ABSTRACT

Title of Dissertation: SIMULATION OPTIMIZATION: EFFICIENT SELECTION AND DYNAMIC PROGRAMMING

Yi Zhou
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Michael Fu
Department of Decision, Operations,
and Information Technologies

Many real-world problems can be modeled as stochastic systems, whose performances can be estimated by simulations. Important topics include statistical ranking and selection (R&S) problems, response surface problems, and stochastic estimation problems.

The first part of the dissertation focuses on the R&S problems, where there is a finite set of feasible solutions ("systems" or "alternatives") with unknown values of performances that must be estimated from noisy observations, e.g., from expensive stochastic simulation experiments. The goal in R&S problems is to identify the highest-valued alternative as quickly as possible. In many practical settings, alternatives with similar attributes could have similar performances; we thus focus on such settings where prior similarity information between the performance outputs of different alternatives can be learned from data. Incorporating similarity information enables efficient budget allocation for faster identification of the best alternative in sequential selection. Using a new selection criterion, the similarity selection index, we develop two new allocation

methods: one based on a mathematical programming characterization of the asymptotically optimal budget allocation, and the other based on a myopic expected improvement measure. For the former, we present a novel sequential implementation that provably learns the optimal allocation without tuning. For the latter, we also derive its asymptotic sampling ratios. We also propose a practical way to update the prior similarity information as new samples are collected.

The second part of the dissertation considers the stochastic estimation problems of estimating a conditional expectation. We first formulate the conditional expectation as a ratio of two stochastic derivatives. Applying stochastic gradient estimation techniques, we express the two derivatives using ordinary expectations, which can be then estimated by Monte Carlo methods. Based on an empirical distribution estimated from simulation, we provide guidance on selecting the appropriate formulation of the derivatives to reduce the variance of the ratio estimator.

The third part of this dissertation further explores the application of estimators for conditional expectations in policy evaluation, an important topic in stochastic dynamic programming. In particular, we focus on a finite-horizon setting with continuous state variables. The Bellman equation represents the value function as a conditional expectation. By using the finite difference method and the generalized likelihood ratio method, we propose new estimators for policy evaluation and show how the value of any given state can be estimated using sample paths starting from various other states.

SIMULATION OPTIMIZATION: EFFICIENT SELECTION AND DYNAMIC PROGRAMMING

by

Yi Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Michael Fu, Chair/Advisor
Professor Ilya Ryzhov
Professor Steve Marcus
Professor Vincent Lyzinski
Professor Dilip B. Madan

Dedication

I dedicate this dissertation to my parents, Wensheng Zhou and Qiaohuan Fan.

Acknowledgments

First and foremost, I am deeply grateful to my advisor, Prof. Michael Fu, for his exceptional guidance and invaluable advice throughout my Ph.D. journey. He has been incredibly responsive, replying to my emails promptly and providing insightful feedback while editing my research papers. His encouragement and guidance were instrumental during many challenging moments of my research. Moreover, he has generously provided me with great opportunities to shape my career growth beyond the scope of my research projects. Without his constant patience and tremendous support, I would not have been able to complete this dissertation. Above all, I will always appreciate and learn from his exceptional character.

I would like to express my gratitude to Prof. Ilya Ryzhov for his previous time and great efforts in discussing my research and offering new ideas. I have learned a lot from him in academic writing. His energy and dedication to research are truly inspiring. I would also like to extend my thanks to Prof. Steve Marcus for taking the time to attend my research discussions. In addition to Prof. Michael Fu, Prof. Ilya Ryzhov, Prof. Steve Marcus, I would also like to thank Prof. Vincent Lyzinski and Prof. Dilip B. Madan for serving on my committee and devoting their time to examining my dissertation. It is an honor to have them on my committee.

Life would be dull without friendship. I am especially grateful for the inspiration brought to me by Qianni Jiang, whom I have known for many years. I also sincerely thank Yiran Zhang for her companionship and help during my Ph.D. years.

Most importantly, I owe a great deal to my parents for their unwavering love and unconditional support. Without their encouragement, I would not have been able to complete my Ph.D. study. I will be forever grateful to them and can never express my gratitude enough.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Tables vii

List of Figures viii

Chapter 1: Introduction 1
- 1.1 Ranking and Selection . . . 1
- 1.2 Estimating a Conditional Expectation With the Generalized Likelihood Ratio Method . . . 3
- 1.3 Policy Evaluation With Stochastic Gradient Estimation Techniques . . . 4

Chapter 2: Sequential Learning With a Similarity Selection Index 5
- 2.1 Introduction . . . 5
- 2.2 The S-index: Definition and Properties . . . 10
- 2.3 Optimal Budget Allocation Under S-index Selection . . . 21
  - 2.3.1 Geometric Approximation of the Probability of Correct Selection . . . 22
  - 2.3.2 Large Deviations Analysis of Probability of Incorrect Selection . . . 26
  - 2.3.3 A Provably Convergent Sequential Implementation Under S-index Selection (SIGD) . . . 30
  - 2.3.4 Convergence Analysis of SIGD . . . 38
- 2.4 Myopic Allocation Under S-index Selection (SIMA) . . . 58
  - 2.4.1 Formulation of An Myopic Allocation Policy . . . 58
  - 2.4.2 Convergence Analysis of SIMA . . . 60
- 2.5 Dynamic Updating of the Similarity Graph . . . 67
- 2.6 Handling Alternatives with Equal Values . . . 74
- 2.7 Numerical Experiments . . . 75
  - 2.7.1 Experiment 1: A Servo System Selection Problem . . . 76
  - 2.7.2 Experiment 2: A Dose Selection Problem . . . 79
  - 2.7.3 Experiment 3: M/G/1 Queues . . . 81
  - 2.7.4 Experiment 4: Rosenbrock Test Function . . . 82
  - 2.7.5 Experiments on Sensitivity to Hyperparameters . . . 83
  - 2.7.6 Summary of Results and Computation Time . . . 85

2.8 Conclusion . . . 86

Chapter 3: Estimating a Conditional Expectation With the Generalized Likelihood Ratio Method 88
3.1 Introduction . . . 88
3.2 Estimating a Conditional Mean . . . 90
3.2.1 Problem Formulation . . . 90
3.2.2 Stochastic Gradient Estimation . . . 92
3.3 Main Results . . . 93
3.3.1 Globally Invertible g_2 . . . 94
3.3.2 g_2 with Stationary Points . . . 99
3.3.3 Confidence Interval of the Ratio Estimator . . . 101
3.3.4 Variance Reduction . . . 102
3.4 Simulation Experiments . . . 103
3.4.1 Experiment 1: Financial Option . . . 103
3.4.2 Experiment 2: g_2 without Explicit Inverse . . . 105
3.4.3 Experiment 3: g_2 with Stationary Points . . . 107
3.5 Conclusion . . . 108

Chapter 4: Policy Evaluation With Stochastic Gradient Estimation Techniques 110
4.1 Introduction . . . 110
4.2 Problem Formulation . . . 112
4.2.1 Monte Carlo (MC) and Temporal Difference (TD) Methods . . . 113
4.3 New Estimators of the Value Function . . . 115
4.3.1 Ratio representation of a conditional expectation . . . 115
4.3.2 Stochastic gradient estimation (SGE) . . . 116
4.4 Algorithms . . . 119
4.5 Numerical Experiments . . . 122
4.5.1 Optimal Asset Allocation and Consumption . . . 122
4.5.2 (s, S) Inventory Policy Evaluation . . . 125
4.6 Conclusion . . . 127

Bibliography 129

List of Tables

2.1 Average computation time (ms) (including simulation) over 100 independent macro-replications. 86

3.1 Put option price estimates in Experiment 1 ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation). 105

3.2 Estimations in Experiment 2 ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation). 107

3.3 Estimation results in Experiment 3 with 10^6 independent runs ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation, $b = 2.3$ for Setting (ii)). 108

