] \tag{5.270}$$

$$+ {\beta_w}^{(t+i)} \mathbb{E}_{\mathbb{P}_{\mathrm{B}}} \Bigg[f(x^{(t+i-1)}) - \eta^{(t+i)} \|\nabla f(x^{(t)})\|^2 \tag{5.271}$$

$$- {\beta_g}^{(t+i)} (\eta^{(t+i)})^2 \|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} \mid x^{(t+1)} \Bigg] \tag{5.272}$$

$$= -{\beta_w}^{(t+i)} \eta^{(t+1)} \left(\|\nabla f(x^{(t)})\|^2 + {\beta_g}^{(t+i)} \eta^{(t+i)} \|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} \right) \tag{5.273}$$

$$+ \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}'_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x^{(t+1)}], \tag{5.274}$$

where $\mathbb{\beta}'_w{}^{(T)} := \{{\beta'_w}^{(s)}\}_s$ satisfies ${\beta'_w}^{(t+i-1)} = {\beta_w}^{(t+i-1)} + {\beta_w}^{(t+i)}$ and ${\beta'_w}^{(s)} = {\beta_w}^{(s)}$ otherwise. By the induction hypothesis for $i-1$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}'_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x^{(t+1)}] \geq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}'_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x'^{(t+1)}]. \tag{5.275}$$

We therefore have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}] \mid x^{(t+1)}] \tag{5.276}$$

$$= -{\beta_w}^{(t+i)} \eta^{(t+1)} \left(\|\nabla f(x^{(t)})\|^2 + {\beta_g}^{(t+i)} \eta^{(t+i)} \|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} \right) \tag{5.277}$$

$$+ \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}'_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x^{(t+1)}] \tag{5.278}$$

$$\geq -{\beta_w}^{(t+i)} \eta^{(t+1)} \left(\|\nabla f(x^{(t)})\|^2 + {\beta_g}^{(t+i)} \eta^{(t+i)} \|\nabla f(x^{(t)})\|^2_{\Sigma(x^{(t)})} \right) \tag{5.279}$$

$$+ \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}'_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i-1)}] \mid x'^{(t+1)}] \tag{5.280}$$

$$= \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[g_{\mathbb{\beta}_w{}^{(T)}}[\mathbb{X}^{(t+1:t+i)}] \mid x'^{(t+1)}], \tag{5.281}$$

which is the induction hypothesis for i. □

Proof of Theorem 5.20. Since f_A is $(A\alpha)$-strongly convex and $(A\gamma)$-smooth, Theorem 3.10 for f_A implies that

$$\mathbb{E}_{\mathbb{P}_{\mathrm{F}}^{f_A}}[f_A(x^{(T)})] - f_A(x^*) \leq C \left(\frac{\sigma^2}{A\alpha T} + \frac{AD^2(\alpha+\gamma)}{T^2} \right). \tag{5.282}$$

From the discussion above Theorem 5.20,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{F}}^{f_A}}[f_A(x^{(t)})] = \mathbb{E}_{\mathbb{P}_{\mathrm{R}}^{f}}[f_A(x^{(t)})]. \tag{5.283}$$

By the definition of f_A,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}^{f}}[f_A(x^{(T)})] - f_A(x^*) = A \left(\mathbb{E}_{\mathbb{P}_{\mathrm{R}}^{f}}[f(x^{(T)})] - f(x^*) \right). \tag{5.284}$$

By combining these equations, we have the statement of the theorem. □

To prove Lemma 5.23, we need the following lemmas.

Lemma 5.28. For any $a, b \in \mathbb{R}^d$,

$$\|a+b\|^2 \leq 4\|a\|^2 + 4\|b\|^2. \tag{5.285}$$

Proof. By triangle inequality, we have

$$\|a+b\|^2 \leq \left(\|a\| + \|b\|\right)^2. \tag{5.286}$$

Since

$$\|a\| + \|b\| \leq 2\max(\|a\|, \|b\|), \tag{5.287}$$

we have

$$\|a+b\|^2 \leq 4\max(\|a\|^2, \|b\|^2) \leq 4\left(\|a\|^2 + \|b\|^2\right). \tag{5.288}$$

□

Lemma 5.29. If

$$\|x^{(t)} - x^*\|^2 > \frac{G_{\mathrm{B}}^2 \kappa^2}{t+2}, \tag{5.289}$$

then

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \mid x^{(t)}\right] \geq \kappa^{-1}\|x^{(t)} - x^*\|^2. \tag{5.290}$$