List of Figures

2.1 The numbers above the markers represent the indices of the alternatives. The S-indices smooth the sample means and correct the errors between the relative order of the largest two alternatives. 19
2.2 PCS and EOC results with standard error bounds obtained by 10000 independent runs (40 macro-replications, 250 micro-replications) with an unaligned prior similarity matrix. 78
2.3 Response curves for the dose selection problem. $\chi_1(v) = \chi(v; [2, 80, 0.3, 600, 4])$, $\chi_2(v) = \chi(v; [2, 100, 0.2, 400, 5])$. 79
2.4 PCS and EOC results with standard error bounds for Experiment 2 obtained by 10000 independent runs (40 macro-replications, 250 micro-replications) 80
2.5 PCS and EOC results with standard error bounds for Experiment 3 obtained by 10000 independent runs (40 macro-replications, 250 micro-replications) 81
2.6 PCS and EOC results with standard error bounds for Experiment 4 based on 10000 independent replications (40 macro-replications, 250 micro-replications.) . 83
2.7 Servo System. Average PCS and EOC results for different values of λ using SIGD, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors. 84
2.8 Servo System. Average PCS and EOC results for different values of λ using SIMA, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors. 84
2.9 M/G/1 queue selection problem. Average PCS and EOC results for different values of κ^{tol} using SIGD, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors. 85
2.10 M/G/1 queues selection problem. Average PCS and EOC results for different values of κ^{tol} using SIMA, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors. 85

3.1 Plot of $g_2(u)$ with stationary points. 100

4.1 Boxplot of $F_t(\cdot)$, distribution of the state variable X_t. Parameters used: $T = 0.05$, $\Delta t = 0.01$, $x_0 = 10$, $\mu = 1$, $\sigma = 2$, $r = 0.8$, $\rho = 1.2$, $\eta = 0.4$, $\epsilon = 0.01$. 125

4.2 The left and right plots represent the logarithm of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Results are reported every 10 epochs. For FD, $M_0 = 200$, $M = 1$, $N_0 = 200$. For GLR, $M_0 = \tilde{M}_0 = 200$, $M = \tilde{M} = 1$, $N_0 = 200$. 126
4.3 (s, S) policy evaluation. The left and right plots represent the logarithm of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Parameters used: $T = 5$, $\gamma = 0.9$, $\lambda = 0.2$, $c_h = 3$, $c_s = 5$, $c = 3$, $K = 5$. Results are reported every 10 epochs. For FD, $M_0 = 200$, $M = 1$, $N_0 = 200$. 127

Chapter 1: Introduction

1.1 Ranking and Selection

Ranking and Selection (R&S) is the process of selecting the best-performing solution (or alternative) from a finite set of candidates, where the performance can be estimated using simulations. Many real-world problems involve random factors that lead to noisy observations during the simulation process. For example, in a queueing system, the arrival process and service times are stochastic, resulting in random waiting times. To estimate the mean waiting time, we need to run the simulation many times. If we wish to select a system with the smallest mean waiting time among different queues, it would be expensive to simulate each of them sufficiently many times to obtain accurate estimates. Therefore, the goal of R&S is to identify the optimal solution using as few simulation replications as possible. To achieve this, algorithms have been developed to determine the strategy of allocating simulation budgets to each alternative. Here, budget refers to a collection of simulation replications.

Mathematically, assume there are k alternatives with true performance values μ_i, $i = 1, \ldots, k$, and the optimal solution i^* is defined as $i^* = \arg\max_{i=1,\ldots,k} \mu_i$. One commonly used estimator of the performance is the sample average, which we denote by $\bar{y}_i$. Chen et al. [1] developed an Optimal Computing Budget Allocation (OCBA) by maximizing the probability of

correct selection (PCS), that is

$$\text{PCS} = P\left(\arg\max_{i=1,\ldots,k} \bar{y}_i = i^*\right)$$

with a constraint of a fixed budget, $\sum_{i=1}^{k} N_i = M$, where M is the total budget and N_i is the budget for each alternative i.

Another line of research employs a Bayesian framework, where the estimator is the posterior mean. Frazier et al. [2] designed a knowledge-gradient (KG) allocation policy by maximizing the expected value of the chosen alternative, that is,

$$\max_{i=1,\ldots,k} \mathbb{E}\left[y_i^M\right],$$

where y_i^M is the posterior mean for alternative i where the used budget is M.

These and many other papers assume that information is independent across alternatives. However, there are circumstances where experiments with one alternative can provide information about others. For instance, systems that have similar structures could have similar performance values. In such cases, leveraging similarity information could help faster identification of the optimal solution. Assuming that correlations exist between the performances of alternatives, Frazier et al. [3] further extended KG for correlated normal beliefs.

In Chapter 2, which is based on Zhou et al. [4], we investigate the R&S problem where we exploit similarity information differently by using a new estimator, the similarity selection index [5], which combines similarities with the standard sample mean or posterior mean estimator, and based on this selection criterion, we propose two new allocation algorithms. In addition to this,

we also develop a practical approach for learning similarities between alternatives.

1.2 Estimating a Conditional Expectation With the Generalized Likelihood Ratio Method

Many problems in forecasting stochastic systems can be formulated as the estimation of expected performance in the future, given information collected up to the current period. One common way to estimate the conditional expectation is to simulate sufficient sample paths starting from the given state and compute the sample average of the target metric. However, in applications such as portfolio risk management, where we want to learn the expected performances from various starting points, this method would require a substantial number of simulations. To address this issue, in Chapter 3, which is based on Zhou et al. [6], we propose a new approach that represents the conditional expectation as a ratio of two stochastic derivatives under appropriate assumptions.

Stochastic derivatives can be estimated using various techniques, such as the finite difference (FD) method, infinitesimal perturbation analysis (IPA) [7] and the likelihood ratio (LR) method [8]. However, all of them have some undesirable properties that would not be suitable for our purpose. Therefore, in our work, we apply a new stochastic gradient technique, the generalized likelihood ratio (GLR) method [9], to obtain estimators of the stochastic derivatives. A variance reduction method is also proposed in Chapter 3.

1.3 Policy Evaluation With Stochastic Gradient Estimation Techniques

Reinforcement learning (RL) techniques have been widely successful in decision-making problems, including optimal asset allocation [10], inventory management in supply chains [11], and option pricing and hedging [12]. These problems are formulated using finite-horizon stochastic processes, and the goal is to find the optimal action policy that can maximize a target objective, such as the cumulative return of assets. Popular RL techniques include Q-learning [13] and policy gradient methods [14, 15].

In Chapter 4, which is based on Zhou et al. [16], we investigate the problem of policy evaluation, which involves estimating the value function for a fixed action policy. Once the value function is estimated, we can iteratively improve the policy based on the estimated values. By applying the approach developed in Chapter 3, which introduces new estimators for the value function based on conditional expectation, we propose new algorithms for policy evaluation. We compare our methods with widely-used approaches such as the Monte Carlo (MC) method and the temporal difference (TD) method [17].

Chapter 2: Sequential Learning With a Similarity Selection Index

2.1 Introduction

In the ranking and selection problem [18], there is a finite set of feasible solutions ("systems" or "alternatives") with unknown performances that must be estimated from noisy observations, e.g., from expensive stochastic simulation replications. A single replication only provides information about a single alternative, creating a tradeoff: spending resources, such as simulation time, to learn about one alternative means that less information can be collected about the others. Much of the research in this area focuses on designing algorithms that sequentially allocate replications to alternatives (adapting to the results of past assignments) for the purpose of identifying the highest-valued alternative as quickly as possible.

This problem can be approached using a wide variety of algorithmic concepts, including indifference-zone selection [19], value of information [2, 20], optimal computing budget allocation [1, 21], procedures based on stopping boundaries [22, 23], and optimal sampling laws [24, 25]. These and many other papers assume that information is completely independent across alternatives, i.e., a replication with one alternative provides no information about any others. Often, however, it is possible to exploit structural similarities or differences between alternatives to solve the problem more efficiently. For example, alternatives may have known common attributes that influence their performance. We may also have past performance data about other

alternatives that come from the same context, giving us some prior knowledge about the problem at hand. Two examples of such settings are:

- In the simulation of automatic control systems, performance is affected by mechanical settings that are known to users before simulation. These settings may have been tested before for other systems. Thus, when optimizing a new system for which no test results are available, we may use these past data to help the simulation process [26].

- In a clinical trial, an important concern is estimation of the dose-response relationship, which is influenced by factors such as age and gender. When determining the optimal dose for a new patient group (e.g., children), for which no data have yet been collected, we may use test results from other groups (e.g., adults) as prior information since the dose-response curves for both groups may follow similar patterns [27].

In these and other settings, past data provides prior information about new alternatives of interest – not only about their performance but also about similarities between these performance values. As a consequence, when we conduct a new replication on one alternative, the similarity information allows us to learn about other, similar alternatives. This can significantly reduce the expenditure of resources needed to identify the best alternative reliably, greatly improving the ability to solve large instances where exhaustive simulation may not be practical. The literature has recognized this advantage of similarity information and has generally handled it in two ways. In situations where alternatives are naturally described by certain attributes or features, one can relate their values to these features using linear regression; see [28] or [29] for examples of such parametric models, and [30] for a setting where the quadratic structure is imposed on the performance values. If no linear or quadratic structure is available, one can instead

quantify pairwise similarities in the form of a matrix. This approach is often adopted by Bayesian optimization methods, e.g., by [3] for ranking and selection or [31] for continuous simulation optimization. In these methods, the matrix is included in the decision-maker's Bayesian belief about the unknown values and is updated after each replication; furthermore, these "correlated beliefs" are also used inside the allocation procedure, leading to a significant increase in overall computational cost.

We investigate a different way to model similarities inside ranking and selection, which we call the *similarity selection index*, or *S-index* for short. This approach is motivated by spectral clustering techniques [32], and was first proposed by Sun et al. [5], where they introduced the concept under the name "spectral index". We use the "similarity selection index" to avoid confusion with the spectrum of a matrix. Sun et al. [5] provided some preliminary theoretical results and numerical experiments in support of the S-index as a selection criterion, but did not develop any new budget allocation algorithms.

In S-index selection, similarity information is represented by weights assigned to edges in a graph whose vertices are the alternatives. The linear transformation then adjusts the estimated values of these alternatives based on the weights. Thus, the weights play a similar role to that of correlated beliefs in Bayesian methods. However, the main distinguishing characteristic of this approach, relative to the simulation literature, is that similarity information is utilized only inside the selection criterion used to return the best alternative after the replications have concluded, rather than inside the statistical model used to learn the unknown values. The learning process still assumes that the values of individual alternatives are estimated independently, as in traditional ranking and selection, but in the selection step, each alternative is evaluated using the S-index rather than the sample mean. In this way, we fully use the power of similarity information while

reducing the computational cost of the allocation procedure.

Since the S-index is used for selection, rather than allocation, it can potentially be used in conjunction with any existing allocation method, which is how it was implemented by Sun et al. [5]. However, allocations should perform better if they are tailored to the new selection criterion. We propose two such allocation methods in this chapter. The first method calculates a budget allocation by optimizing an approximation of the probability of correct selection (PCS) from a geometric perspective. This approximate problem, however, is shown to yield an asymptotically optimal solution, because it can be shown to be equivalent to the well-established problem of optimizing large deviations rates of PCS, first posed by Glynn and Juneja [24].

The optimal budget allocation is a function of the unknown values, and cannot be directly implemented; one could approximate it by solving the optimization problem with plug-in estimators, but this is computationally cumbersome. We address this difficulty by developing a sequential implementation that provably converges to the optimal allocation, without tuning, and without exactly solving a convex optimization problem in every step. The design of this algorithm is quite distinct from existing sequential procedures, and leverages techniques from nonlinear optimization in a novel way. We also develop a second heuristic approach, based on value of information, which is simple to compute and performs remarkably well in numerical experiments.

The advantages of the S-index depend on the availability of accurate similarity information. To make this approach more robust against misspecification, we also develop a procedure for learning prior similarities from existing data and sequentially updating them as new information is collected. It should be noted that, while correlated Bayesian beliefs have to deal with the same concern, in the simulation literature only [33] and [34] have dealt with the problem of learning similarity structures to a significant degree.

In sum, our work makes the following contributions:

- We formulate an optimal budget allocation problem using a geometric approximation of PCS, and show that the solution to this problem optimizes the asymptotic convergence rate of PCS characterized using large deviations theory. There is a separate stream of literature focusing on large deviations-based allocations, so this result is of stand-alone interest.

- We propose a sequential implementation, leveraging the idea of reduced gradient methods in nonlinear optimization, and prove that this method asymptotically learns the optimal budget allocation with probability 1. We also propose an additional sequential heuristic that is computationally efficient and easy to implement.

- We develop a dynamic update procedure that learns a prior similarity graph from data and sequentially updates it as more samples are collected, significantly strengthening the applicability of our proposed algorithms.

- Numerical experiments comparing our proposed methods to each other and also to several existing benchmarks indicate that the new algorithms perform well. Furthermore, using the S-index for selection can also improve the performance of existing allocation methods.

This chapter is organized as follows. Section 2.2 introduces the S-index and a class of similarity structures in which consistent selection can be guaranteed. Section 2.3 defines the notion of an optimal budget allocation under S-index selection, and presents a computationally efficient algorithm that can be guaranteed to learn that allocation asymptotically. Section 2.4 presents a second approach based on Bayesian value of information. Section 2.5 describes one way in which similarity structures can be updated over time as new information is acquired.

Section 2.6 considers the case where there exist alternatives sharing the same value. Numerical experiments are given in Section 2.7.

2.2 The S-index: Definition and Properties

We assume there are k alternatives with unknown values $\mu_1 > \mu_2 > \ldots > \mu_k$. For simplicity, we will develop the main concepts of this chapter assuming that all k values are distinct. However, there is no loss of generality in this assumption, and Section 2.6 gives some additional discussion for the case where multiple alternatives may have equal values. We model similarity information using a weighted graph representation $G = (V, S)$, where each alternative is viewed as a vertex $i \in V$. Let S be the similarity matrix, with $S_{i,j} = S_{j,i} \geq 0$ being the weight assigned to the edge between vertices $i, j \in V$.

Suppose now that the values μ_i are unknown, but we have access to estimates y_i of μ_i. In ranking and selection, y_i is usually the sample mean of all replications conducted with alternative i (or the posterior mean, in a Bayesian framework). Now, consider the optimization problem

$$z = \arg\min_{u \in \mathbb{R}^k} \sum_{i=1}^{k} (u_i - y_i)^2 + \frac{\lambda}{2} \sum_{1 \leq i,j \leq k} S_{i,j}(u_i - u_j)^2. \tag{2.1}$$

The quantity z_i is called the *S-index* of alternative i. If $\lambda = 0$, we will have $z = y$, but when $\lambda > 0$ the solution of (2.1) will be regularized by the similarity information; consequently, two alternatives with larger similarity will also have similar S-indices. The choice of λ, which controls the relative importance of the similarity information, is left up to the user. We can rewrite (2.1)

in matrix notation as

$$z = \arg\min_{u \in \mathbb{R}^k} (u - y)^T (u - y) + \lambda u^T L u, \tag{2.2}$$

where $L = L(S) = D - S$ and D is a diagonal matrix with $D_{i,i} = \sum_{j=1}^{k} S_{i,j}$. For simplicity, we let $S_{i,i} = 0$, since the diagonal entries of S do not affect L. The solution to (2.2) can be expressed in closed form as

$$z = (I + \lambda L)^{-1} y, \tag{2.3}$$

which allows us to compute the S-indices efficiently, given estimates y and a similarity matrix S. From [32], it can be seen that the matrix $I + \lambda L$ is symmetric positive definite (thus invertible) with minimum eigenvalue 1 and corresponding eigenvector $\mathbf{1}$. Sun et al. [5] discusses alternate ways to compute (2.3) that do not require matrix inversion.

In much of what follows, we will make use of several properties of the inverse $Q = Q(S; \lambda) = (I + \lambda L(S))^{-1}$, which are stated below. These properties hold for all similarity graphs.

Theorem 2.1. *For a graph G with $S_{i,j} \geq 0$ and $S_{i,j} = S_{j,i}$, $Q = (I + \lambda(D - S))^{-1}$ has the following properties:*

(1) Q is symmetric positive definite, having largest eigenvalue 1 with eigenvector $\mathbf{1}$.

(2) $Q_{i,j} \geq 0$, $\forall i, j = 1, ..., k$, i.e., every entry of matrix Q is non-negative.