Proof. By Lemma 5.17, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \mid x^{(t)}\right] && (5.291)\\
&\geq \left\langle \nabla f(x^{(t)}) + \beta^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle && (5.292)
\end{aligned}$$

By (3.54) for $\alpha = 0$ and the optimality condition $\nabla f(x^*) = 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \mid x^{(t)}\right] && (5.293)\\
&= f(x^{(t)}) - f(x^*) + \left\langle \beta^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle && (5.294)\\
&\geq \left\langle \beta^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle. && (5.295)
\end{aligned}$$

By the definition of $\eta^{(t+1)}$ and $\beta^{(t+1)}$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \mid x^{(t)}\right] \geq \frac{(t+2)^{\theta_1}}{\alpha\kappa(G_{\mathrm{B}}\kappa)^{2\theta_1}} \left\langle \Sigma(x^{(t)})\nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle. \tag{5.296}$$

By applying $(\alpha, \theta_1, \Sigma)$-strong convexity, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^* \right\rangle \mid x^{(t)}\right] &\geq \frac{(t+2)^{\theta_1}}{\alpha\kappa(G_{\mathrm{B}}\kappa)^{2\theta_1}} \alpha\|x^{(t)} - x^*\|^{2(1+\theta_1)} && (5.297)\\
&\geq \kappa^{-1}\|x^{(t)} - x^*\|^2. && (5.298)
\end{aligned}$$

In the last transformation, we used the following consequence of the assumption:

$$\|x^{(t)} - x^*\|^{2\theta_1} \geq \left(\frac{G_{\mathrm{B}}^2\kappa^2}{t+2}\right)^{\theta_1}. \tag{5.299}$$

□

Lemma 5.30.

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] \leq 4G^2 + \frac{4\gamma^2(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|x^{(t)} - x^*\|^{2\theta_2}. \tag{5.300}$$

Proof. By Lemma 5.17 and the formula of the second moment of the normal distribution [83], we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{F}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] = \|\nabla f(x^{(t)})\|^2 + \mathrm{Tr}\left(\Sigma(x^{(t)})\right) \leq G^2, \tag{5.301}$$

and

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] \leq \left\|\left(I + \beta^{(t+1)}\eta^{(t+1)}\Sigma(x^{(t)})\right)\nabla f(x^{(t)})\right\|^2 + \mathrm{Tr}\left(\Sigma(x^{(t)})\right), \tag{5.302}$$

$$= \left\|\left(I + \frac{(t+2)^{\theta_1}}{\alpha\kappa^2(G_{\mathrm{B}}\kappa)^{2\theta_1}}\Sigma(x^{(t)})\right)\nabla f(x^{(t)})\right\|^2 + \mathrm{Tr}\left(\Sigma(x^{(t)})\right). \tag{5.303}$$

By Lemma 5.28, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] && (5.304)\\
&\leq 4\|\nabla f(x^{(t)})\|^2 + \frac{4(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|\Sigma(x^{(t)})\nabla f(x^{(t)})\|^2 + \mathrm{Tr}\left(\Sigma(x^{(t)})\right) && (5.305)\\
&\leq 4\|\nabla f(x^{(t)})\|^2 + \frac{4(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|\Sigma(x^{(t)})\nabla f(x^{(t)})\|^2 + 4\mathrm{Tr}\left(\Sigma(x^{(t)})\right) && (5.306)\\
&\leq 4\left(\|\nabla f(x^{(t)})\|^2 + \mathrm{Tr}\left(\Sigma(x^{(t)})\right)\right) + \frac{4(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|\Sigma(x^{(t)})\nabla f(x^{(t)})\|^2. && (5.307)
\end{aligned}$$

By (5.301),

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] \leq 4G^2 + \frac{4(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|\Sigma(x^{(t)})\nabla f(x^{(t)})\|^2. \tag{5.308}$$

By using $(\gamma, \theta_2, \Sigma)$-smoothness, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] \leq 4G^2 + \frac{4\gamma^2(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\|x^{(t)} - x^*\|^{2\theta_2}. \tag{5.309}$$

□

Lemma 5.31. Let $C > 0$ and $t \in \mathbb{N} \cup \{0\}$. Suppose that $Z^{(t)}, Z^{(t+1)} \in \mathbb{R}$ satisfy the following inequalities:

$$Z^{(t+1)} \leq \left(1 - \frac{2}{t+2}\right) Z_t + \frac{C}{(t+2)^2}, \tag{5.310}$$

$$Z^{(t)} \leq \frac{C}{t+2}. \tag{5.311}$$

Then,

$$Z^{(t+1)} \leq \frac{C}{t+3}. \tag{5.312}$$

Proof. By the assumption,

$$Z^{(t+1)} \leq \left(1 - \frac{2}{t+2}\right)\frac{C}{t+2} + \frac{C}{(t+2)^2} \tag{5.313}$$

$$= \frac{C}{t+2} - \frac{C}{(t+2)^2}. \tag{5.314}$$

By a direct calculation, we have

$$\frac{C}{t+3} - \left(\frac{C}{t+2} - \frac{C}{(t+2)^2}\right) = \frac{C}{(t+2)^2(t+3)} \geq 0. \tag{5.315}$$

Therefore, we have

$$Z^{(t+1)} \leq \frac{C}{t+3}. \tag{5.316}$$

□

Proof of Lemma 5.23 Theorem 5.24. This proof is an extension of [88, Lemma 1].

We prove Lemma 5.23 and Theorem 5.24 at the same time by induction on t. For the base case $t = 0$, Theorem 5.24 holds by the definition of G_{B}.

We next consider the step case. We assume that Theorem (5.24) holds for t and prove (5.112) and (5.113) holds for $t + 1$. We first prove Lemma (5.112). By Lemma 5.30,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2] \leq 4G^2 + \frac{4\gamma^2(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^{2\theta_2}]. \tag{5.317}$$

Since $0 < \theta_2 \leq 1$, the function $h(x) = x^{\theta_2}$ is concave. Therefore, Jensen's inequality implies that

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^{2\theta_2}] \leq \left(\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2]\right)^{\theta_2}. \tag{5.318}$$

By this inequality,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2] \leq 4G^2 + \frac{4\gamma^2(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\left(\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2]\right)^{\theta_2}. \tag{5.319}$$

By the induction hypothesis,

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2] \leq 4G^2 + \frac{4\gamma^2(t+2)^{2\theta_1}}{\alpha^2\kappa^2(G_{\mathrm{B}}\kappa)^{4\theta_1}}\left(\frac{3G_{\mathrm{B}}^2\kappa^2}{t+2}\right)^{\theta_2} \tag{5.320}$$

$$= 4G^2 + 4\frac{3^{\theta_2}\gamma^2}{\alpha^2\kappa^2}(G_{\mathrm{B}}\kappa)^{2\theta_2 - 4\theta_1}(t+2)^{2\theta_1 - \theta_2}. \tag{5.321}$$

Since $t + 2 \geq 1$ and $\theta_2 - 2\theta_1 \geq 0$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2] \leq 4G^2 + 4\frac{3^{\theta_2}\gamma^2}{\alpha^2\kappa^2}(G_{\mathrm{B}}\kappa)^{2\theta_2 - 4\theta_1} \leq G_{\mathrm{B}}^2. \tag{5.322}$$

In the last transformation, we used the definition of G_{B}.

We next prove (5.113) for $t+1$ by case analysis. We first treat the case where $\|x^{(t)} - x^*\|^2 \leq G_{\mathrm{B}}^2\kappa^2/(t+2)$. In this case, by the triangle inequality and the contracting property of $\Pi_{\mathcal{X}}$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2 \mid x^{(t)}] \tag{5.323}$$
$$\leq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|\Pi_{\mathcal{X}}\left(x^{(t)} - \eta^{(t+1)} g^{(t+1)}\right) - x^*\|^2 \mid x^{(t)}] \tag{5.324}$$
$$\leq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - \eta^{(t+1)} g^{(t+1)} - x^*\|^2 \mid x^{(t)}] \tag{5.325}$$
$$\leq \|x^{(t)} - x^*\|^2 + \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[(\eta^{(t+1)})^2 \|g^{(t+1)}\|^2] \mid x^{(t)}] \tag{5.326}$$
$$\leq \left(1 - \frac{2}{t+2}\right)\|x^{(t)} - x^*\|^2 + \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\| \mid x^{(t)}] + \frac{2G_{\mathrm{B}}^2\kappa^2}{(t+2)^2}. \tag{5.327}$$

We next treat the other case where $\|x^{(t)} - x^*\|^2 > G_{\mathrm{B}}^2\kappa^2/(t+1)$. By a similar argument, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2 \mid x^{(t)}] \tag{5.328}$$
$$\leq \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - \eta^{(t+1)} g^{(t+1)} - x^*\|^2 \mid x^{(t)}] \tag{5.329}$$
$$= \|x^{(t)} - x^*\|^2 - 2\eta^{(t+1)}\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle \mid x^{(t)}\right] \tag{5.330}$$
$$+ \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}]. \tag{5.331}$$