(3) $\forall i = 1, ..., k$, $Q_{i,i} > \max_{j \neq i} Q_{i,j}$, i.e., the diagonal entry is the largest entry in both its

column and its row.

Proof. (1) can be directly concluded from the fact that the inverse of Q, $I + \lambda L$, is positive semidefinite and has largest eigenvalue 1 with eigenvector $\mathbf{1}$.

To show (2), we recall that Sun et al. [5] showed that the linear system $(I + \lambda L)z = y$ can be solved using an iterative algorithm

$$z^{(0)} = y, \quad z^{(t)} = z^{(t-1)} + a\left(y - (I + \lambda L)z^{(t-1)}\right), \quad t \geq 1, \tag{2.4}$$

which is guaranteed to converge to the true solution if $a > 0$ is sufficiently small. For any i, (2.4) can be rewritten as

$$z_i^{(t)} = ay_i + \left(1 - a - a\sum_{j \neq i} S_{i,j}\right) z_i^{(t-1)} + a\lambda \sum_{j \neq i} S_{i,j} z_j^{(t-1)}. \tag{2.5}$$

Consequently, when a is sufficiently small, $y_i \geq 0$ implies $z_i^{(t)} \geq 0$, $i = 1, ..., k$. This implies that every entry of Q is non-negative.

Now we show part (3). Let Q_l denote the lth column of $Q = (I + \lambda L)^{-1}$, then we have $(I + \lambda L)Q_l = e_l$, where $e_l = [0, ..., 0, 1, 0, ..., 0]^T$, 1 is the lth element. From equation (2.2), we know Q_l is the solution to the following optimization problem:

$$\arg\min_{u \in \mathbb{R}^k} \sum_{i \neq l}^{k} u_i^2 + (u_l - 1)^2 + \frac{\lambda}{2} \sum_{1 \leq i,j \leq k} S_{i,j}(u_i - u_j)^2. \tag{2.6}$$

Then we prove the results by contradiction. Assume the optimal solution is u^* and define $\Omega_1 =$

$\{i : u_i^* > u_l^*\}$. Suppose $u_l^* \neq \max_i u_i^*$, i.e., $\Omega_1 \neq \emptyset$. Now we construct $\bar{u}$ by

$$\bar{u}_i = \begin{cases} u_i^*, & i \notin \Omega_1 \\ \\ u_l^*, & i \in \Omega_1 \end{cases}.$$

Denote the value of the objective in (2.6) by $\psi(u)$ for a given u. Then,

$$\psi(\bar{u}) = \sum_{i \neq l}^{k} \bar{u}_i^2 + (\bar{u}_l - 1)^2 + \frac{\lambda}{2} \sum_{1 \leq i,j \leq k} S_{i,j}(\bar{u}_i - \bar{u}_j)^2 \tag{2.7}$$

$$< \sum_{i \neq l}^{k} (u_i^*)^2 + (u_l^* - 1)^2 + \frac{\lambda}{2} \sum_{1 \leq i,j \leq k} S_{i,j}(\bar{u}_i - \bar{u}_j)^2. \tag{2.8}$$

Now we discuss the term $S_{i,j}(\bar{u}_i - \bar{u}_j)^2$ for four cases.

1) for $i \notin \Omega_1$ and $j \in \Omega_1$, we have $u_i^* \leq u_l^* < u_j^*$, therefore, $(\bar{u}_i - \bar{u}_j)^2 = (u_i^* - u_l^*)^2 < (u_i^* - u_j^*)^2$;

2) for $i \in \Omega_1$ and $j \notin \Omega_1$, same as case (1);

3) for $i \notin \Omega_1$ and $j \notin \Omega_1$, $(\bar{u}_i - \bar{u}_j)^2 = (u_i^* - u_j^*)^2$;

4) for $i \in \Omega_1$ and $j \in \Omega_1$, $(\bar{u}_i - \bar{u}_j)^2 = (u_l^* - u_l^*)^2 = 0 \leq (u_i^* - u_j^*)^2$.

In summary, we have

$$\psi(\bar{u}) < \sum_{i \neq l}^{k} (u_i^*)^2 + (u_l^* - 1)^2 + \frac{\lambda}{2} \sum_{1 \leq i,j \leq k} S_{i,j}(u_i^* - u_j^*)^2 = \psi(u^*),$$

which contradicts the assumption that u^* is optimal. Therefore, $u_l^* = \max_j u_j^*$. We further prove that $u_l^* > \max_{i \neq l} u_i^*$. From (1) and (2) in this theorem, we know that $u_i^* \in [0, 1]$ and u^* can't be

a multiple of $\mathbf{1}$. Therefore, there must be a positive gap between u_l^* and some other component of u^*. Let $\Omega_2 = \{i \neq l : u_i^* = u_l^*\}$ and suppose $\Omega_2 \neq \emptyset$. Let ε be a positive number less than $\min\{u_l^* - \max_{j\in\Omega_2^c\setminus\{l\}} u_j^*, \min_{i\in\Omega_2}\{\frac{2u_l^*}{1+\lambda S_{l,i}}\}\}$. Construct $\bar{u}$ by

$$\bar{u}_i = \begin{cases} u_i^*, & i \notin \Omega_2 \\ u_l^* - \varepsilon, & i \in \Omega_2 \end{cases}.$$

Notice that for $i \in \Omega_2$ and $j \in \Omega_2^c \setminus \{l\}$,

$$(\bar{u}_i - \bar{u}_j)^2 < (u_i^* - u_j^*)^2.$$

Then we have

$$\begin{aligned} \psi(\bar{u}) - \psi(u^*) &< \sum_{i\in\Omega_2} \bar{u}_i^2 - \sum_{i\in\Omega_2} {u_i^*}^2 + \lambda \sum_{i\in\Omega_2} S_{l,i}(u_l^* - \bar{u}_i)^2 \\ &= \sum_{i\in\Omega_2} \left((u_l^* - \varepsilon)^2 - (u_l^*)^2 + \lambda S_{l,i}(u_l^* - (u_l^* - \varepsilon))^2\right) \\ &= \sum_{i\in\Omega_2} \varepsilon\left((1 + \lambda S_{l,i})\varepsilon - 2u_l^*\right) < 0, \end{aligned}$$

which leads to a contradiction. □

Classical ranking and selection uses $\arg\max_i y_i$ as the selection criterion, i.e., it predicts the alternative with the highest sample mean as being the best. Given enough samples, $y_i \to \mu_i$ and so the true best alternative will eventually be discovered under this criterion. We propose to use S-indices instead of sample means, so that $\arg\max_i z_i$ becomes the selection rule. Then the

first important question is whether the true best alternative can still be recovered using the linear transformation (2.3). To that end, [5] introduced a particular class of similarity structures under which this can be guaranteed. The following is a slightly relaxed version of Definition 1 in [5].

Definition 2.1. *A graph G is aligned if for every fixed $i = 1, ..., k$, the similarities $\{S_{i,j}\}$ satisfy $S_{i,j} \leq S_{i,m}$ for any $j < m < i$ or $j > m > i$.*

We can then show that if the original estimates are sufficiently accurate to recover the correct ordering of the alternatives, this ordering will also be preserved under S-indices. Consequently, selecting alternative $\arg\max_i z_i$ will still lead us to the true best alternative.

Theorem 2.2. *[Theorem 3 in Sun et al. [5]] For an aligned graph G and estimates $y_1 \geq y_2 \geq \cdots \geq y_k$, the S-index defined by $z = Qy = (I + \lambda L)^{-1} y$ preserves the ordering, i.e., $z_1 \geq z_2 \geq \cdots \geq z_k$.*

Here, we give a corrected version of the proof of Theorem 3 in Sun et al. [5], which had $D_{i,i} = \sum_{j \neq i} S_{i,j}$, and $S_{i,i}$ is not included in the sum. In their proof, they manipulated $S_{i,i}$ such that $D_{i,i} = D_{j,j}$, $i \neq j$, which is not consistent with their setting. We follow the same general arguments, but correct the technical issues.