By Lemma 5.29, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2 \mid x^{(t)}] \tag{5.332}$$
$$\leq \|x^{(t)} - x^*\|^2 - \frac{2\eta^{(t+1)}}{\kappa}\|x^{(t)} - x^*\|^2 + \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] \tag{5.333}$$
$$= \left(1 - \frac{2}{t+2}\right)\|x^{(t)} - x^*\|^2 + \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}]. \tag{5.334}$$

Therefore, for all cases, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2 \mid x^{(t)}] \leq \left(1 - \frac{2}{t+2}\right)\|x^{(t)} - x^*\|^2 \tag{5.335}$$
$$+ \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2 \mid x^{(t)}] + \frac{2G_{\mathrm{B}}^2\kappa^2}{(t+2)^2}. \tag{5.336}$$

By taking the expectation with respect to $x^{(t)}$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2] \tag{5.337}$$
$$\leq \left(1 - \frac{2}{t+2}\right)\mathbb{E}\left[\|x^{(t)} - x^*\|^2\right] + \left(\eta^{(t+1)}\right)^2 \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|g^{(t+1)}\|^2] + \frac{2G_{\mathrm{B}}^2\kappa^2}{(t+2)^2}. \tag{5.338}$$

By (5.112) and the definition of $\eta^{(t+1)}$, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t+1)} - x^*\|^2] \leq \left(1 - \frac{2}{t+2}\right)\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2] + \frac{3G_{\mathrm{B}}^2\kappa^2}{(t+2)^2}. \tag{5.339}$$

By Lemma 5.31 for $Z^{(t)} := \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2]$, we have (5.113) for $t+1$. □

Proof of Theorem 5.25. By Theorem 3.11, it suffices to prove (3.75) for $\mathbb{P} := \mathbb{P}_{\mathrm{R}}$ and $C := 3\kappa G_{\mathrm{B}}^2$. By Lemma 5.17, we have

$$\mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle g^{(t+1)}, x^{(t)} - x^*\right\rangle\right] \tag{5.340}$$
$$= \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}\left[\left\langle \nabla f(x^{(t)}), x^{(t)} - x^*\right\rangle + \frac{\beta^{(t+1)}\eta^{(t+1)}}{2}\left\langle \Sigma(x^{(t)})\nabla f(x^{(t)}), x^{(t)} - x^*\right\rangle\right]. \tag{5.341}$$

We evaluate two terms in the right hand side. For the first term, the convexity of f implies that

$$\left\langle \nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle \geq f(x^{(t)}) - f(x^*), \tag{5.342}$$

since $\nabla f(x^*) = 0$. For the second term, the $(\alpha, \theta_1, \Sigma)$-strong convexity implies that

$$\left\langle \Sigma(x^{(t)}) \nabla f(x^{(t)}), x^{(t)} - x^* \right\rangle \geq \alpha \|x^{(t)} - x^*\|^{2(1+\theta_1)} \geq 0. \tag{5.343}$$

Theorem 5.24 implies that

$$\frac{1}{\kappa} \mathbb{E}_{\mathbb{P}_{\mathrm{R}}}[\|x^{(t)} - x^*\|^2] \leq \frac{3\kappa G_{\mathrm{B}}^2}{t+2}. \tag{5.344}$$

By these three inequalities, we know that (3.75) holds for the probability measure $\mathbb{P}_{\mathrm{R}}$ and $C = 3\kappa G_{\mathrm{B}}^2$. □

Chapter 6

Conclusion

In this thesis, we solved the problem about the coordination between individual learning and populational evolution in each field from a unified point of view. In Chapter 4, we showed that individual learning from experience can accelerate populational evolution in biological evolution. We also quantified the acceleration by extending FF-thm. In Chapter 5, we consider the converse: We showed populational evolution of the BA can accelerate individual learning. Concretely, we proved that iterative optimization algorithms are accelerated by incorporating populational evolution in the form of the BA.

What played an important role in our thesis is the bridging of the concepts and techniques in each field. For example, we observed that the perspective of HMM was useful to analyze population dynamics in Chapter 4. Moreover, concepts from population dynamics like FF-thm and retrospective process was useful to analyze the performance of the BA in Chapter 5. Background of the bridging is the Feynman-Kac formula [68]: All population dynamics (3.11), the BA (5.7), HMM (3.46) considered the measure transformation from $\mathbb{P}_{\mathrm{F}}[\mathbb{X}^{(T)}]$ to $\mathbb{P}_{\mathrm{B}}[\mathbb{X}^{(T)} \mid \mathbb{Y}^{(t)}]$ by a multiplicative factor $e^{k[\mathbb{X}^{(T)},\mathbb{Y}^{(t)}]}$; This type of measure transformation is called the Feynman-Kac formula and appears in various fields beyond what we have considered in this thesis. Therefore, our analysis of the coordination between individual learning and populational evolution has a potential to further apply to different fields. Moreover, importing ideas from other fields might advance the results in this thesis.

A possible application of our analysis is parameter estimation by SMC methods. As discussed in Chapter 1, we can propose a new SMC method to estimate parameters of HMM by considering coordination between individual learning and populational evolution. To theoretically analyze the performance of this method, the techniques we developed in Chapter 4 might be useful. For example, the extended FF-thm of ancestral learning might have some implications on the performance since the population fitness corresponds to log-likelihood.

Another possible application is the acceleration of individual learning by populational evolution in biology. When individual learning of an organism is at the same time scale as replication, the individual learning might be accelerated by populational evolution. Indeed, some fraction of the organism that happens to learn more successfully has more daughters and thus such events are emphasized by the survivorship bias caused by populational evolution. Analyzing the acceleration is equivalent to quantify the performance of the memetic algorithms. Therefore, the results in Chapter 5 might be useful for this analysis.

Since the remaining problems of the thesis in each field might are related to each other due to the bridging, a unified view like this thesis might be useful to solve. For example, the dependency of the agents in the population is a difficult problem in each field. In biology, we can consider communication among agents, which might improve the acceleration of evolutionary process by learning as indicated in Section 4.10. However, the analysis in this situation is difficult due to the dependency of agents caused by communication. In information systems, recombination induces the correlation between solutions, which makes the theoretical analysis of the performance difficult. A unified analysis based on new techniques, like mean-field approximation and perturbation theory, might advance the analysis of both topics.

References

[1] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[2] Lee Altenberg. The schema theorem and price's theorem. volume 3 of *Foundations of Genetic Algorithms*, pages 23–49. Elsevier, 1995.

[3] Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 02 1998.

[4] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[5] Ludwig Arnold, Volker Matthias Gundlach, and Lloyd Demetrius. Evolutionary formalism for products of positive random matrices. *Ann. Appl. Probab.*, 4(3):859–901, 08 1994.

[6] Ellen Baake and Hans-Otto Georgii. Mutation, selection, and ancestry in branching models: a variational approach. *Journal of Mathematical Biology*, 54(2):257–303, Feb 2007.

[7] Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627, 2014.

[8] Vincent Bansaye. Ancestral lineages and limit theorems for branching markov chains in varying environment. *Journal of Theoretical Probability*, 32(1):249–281, Mar 2019.

[9] James O Berger. *Statistical decision theory and Bayesian analysis (2nd edition)*. Springer Science & Business Media New York, 1985.

[10] Hans-Georg Beyer. Toward a theory of evolution strategies: Some asymptotical results from the $(1 \underset{,}{+} \lambda)$-theory. *Evolutionary Computation*, 1(2):165–188, 1993.

[11] Hans-Georg Beyer. Toward a theory of evolution strategies: The (μ, λ)-theory. *Evolutionary Computation*, 2(4):381–407, 1994.

[12] Hans-Georg Beyer. Toward a theory of evolution strategies: Self-adaptation. *Evol. Comput.*, 3(3):311–347, September 1995.

[13] J.D. Biggins, H. Cohn, and O. Nerman. Multi-type branching in varying environment. *Stochastic Processes and their Applications*, 83(2):357 – 400, 1999.

[14] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[15] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[16] Léon Bottou and Yann Cun. Large scale online learning. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

[17] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[18] Clayton L Bridges and David E. Goldberg. An analysis of reproduction and crossover in a binary-coded genetic algorithm. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic Algorithms and Their Application*, page 9–13, USA, 1987. L. Erlbaum Associates Inc.

[19] B. M. Brown. Martingale Central Limit Theorems. *The Annals of Mathematical Statistics*, 42(1):59 – 66, 1971.

[20] Harry Cohn. On the growth of the multitype supercritical branching process in a random environment. *Ann. Probab.*, 17(3):1118–1123, 07 1989.

[21] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms.* MIT press, 2009.