Proof. At iteration 0, let $z^{(0)} = y$, we have $z_1^{(0)} \geq z_2^{(0)} \geq \cdots \geq z_k^{(0)}$. At iteration $t > 0$, assume $z_1^{(t-1)} \geq z_2^{(t-1)} \geq \cdots \geq z_k^{(t-1)}$ holds, using the iterative algorithm in (2.4), we need to show that $z_i^{(t)} \geq z_{i+1}^{(t)}$ holds for any $1 \leq i \leq k-1$. Since

$$z_i^{(t)} = a y_i + (1-a) z_i^{(t-1)} + a\lambda \sum_{j \neq i} S_{i,j} \left(z_j^{(t-1)} - z_i^{(t-1)} \right)$$

$$z_{i+1}^{(t)} = a y_{i+1} + (1-a) z_{i+1}^{(t-1)} + a\lambda \sum_{j \neq i+1} S_{i+1,j} \left(z_j^{(t-1)} - z_{i+1}^{(t-1)} \right),$$

taking the difference, we have:

$$\begin{aligned} z_i^{(t)} - z_{i+1}^{(t)} =& a(y_i - y_{i+1}) + (1-a)\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right) \\ &+ a\lambda\left(\sum_{j\neq i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) - \sum_{j\neq i+1} S_{i+1,j}\left(z_j^{(t-1)} - z_{i+1}^{(t-1)}\right)\right). \end{aligned}$$

Since

$$\begin{aligned} &\sum_{j\neq i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) - \sum_{j\neq i+1} S_{i+1,j}\left(z_j^{(t-1)} - z_{i+1}^{(t-1)}\right) \\ =& \sum_{j<i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) - \sum_{j<i} S_{i+1,j}\left(z_j^{(t-1)} - z_{i+1}^{(t-1)}\right) + \sum_{j>i+1} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) \\ &- \sum_{j>i+1} S_{i+1,j}\left(z_j^{(t-1)} - z_{i+1}^{(t-1)}\right) - 2S_{i,i+1}\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right) \\ =& \sum_{j<i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) - \sum_{j<i} S_{i+1,j}\left(z_j^{(t-1)} - z_{i+1}^{(t-1)}\right) + \sum_{j>i+1} S_{i+1,j}\left(z_{i+1}^{(t-1)} - z_j^{(t-1)}\right) \\ &- \sum_{j>i+1} S_{i,j}\left(z_i^{(t-1)} - z_j^{(t-1)}\right) - 2S_{i,i+1}\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right) \\ =& \sum_{j<i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right) - \sum_{j<i} S_{i+1,j}\left(z_j^{(t-1)} - z_i^{(t-1)} + z_i^{(t-1)} - z_{i+1}^{(t-1)}\right) + \sum_{j>i+1} S_{i+1,j}\Big(z_{i+1}^{(t-1)} \\ &- z_j^{(t-1)}\Big) - \sum_{j>i+1} S_{i,j}\left(z_i^{(t-1)} - z_{i+1}^{(t-1)} + z_{i+1}^{(t-1)} - z_j^{(t-1)}\right) - 2S_{i,i+1}\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right) \\ =& \sum_{j<i} (S_{i,j} - S_{i+1,j})\left(z_j^{(t-1)} - z_i^{(t-1)}\right) + \sum_{j>i+1} (S_{i+1,j} - S_{i,j})\left(z_{i+1}^{(t-1)} - z_j^{(t-1)}\right) \\ &- \left(\sum_{j<i} S_{i+1,j} + \sum_{j>i+1} S_{i,j} + 2S_{i,i+1}\right)\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right), \end{aligned}$$

we have

$$z_i^{(t)} - z_{i+1}^{(t)} = a(y_i - y_{i+1}) + \left[1 - a - a\lambda\left(\sum_{j<i} S_{i+1,j} + \sum_{j>i+1} S_{i,j} + 2S_{i,i+1}\right)\right]\left(z_i^{(t-1)} - z_{i+1}^{(t-1)}\right)$$

$$+a\lambda\left[\sum_{j<i}\left(S_{i,j}-S_{i+1,j}\right)\left(z_j^{(t-1)}-z_i^{(t-1)}\right)+\sum_{j>i+1}\left(S_{i+1,j}-S_{i,j}\right)\left(z_{i+1}^{(t-1)}-z_j^{(t-1)}\right)\right].$$

Since $S_{i,j} \geq S_{i+1,j}$, $z_j^{(t-1)} \geq z_i^{(t-1)}$ when $j < i$, and $S_{i+1,j} \geq S_{i,j}$, $z_{i+1}^{(t-1)} \geq z_j^{(t-1)}$, when $j > i+1$,

$$\sum_{j<i}\left(S_{i,j}-S_{i+1,j}\right)\left(z_j^{(t-1)}-z_i^{(t-1)}\right)+\sum_{j>i+1}\left(S_{i+1,j}-S_{i,j}\right)\left(z_{i+1}^{(t-1)}-z_j^{(t-1)}\right)\geq 0.$$

Also, $y_i \geq y_{i+1}$ and $z_i^{(t-1)} \geq z_{i+1}^{(t-1)}$, if a is sufficiently small, we have

$$z_i^{(t)} \geq z_{i+1}^{(t)},$$

which completes the proof. □

In the rest of the chapter, y (and therefore z as well) will be a random vector, whose distribution depends on the number and outcomes of simulation replications. The performance of a selection criterion based on such a vector can be measured in terms of the probability of correct selection (PCS), i.e., the probability that the first component of the vector is also the highest-valued. For example, under S-index selection, the PCS can be written as

$$\text{PCS}(z) = P\left(z_1 - z_i \geq 0, i \neq 1\right),$$

and analogously PCS (y) is the PCS under the original estimates y. If the similarity graph is very informative about the relationships between the values of alternatives, it is possible to show that $\text{PCS}(z) \geq \text{PCS}(y)$, though this cannot be guaranteed in general. Theorem 2.3 gives one example of a graph structure that has this property. Although most of our analysis, from Section 2.3

onwards, assumes that y and z are normally distributed, Theorem 2.3 does not require specific distributional assumptions, but rather arises from the structure of the similarity measure. The result also does not require the graph to be aligned (and does not hold for all aligned graphs).

Theorem 2.3. *If $S_{1,j} \leq S_{i,j}$ for any $i \neq j, i, j \neq 1$, for any allocation policy, PCS$(z) \geq$ PCS(y).*

Proof. We use the same induction approach as in the proof of Theorem 2.2 by assuming $z^{(0)} = y$. Suppose that $z_1^{(0)} \geq z_i^{(0)}$ for any $i = 1, \ldots, k$. Assume that at iteration $t-1$, $z_1^{(t-1)} \geq z_i^{(t-1)}$. Then at iteration t, we have

$$
\begin{aligned}
z_1^{(t)} - z_i^{(t)} =& a(y_1 - y_i) + (1-a)\left(z_1^{(t-1)} - z_i^{(t-1)}\right) \\
&+ a\lambda\left(\sum_{j\neq 1} S_{1,j}\left(z_j^{(t-1)} - z_1^{(t-1)}\right) - \sum_{j\neq i} S_{i,j}\left(z_j^{(t-1)} - z_i^{(t-1)}\right)\right) \\
=& a(y_1 - y_i) + (1-a)\left(z_1^{(t-1)} - z_i^{(t-1)}\right) \\
&+ a\lambda\left(\sum_{j\neq 1} S_{1,j}\left(z_j^{(t-1)} - z_1^{(t-1)}\right) - \sum_{j\neq i} S_{i,j}\left(\left(z_1^{(t-1)} - z_i^{(t-1)}\right) + \left(z_j^{(t-1)} - z_1^{(t-1)}\right)\right)\right) \\
=& a(y_1 - y_i) + (1 - a - a\lambda\sum_{j\neq i} S_{i,j})\left(z_1^{(t-1)} - z_i^{(t-1)}\right) \\
&+ \lambda\left(\sum_{j\neq 1,i} (S_{i,j} - S_{1,j})\left(z_1^{(t-1)} - z_j^{(t-1)}\right)\right).
\end{aligned}
$$

Since $S_{i,j} \geq S_{1,j}$, $z_1^{(t-1)} \geq z_j^{(t-1)}$, if a is sufficiently small, we have

$$
z_1^{(t)} \geq z_i^{(t)}, \quad i = 1, \ldots, k.
$$

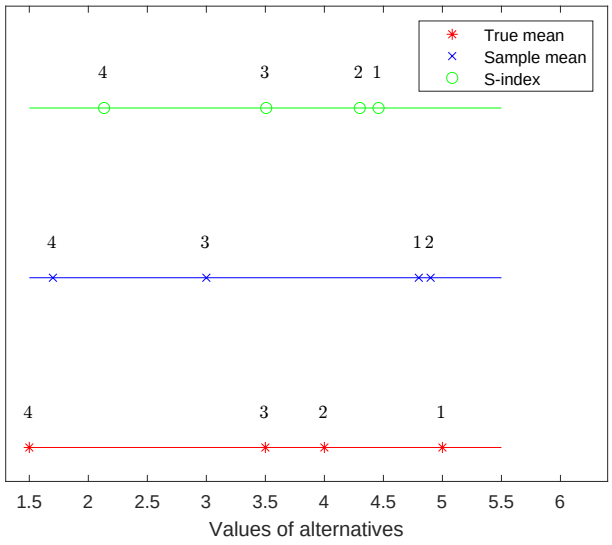

Figure 2.1: The numbers above the markers represent the indices of the alternatives. The S-indices smooth the sample means and correct the errors between the relative order of the largest two alternatives.