[22] Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

[23] Ashley M. Cunningham, Deena M. Walker, Aarthi Ramakrishnan, Marie. A. Doyle, Rosemary C. Bagot, Hannah M. Cates, Catherine J. Peña, Orna Issler, Casey Lardner, Caleb Browne, Scott J. Russo, Li Shen, and Eric J. Nestler. Sperm transcriptional state associated with paternal transmission of stress phenotypes. *Journal of Neuroscience*, 2021.

[24] Thomas Elder Davis. *Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm.* PhD thesis, University of Florida, 1991.

[25] Donald A. Dawson. Introductory lectures on stochastic population systems, 2017.

[26] Miguel de Carvalho. Mean, what do you mean? *The American Statistician*, 70(3):270–274, 2016.

[27] Imke G. de Jong, Patsy Haccou, and Oscar P. Kuipers. Bet hedging or not? a guide to proper classification of microbial survival strategies. *BioEssays*, 33(3):215–223, 2011.

[28] Del Moral, Pierre, Doucet, Arnaud, and Singh, Sumeetpal S. A backward particle interpretation of feynman-kac formulae. *ESAIM: M2AN*, 44(5):947–975, 2010.

[29] Joseph Leo "Joe" Doob. *Stochastic Processes.* WILEY, USA, 1991.

[30] Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples.* Chapman and Hall/CRC, 2014.

[31] A. E. Eiben, E. H. L. Aarts, and K. M. Van Hee. Global convergence of genetic algorithms: A markov chain analysis. In Hans-Paul Schwefel and Reinhard Männer, editors, *Parallel Problem Solving from Nature*, pages 3–12, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[32] A. E. Eiben and G. Rudolph. Theory of evolutionary algorithms: A bird's eye view. *Theor. Comput. Sci.*, 229(1–2):3–9, November 1999.

[33] Ronald Aylmer Fisher. *The genetical theory of natural selection.* The Clarendo Pressn, 1930.

[34] H. Furstenberg and H. Kesten. Products of random matrices. *Ann. Math. Statist.*, 31(2):457–469, 06 1960.

[35] Hans-Otto Georgii and Ellen Baake. Supercritical multitype branching processes: the ancestral types of typical individuals. *Advances in Applied Probability*, 35(4):1090–1110, 2003.

[36] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[37] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[38] P. Haccou and Y. Iwasa. Optimal mixed strategies in stochastic environments. *Theoretical Population Biology*, 47(2):212–243, 1995.

[39] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[40] Theodore Edward Harris. *The Theory of Branching Process.* RAND Corporation, Santa Monica, CA, 1964.

[41] Mikihiro Hashimoto, Takashi Nozoe, Hidenori Nakaoka, Reiko Okura, Sayo Akiyoshi, Kunihiko Kaneko, Edo Kussell, and Yuichi Wakamoto. Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences*, 113(12):3251–3256, 2016.

[42] H. Hennion. Limit theorems for products of positive random matrices. *Ann. Probab.*, 25(4):1545–1587, 10 1997.

[43] Joachim Hermisson, Oliver Redner, Holger Wagner, and Ellen Baake. Mutation–selection balance: Ancestry, load, and maximum principle. *Theoretical Population Biology*, 62(1):9 – 46, 2002.

[44] John A. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the theory of neural computation (1st ed.).* CRC Press, 1991.

[45] John Henry Holland et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* MIT press, 1992.

[46] Chonghai Hu, James Tin-Yau Kwok, and Weike Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[47] Galin L. Jones. On the markov chain central limit theorem. *Probab. Surveys*, 1:299–320, 2004.

[48] Owen Dafydd Jones. On the convergence of multitype branching processes with varying environments. *Ann. Appl. Probab.*, 7(3):772–801, 08 1997.

[49] Jürgen Jost. *Postmodern analysis.* Springer Science & Business Media, 2006.

[50] N. Kantas, A. Doucet, S.S. Singh, and J.M. Maciejowski. An overview of sequential monte carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes*, 42(10):774 – 785, 2009. 15th IFAC Symposium on System Identification.

[51] Shauharda Khadka, Somdeb Majumdar, Tarek Nassar, Zach Dwiel, Evren Tumer, Santiago Miret, Yinyin Liu, and Kagan Tumer. Collaborative evolutionary reinforcement learning. In *International Conference on Machine Learning*, pages 3341–3350. PMLR, 2019.

[52] Yuri Kifer. Perron-frobenius theorem, large deviations, and random perturbations in random environments. *Mathematische Zeitschrift*, 222(4):677–698, Jul 1996.

[53] Tetsuya J. Kobayashi and Yuki Sughiyama. Fluctuation relations of fitness and information in population dynamics. *Phys. Rev. Lett.*, 115:238102, Dec 2015.

[54] Tetsuya J. Kobayashi and Yuki Sughiyama. Fitness gain of individually sensed information by cells. *Entropy*, 21(10), 2019.

[55] Edo Kussell and Stanislas Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.

[56] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[57] Stanislas Leibler and Edo Kussell. Individual histories and selection in heterogeneous populations. *Proceedings of the National Academy of Sciences*, 107(29):13183–13188, 2010.

[58] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110. PMLR, 2017.

[59] Ben D. MacArthur, Avi Ma'ayan, and Ihor R. Lemischka. Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10):672–681, Oct 2009.

[60] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, January 2017.

[61] A. Marguet. Uniform sampling in a structured branching population. *ArXiv e-prints*, September 2016.

[62] A. Marguet. A law of large numbers for branching Markov processes by the ergodicity of ancestral lineages. *ArXiv e-prints*, July 2017.

[63] Heinz Mühlenbein and Dirk Schlierkamp-Voosen. Predictive models for the breeder genetic algorithm i. continuous parameter optimization. *Evolutionary Computation*, 1(1):25–49, 1993.

[64] D. J. McFarland. Decision making in animals. *Nature*, 269(5623):15–21, Sep 1977.

[65] D. L. McLeish. Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(3):165–178, Sep 1975.

[66] Taejin L. Min, Patrick J. Mears, Ido Golding, and Yann R. Chemla. Chemotactic adaptation kinetics of individual escherichia coli cells. *Proceedings of the National Academy of Sciences*, 109(25):9869–9874, 2012.

[67] A. G. M'Kendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925.

[68] Pierre Moral. *Feynman-Kac Formulae*. Springer-Verlag,New York, 2004.

[69] Pablo Moscato et al. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. *Caltech concurrent computation program, C3P Report*, 826:1989, 1989.

[70] Noboru Murata and Shun ichi Amari. Statistical analysis of learning dynamics. *Signal Processing*, 74(1):3–28, 1999.

[71] Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences*, 107(9):4248–4253, 2010.

[72] Ravinder Nagpal, Hirokazu Tsuji, Takuya Takahashi, Kazunari Kawashima, Satoru Nagata, Koji Nomoto, and Yuichiro Yamashiro. Sensitive quantitative analysis of the meconium bacterial microbiota in healthy term infants born vaginally or by cesarean section. *Frontiers in Microbiology*, 7:1997, 2016.

[73] Kento Nakamura and Tetsuya J. Kobayashi. Connection between the bacterial chemotactic network and optimal filtering. *Phys. Rev. Lett.*, 126:128102, Mar 2021.

[74] Kento Nakamura and Tetsuya J. Kobayashi. Optimal sensing and control of run-and-tumble chemotaxis, 2021.

[75] So Nakashima and Tetsuya J. Kobayashi. Acceleration of evolutionary processes by learning and extended fisher's fundamental theorem. *Phys. Rev. Research*, 4:013069, Jan 2022.

[76] So Nakashima, Yuki Sughiyama, and Tetsuya J Kobayashi. Lineage EM algorithm for inferring latent states from cellular lineage trees. *Bioinformatics*, 36(9):2829–2838, 01 2020.

[77] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[78] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[79] Allen E. Nix and Michael D. Vose. Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1):79–88, Mar 1992.

[80] Ryo Oizumi and Takenori Takada. Optimal life schedule with stochastic growth in age-size structured models: Theory and an application. *Journal of Theoretical Biology*, 323:76–89, 2013.

[81] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *Journal of Machine Learning Research*, 18(18):1–65, 2017.

[82] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[83] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook (version: November 15, 2012), 2012.

[84] Anya Plutynski. What was fisher's fundamental theorem of natural selection and what was it for? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 37(1):59–82, 2006.

[85] GEORGE R. PRICE. Fisher's ‘fundamental theorem’ made clear. *Annals of Human Genetics*, 36(2):129–140, 1972.

[86] Rami Pugatch. Greedy scheduling of cellular self-replication leads to optimal doubling times with a log-frechet distribution. *Proceedings of the National Academy of Sciences*, 112(8):2611–2616, 2015.

[87] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[88] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[89] Ingo Rechenberg. Evolutionsstrategie: Optimierung technischer systeme nach prinzipien der biologischen evolution. 1973. *frommann-holzbog, Stuttgart.*

[90] Olivier Rivoire and Stanislas Leibler. The value of information for populations in varying environments. *Journal of Statistical Physics*, 142(6):1124–1166, Apr 2011.