Thus, we have shown

$$\{y_1 \geq y_i,\, i = 1, \ldots, k\} \subseteq \{z_1 \geq z_i,\, i = 1, \ldots, k\}\,,$$

which completes the proof. □

Intuitively, if the best alternative is "separated" from the others by the similarity measure, PCS will be improved because our estimates of μ_i for suboptimal i will smooth out each other, while exerting little impact on our estimate of μ_1. This may give some intuition why, even if y_i is a minimum-variance unbiased estimator of μ_i, the S-indices z may nonetheless be more accurate than y in identifying the highest-valued alternative. A simple illustration is given in Figure 2.1, where the sample means suggest a wrong best alternative, while the S-indices give the correct best alternative by smoothing the sample means.

The preceding results treat the similarity graph as given. In practice, if there is reason to

believe that alternatives have common attributes that strongly influence their values, one might construct the graph from past data on similar alternatives (see Section 2.7 for concrete examples). Such a graph generally will not be aligned, nor will it satisfy the conditions of Theorem 2.3. Nonetheless, it may still contain helpful information that can improve the performance of ranking and selection methods, especially in the early stages. Later, in Section 2.5, we will present a practical approach for updating the similarity graph over time, in a manner that eventually makes it aligned.

In this methodology, the similarity measure is separate from the statistical mechanism used to estimate the unknown values. The computation of y itself need not use the similarity measure at all; the latter is only used to transform y in the final selection step. In this way, even though the concept of similarity between alternatives is used both in S-index selection and in correlated Bayesian belief models, they use it in different ways. In Bayesian methods, covariance matrices are built into the statistical model used to estimate values. Prior covariances are typically obtained from spatial similarities that are not directly connected to the performance values, and posterior covariances decay over time (converge to zero) as more information is collected. On the other hand, S-index selection views the similarity structure (represented by an aligned graph) as a property of the true values. In the frequentist view, these values are fixed and do not change over time. Therefore, the similarities between them do not vanish either, and our selection decision should be able to benefit from them regardless of how the samples were collected. Thus, one can view S-indices as a purely frequentist framework for modeling correlated beliefs, but the similarity graph is constructed in a different way than in Bayesian methods; see Section 2.7 for practical examples.

2.3 Optimal Budget Allocation Under S-index Selection

In and of itself, the S-index is only used to select an alternative given estimated values y_i. We may compute S-indices without any assumptions about how these estimates were obtained. Thus, potentially, S-index selection is compatible with any existing procedure for allocating the budget. At the same time, it is also possible to define a strong notion of optimality for a budget allocation in the presence of S-indices; the principle behind this notion is similar to the optimal computing budget allocation methodology of [35], but the optimal allocation itself is influenced by the similarity structure. It is not possible to implement this "ideal" allocation directly, because it also depends on the unknown values μ, but we will design efficient sequential procedures that can be guaranteed to converge to it asymptotically.

In the following, we often denote the index of the best alternative by i^* instead of 1 to make the eventual implementation clearer. We suppose that the estimates y are constructed using simple frequentist statistics. Let $w_1, ..., w_k$ be a vector of i.i.d. standard normal random variables, and assume that

$$y = \mu + \Lambda w, \qquad \Lambda = \text{diag}\left(\frac{\sigma_1}{\sqrt{N_1}}, ..., \frac{\sigma_k}{\sqrt{N_k}}\right),$$

where $\sigma_1, ..., \sigma_k$ are fixed positive constants and $N_1, ..., N_k$ are positive integers, i.e., the simulation output is assumed to be normally distributed with standard deviation σ_i, and N_i being the number of replications allocated to alternative i.

2.3.1 Geometric Approximation of the Probability of Correct Selection

From (2.3), we have $z = Q\mu + Q\Lambda w$, so $z \sim \mathcal{N}(Q\mu, Q\Lambda^2 Q^T)$. Let $P_{i,j} = Q_{i^*,j} - Q_{i,j}$, and define $m = P\mu$ and $\Sigma = P\Lambda^2 P^T$. Then, the joint distribution of the vector $\{z_{i^*} - z_i, i \neq i^*\}$ is $\mathcal{N}(m_{-i^*}, \Sigma_{-i^*})$, where the mean vector m_{-i^*} is obtained by deleting the i^*th element of m, and the covariance matrix Σ_{-i^*} is obtained by deleting the i^*th row and column of Σ.

Suppose P_{-i^*}, defined similarly, has full rank. By the Cholesky decomposition, we can represent $\Sigma = U^T U$, where $U = [U_{i,j}]_{k\times k}$ is an upper triangular matrix. We are interested in the probability that $\arg\max_i z_i = \arg\max_i \mu_i$, which means that the correct alternative is identified by using the S-index for selection. The corresponding PCS is given by

$$\text{PCS}(z) = \left(\frac{1}{\sqrt{2\pi}}\right)^{k-1} \int \cdots \int_{\{\sum_{j\neq i^*}^{i} U_{j,i} w_j \geq -m_i, i\neq i^*\}} \prod_{i\neq i^*} e^{\frac{-w_i^2}{2}} dw_i. \tag{2.9}$$

Since the terms in the integrand in (2.9) decay exponentially, the region near the origin dominates the value of the integral. Therefore, (2.9) can be approximately optimized by maximizing the volume of the hypersphere contained inside the domain of integration and centered at the origin [36]. This approximation methodology was also used in Peng et al. [37] for a traditional ranking and selection problem.

The distance from the origin to the hyperplane $\sum_{j\neq i^*}^{i} U_{j,i} w_j = -m_i$, $i \neq i^*$ is given by

$$\frac{m_i}{\sqrt{\sum_{j\neq i^*}^{i} U_{j,i}^2}} = \frac{m_i}{\sqrt{\Sigma_{i,i}}}, \tag{2.10}$$

where $\Sigma_{i,i} = \sum_{j=1}^{k} \frac{\sigma_j^2}{N_j} P_{i,j}^2$. To maximize the volume of the hypersphere, it is sufficient to solve

$$\min_{\substack{\sum_{i=1}^{k} N_i = M \\ N_j \geq 0}} \max_{i \neq i^*} \frac{1}{(P_i^T \mu)^2} \sum_{j=1}^{k} \frac{\sigma_j^2}{N_j} P_{i,j}^2, \tag{2.11}$$

which is equivalent to maximizing the smallest distance to any of the hyperplanes. The quantity M represents the total simulation budget; alternatively, one could set $M = 1$ and interpret N_j as the proportion of the budget to assign to alternative j. The objective in (2.11) can be linearized, giving rise to the equivalent problem

$$\begin{aligned} \text{minimize} \quad & \xi \\ \text{subject to} \quad & \sum_{j=1}^{k} \frac{a_{i,j}}{N_j} \leq \xi, \quad i \neq i^*, \\ & N_1 + N_2 + \cdots + N_k = M, \\ & N_j \geq 0 \quad j = 1, 2, \ldots, k, \end{aligned} \tag{2.12}$$

where $a_{i,j} = \frac{\sigma_j^2 P_{i,j}^2}{(P_i^T \mu)^2}$, $i \neq i^*$, $j = 1, ..., k$.

From the definition of a convex function [38], it is easy to see that the mapping $x \mapsto \sum_{j=1}^{k} \frac{a_j}{x_j}$, where $a_j \geq 0$, is convex on $\mathbb{R}_+^k$. If we allow N_j to be continuous, (2.12) will be a convex program, guaranteed to have a unique optimal solution. In general, this solution cannot be computed analytically, with the exception of the special case $k = 3$, treated in the following result.

Proposition 2.1. *Let $i_1, i_2 \neq i^*$ denote the two distinct suboptimal alternatives in $\{1, 2, 3\}$, and*

define $r(\gamma) = \sum_{j=1}^{3} \frac{a_{i_1,j} - a_{i_2,j}}{\sqrt{\gamma a_{i_1,j} + (1-\gamma) a_{i_2,j}}}$. *Let*

$$\gamma^* = \begin{cases} 1 & \text{if } r(0) > 0 \text{ and } r(1) > 0 \\ 0 & \text{if } r(0) < 0 \text{ and } r(1) < 0 \\ \gamma_0 & \text{otherwise,} \end{cases}$$

where γ_0 *is the solution of* $r(\gamma) = 0$*, which can be efficiently obtained using a bisection algorithm. Then for* $k = 3$*, the optimal solution of (2.12) is given by*

$$N_j^* = \frac{\sqrt{\gamma^* a_{i_1,j} + (1-\gamma^*) a_{i_2,j}}}{\sum_{j'=1}^{3} \sqrt{\gamma^* a_{i_1,j'} + (1-\gamma^*) a_{i_2,j'}}} M, \quad j = 1, 2, 3.$$

Proof. Here, we let $M = 1$. For a general M, we only need to scale the optimal proportion by M to get the allocation vector. Then, for $k = 3$, the dual problem of (2.11) can be written as

$$\begin{aligned} \max_{\gamma} \quad & \left(\sum_{j=1}^{k} \sqrt{\gamma a_{i_1,j} + (1-\gamma) a_{i_2,j}} \right)^2 \\ \text{subject to} \quad & 0 \le \gamma \le 1, \end{aligned} \tag{2.13}$$

where $i_1, i_2 \neq i^*$ denote the two distinct suboptimal alternatives. The detailed derivation of the dual problem can be found in the proof of Theorem 2.4. Denote by $g(\gamma)$ the objective function of (2.13). It is easy to see that g is concave. Taking the derivative, we have

$$g'(\gamma) = \left(\sum_{j=1}^{k} \sqrt{\gamma a_{i_1,j} + (1-\gamma) a_{i_2,j}} \right) \sum_{j=1}^{k} \left(\frac{a_{i_1,j} - a_{i_2,j}}{\sqrt{\gamma a_{i_1,j} + (1-\gamma) a_{i_2,j}}} \right). \tag{2.14}$$

The primal (2.11) is convex, and it is easy to see that Slater's condition (and thus, strong duality)

holds. Neglecting the factor $\sum_{j=1}^{k} \sqrt{\gamma a_{i_1,j} + (1-\gamma) a_{i_2,j}} > 0$ in (2.14), we obtain the function $r(\cdot)$ in the statement of the theorem. The results then follow from the properties of one-variable concave optimization problems on $[0, 1]$. □

Whether or not we are in this special case, however, the solution depends on the unknown values μ (as well as on σ, which may also be unknown), and thus cannot be implemented directly. Instead, one might replace the unknown parameters in (2.12) with plug-in estimators computed from past replications, and use the solution of the resulting approximate problem to allocate a portion of the budget. By iteratively re-solving (2.12) with updated estimators, one will eventually arrive at the true optimal solution. If $k > 3$, the computational cost will be fairly high, due to the need to repeatedly solve convex programs, but it is possible to do this in principle using a solver such as CVX [39].

Algorithm 2.1 gives one possible implementation, which we call SIOCBA (S-index optimal computing budget allocation), where e_i denotes a k-vector of zeroes with the ith coordinate equal to 1. For simplicity, we treat σ as being known; in practice, the variances can be estimated using sample variances and updated as more data are collected. In the nth stage of sampling, M^n is the total budget spent thus far, y^n is the current vector of sample means, with $z^n = Qy^n$ being the corresponding vector of S-indices. We denote by $i^{*,n} = \arg\max_j (Qy^n)_j$ the index of the alternative currently believed to be the best (note that S-indices are used to identify this alternative), and use the quantities $P_{i,j}^n = Q_{i^{*,n},j} - Q_{i,j}$ and $a_{i,j}^n = \frac{\sigma_j^2 \left(P_{i,j}^n\right)^2}{\left(\left(P_i^n\right)^T y^n\right)^2}$, for all $i \neq i^{*,n}$, in the computation. The inputs i^*, $a_{i,j}$ and M in (2.12) are replaced by $i^{*,n}$, $a_{i,j}^n$ and M^n, respectively. The next replication is then allocated to alternative $j^n = \arg\max_j \frac{N_j^{*,n}}{N_j^n}$, which favors those alternatives whose actual sample sizes N_j^n are small relative to the estimated optimal

Algorithm 2.1: S-index optimal computing budget allocation (SIOCBA).

Input: number of alternatives k, variances σ_i^2, total budget M, number of initial samples n_0, matrix Q.

Output: final selection $i^{*,M}$.

Initialization: $n = 0$, $N_j^n = n_0$, $j = 1, ..., k$, $M^0 = kn_0$.

Collect n_0 samples for each alternative and calculate sample means y^0.

while $M^0 + n < M$ **do**

$i^{*,n} \leftarrow \arg\max_j (Qy^n)_j$; `// find the best alternative using current S-index`

$P^n \leftarrow \text{kron}(Q(i^{*,n}, :), \text{ones}(k, 1))$; `// update the optimization problem`

$a_{i,j}^n \leftarrow \frac{\sigma_j^2 \left(P_{i,j}^n\right)^2}{((P_i^n)^T y^n)^2}$, $i \neq i^{*,n}$, $j = 1, .., k$;

$M^n \leftarrow M^n + 1$;

$N^{n,*} \leftarrow$ optimal solution of the updated optimization problem (2.12);

$j^n = \arg\max_j \frac{N_j^{*,n}}{N_j^n}$;

$N^{n+1} \leftarrow N^n + e_{j^n}$;

Sample alternative j^n and update the sample means y^{n+1};

$n \leftarrow n + 1$;

end

$i^{*,M} \leftarrow \arg\max_j (Qy^M)_j$.

x_j^n (i.e., they are under-sampled).

Of course, even if the unknown values were known, the objective of (2.12) is based on an approximation of (2.9), not the exact PCS. However, it can be shown that the geometric approximation produces an asymptotically exact solution in the regime where $M \to \infty$. To show this, we turn to the asymptotic convergence rates of the probability of *incorrect* selection.

2.3.2 Large Deviations Analysis of Probability of Incorrect Selection

Let us now increase the simulation budget in a way that satisfies $\lim_{M\to\infty} \frac{N_j^M}{M} = x_j$ for each j, where each $x_j > 0$ is a strictly positive constant. In other words, the budget becomes large, but each alternative receives a certain prespecified, nonzero proportion of it in the long term. For now, let us view the proportions x_j as fixed, but our ultimate goal will be to optimize them in a sense to be formalized later.

For $i \neq i^*$, define the "error set" $E_i = \{z : z_i \geq z_{i^*}\}$. We interpret $z \in E_i$ as a vector of possible S-indices under which the suboptimal alternative i appears to be better than the true best alternative i^*. Let $z^M = Qy^M$ be the (random) vector of S-indices obtained after M replications have been conducted. Then,

$$1 - PCS\left(z^M\right) = P\left(z^M \in \bigcup_{i \neq i^*} E_i\right),$$

that is, we select an incorrect alternative if z^M estimates at least one suboptimal alternative as being better than i^*. In the following, we will show that a certain $R_i > 0$ satisfies

$$\lim_{M \to \infty} \frac{1}{M} \log P\left(z^M \in E_i\right) = -R_i, \qquad i \neq i^*, \tag{2.15}$$

where, by the arguments in Glynn and Juneja [24], it straightforwardly follows that

$$\lim_{M \to \infty} \frac{1}{M} \log P\left(z^M \in \bigcup_{i \neq i^*} E_i\right) = -\min_{i \neq i^*} R_i. \tag{2.16}$$

In words, for large M, $P\left(z^M \in E_i\right)$ behaves like $e^{-R_i \cdot M}$, and the overall probability of incorrect selection is governed by the smallest (slowest) of the rate exponents across $i \neq i^*$.