[91] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[92] Susan M. Rosenberg. Evolving responsively: adaptive mutation. *Nature Reviews Genetics*, 2(7):504–515, Jul 2001.

[93] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986.

[94] David Saad. Online algorithms and stochastic approximations. *Online Learning*, 5:6–3, 1998.

[95] Klaus D Schmidt. *On the covariance of monotone functions of a random variable*. Professoren des Inst. für Math. Stochastik, 2003.

[96] Lothar M. Schmitt. Theory of genetic algorithms. *Theoretical Computer Science*, 259(1):1–61, 2001.

[97] Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer-Verlag Berlin Heidelberg, 2003.

[98] Hans-Paul Paul Schwefel. *Evolution and Optimum Seeking: The Sixth Generation*. John Wiley & Sons, Inc., USA, 1993.

[99] Jon Seger. What is bet-hedging? *Oxford surveys in evolutionary biology*, 4:182–211, 1987.

[100] Timo Seppalainen. Large deviations for markov chains with random transitions. *Ann. Probab.*, 22(2):713–748, 04 1994.

[101] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, Mar 2011.

[102] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[103] Yuki Sughiyama, Tetsuya J. Kobayashi, Koji Tsumura, and Kazuyuki Aihara. Pathwise thermodynamic structure in population dynamics. *Phys. Rev. E*, 91:032120, Mar 2015.

[104] Yuki Sughiyama, So Nakashima, and Tetsuya J. Kobayashi. Fitness response relation of a multitype age-structured population dynamics. *Phys. Rev. E*, 99:012413, Jan 2019.

[105] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X. Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[106] Lisa A. Urry, Steven A. Wasserman, Peter V. Minorsky, and Rebecca Orr. *Campbell Biology (12th edition).* Pearson, 2020.

[107] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[108] Vladimir Vatutin and Vitali Wachtel. Multi-type subcritical branching processes in a random environment. *Advances in Applied Probability*, 50(A):281–289, 2018.

[109] H. von Foerster. Some remarks on changing populations. In J. F. Stohlman, editor, *The Kinetics of Cellular Proliferation*, pages 382–407. Grune and Stratton, New York, 1959.

[110] Michael D Vose, Gunar E Liepins, et al. Punctuated equilibria in genetic search. *Complex systems*, 5(1):31–44, 1991.

[111] Yuichi Wakamoto, Alexander Y. Grosberg, and Edo Kussell. Optimal lineage principle for age-structured populations. *Evolution*, 66(1):115–134, 2012.

[112] John Wakeley. *Coalescent Theory.* Robert and Company Publisher, 2009.

[113] Ping Wang, Lydia Robert, James Pelletier, Wei Lien Dang, Francois Taddei, Andrew Wright, and Suckjoon Jun. Robust growth of escherichia coli. *Current Biology*, 20(12):1099–1103, 2010.

[114] Edward W. Weissner. Multitype branching processes in random environments. *Journal of Applied Probability*, 8(1):17–31, 1971.

[115] BingKan Xue and Stanislas Leibler. Evolutionary learning of adaptation to varying environments through a transgenerational feedback. *Proceedings of the National Academy of Sciences*, 113(40):11266–11271, 2016.

[116] Keisuke Yoshida, Shin-ichiro Fujita, Ayako Isotani, Takashi Kudo, Satoru Takahashi, Masahito Ikawa, Dai Shiba, Masaki Shirakawa, Masafumi Muratani, and Shunsuke Ishii. Intergenerational effect of short-term spaceflight in mice. *iScience*, 24(7), Jul 2021.

[117] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 116, New York, NY, USA, 2004. Association for Computing Machinery.

[118] Xiao-Peng Zhang, Feng Liu, Zhang Cheng, and Wei Wang. Cell fate decision mediated by p53 pulses. *Proceedings of the National Academy of Sciences*, 106(30):12245–12250, 2009.

[119] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6792–6792, 2012.

[120] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach,*

California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663. PMLR, 2019.

[121] 庸 巌佐. 数理生物学入門　—生物社会のダイナミクスを探る—. 共立出版株式会社, 1998.

[122] 庸 巌佐. 生命の数理. 共立出版株式会社, 2008.

[123] 田崎晴明. 統計力学〈*1*〉 *(*新物理学シリーズ*)*. 培風館, 日本, 2008.

[124] 鈴木大慈. 確率的最適化 *(*機械学習プロフェッショナルシリーズ*)*. 講談社, 日本, 2015.