The derivation of (2.15) uses the Gärtner-Ellis theorem [40]. Using this result, we can compute

$$R_i = \inf_{z \in E_i} I\left(z\right), \tag{2.17}$$

where I is the large deviations rate function of the sequence $\left\{z^M\right\}$. Since, for any M, $z^M = Qy^M$ is a linear transformation of a vector of independent normal random variables, it follows

a multivariate normal distribution, and therefore I can be computed in closed form, as shown in the following result.

Proposition 2.2. *The sequence* $\left\{z^M\right\}$ *obeys a large deviations law with rate function*

$$I(z) = \frac{1}{2}(z - Q\mu)^T Q\Gamma^{-1}Q^T(z - Q\mu),$$

where Γ *is a diagonal matrix with* $\Gamma_{j,j} = \frac{\sigma_j^2}{x_j}$.

Proof. Let

$$\Psi_M(\theta) = \mathbb{E}\left[e^{\theta z^M}\right] = e^{\theta^T Q\mu + \frac{1}{2}\theta^T Q\Lambda^2 Q^T\theta}$$

be the moment generating function of z^M, where Λ is a diagonal matrix with $\Lambda_{jj} = \frac{\sigma_j}{\sqrt{x_j \cdot M}}$. Here the sample size $N_j \approx x_j \cdot M$ is allowed to be fractional; however, as we will be passing to an asymptotic regime, this is not a major issue. Next, we calculate the scaled limit of the log-mgf, given by

$$\begin{aligned}
\Psi(\theta) &= \lim_{M\to\infty} \frac{1}{M} \log \Psi_M(M\theta) \\
&= \lim_{M\to\infty} \frac{1}{M}\left(M\theta^T Q\mu + \frac{1}{2}M^2\theta^T Q\Lambda^2 Q^T\theta\right) \\
&= \theta^T Q\mu + \lim_{M\to\infty} \frac{1}{2}\theta^T Q\left(M\Lambda^2\right)Q^T\theta \\
&= \theta^T Q\mu + \frac{1}{2}\theta^T Q\Gamma Q^T\theta.
\end{aligned}$$

By the Gärtner-Ellis theorem, the rate function is the Fenchel-Legendre transform of Ψ, defined

as

$$\begin{aligned} I(z) &= \sup_{\theta} \theta^T z - \Psi(\theta) \\ &= \sup_{\theta} \theta^T (z - Q\mu) - \frac{1}{2}\theta^T Q \Gamma Q^T \theta. \end{aligned} \tag{2.18}$$

The supremum is achieved at

$$\theta^* = Q\Gamma^{-1}Q^T (z - Q\mu),$$

and plugging this back into (2.18) yields the desired result. □

Thus, the right-hand side of (2.17) is the optimal value of a convex optimization problem with a quadratic objective and a single linear constraint $v_i^T z \leq 0$, where v_i is a vector of zeros with $v_{i,i} = -1$ and $v_{i,i^*} = 1$. Applying results in Section 2 of [41], we arrive at the closed-form solution

$$R_i = \frac{(v_i^T Q\mu)^2}{v_i^T Q \Gamma Q^T v_i}. \tag{2.19}$$

Note that R_i depends on the vector x of proportions through the diagonal matrix Γ. If $S = 0$, and consequently $Q = I$, we are back in the setting of classical ranking and selection, and it is easy to see that (2.19) reduces to the rate function derived in Example 1 of [24]. One can then formulate the concave maximization problem

$$\begin{aligned} &\text{maximize} \quad \min_{i \neq i^*} R_i(x) \\ &\text{subject to} \quad \mathbf{1}^T x = 1, \quad x_j \geq 0, \quad j = 1, ..., k. \end{aligned} \tag{2.20}$$

Recalling (2.16), we can see that the objective function in (2.20) is the rate exponent for the probability of incorrect selection. By maximizing this quantity, this probability is made to converge to zero at the fastest possible rate.

At the same time, using the notation from Section 2.3.1, it is easy to see that $R_i(x) = 1/\sum_{j=1}^{k} \frac{a_{i,j}}{x_j}$. Therefore, (2.20) is equivalent to (2.11) with the simulation budget scaled to 1. This proves that, by using the geometric approximation of PCS, we obtain a budget allocation that optimizes the asymptotic convergence rate of the probability of incorrect selection.

Within the simulation community, there is an entire stream of literature studying large deviations-based problems similar to (2.20), beginning with the seminal work of Glynn and Juneja [24], with some representative papers being [25], [42] and [43]. We find it is closely connected to the geometric approximation of Section 2.3.1, and the derivation of the optimal allocation under S-index selection is completely new.

2.3.3 A Provably Convergent Sequential Implementation Under S-index Selection (SIGD)

We have proposed SIOCBA to approximately maximize PCS when selecting with S-indices. However, when k is large, solving the nonlinear optimization problem (2.12) is expensive. Therefore, in this section, we propose a new sequential allocation policy that is more computationally efficient. We begin by deriving the convex dual of (2.12), since it will be more convenient to solve sequentially than the primal. We normalize $M = 1$, since we are interested in learning the asymptotically optimal solution as the budget becomes large.

Theorem 2.4. *The primal problem*

$$\begin{aligned} \min_{x} \quad & \max_{i \neq i^*} \sum_{j=1}^{k} \frac{a_{i,j}}{x_j} \\ \textit{subject to} \quad & x_j \geq 0, \quad j = 1, ..., k, \\ & \mathbf{1}^T x = 1, \end{aligned} \tag{2.21}$$

has the convex dual problem

$$\begin{aligned} \max_{\gamma} \quad & \sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i a_{i,j}} \\ \textit{subject to} \quad & \gamma_i \geq 0, \quad i \neq i^*, \\ & \mathbf{1}^T \gamma = 1, \end{aligned} \tag{2.22}$$

and strong duality holds.

Proof. As in (2.12), we can linearize the objective of (2.21) by adding a scalar variable ξ. We then obtain the Lagrangian

$$L(x, \xi, \gamma, \beta) = \xi + \sum_{i \neq i^*} \gamma_i \left(\sum_{j=1}^{k} \frac{a_{i,j}}{x_j} - \xi \right) + \beta \left(\sum_{j=1}^{k} x_j - 1 \right).$$

The dual function is obtained from the Lagrangian (Ch. 5, [38]) by computing

$$\begin{aligned} \tilde{g}(\gamma, \beta) &= \inf_{x,\xi} L(x, \gamma, \beta) \\ &= \inf_{x,\xi} \left(1 - \sum_{i \neq i^*} \gamma_i \right) \xi + \sum_{i \neq i^*} \gamma_i \sum_{j=1}^{k} \frac{a_{i,j}}{x_j} + \beta \left(\sum_{j=1}^{k} x_j - 1 \right) \end{aligned}$$

$$= \inf_{x,\xi} \left(1 - \sum_{i \neq i^*} \gamma_i\right) \xi + \sum_{j=1}^{k} \left(\frac{\sum_{i \neq i^*} \gamma_i a_{i,j}}{x_j} + \beta x_j \right) - \beta. \tag{2.23}$$

From (2.23), we can see that

$$\tilde{g}(\gamma, \beta) = \begin{cases} 2\sqrt{\beta} \sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i a_{i,j}} - \beta, & \sum_{i \neq i^*} \gamma_i = 1 \text{ and } \beta \geq 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

Therefore, the dual problem of (2.21) is given by

$$\begin{aligned} \max_{\gamma, \beta} \quad & 2\sqrt{\beta} \sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i a_{i,j}} - \beta \\ \text{subject to} \quad & \gamma_i \geq 0, \quad i \neq i^*, \\ & \sum_{i \neq i^*} \gamma_i = 1, \\ & \beta \geq 0. \end{aligned} \tag{2.24}$$

It is possible to simplify (2.24) by removing β entirely. Taking the derivative of the dual objective with respect to β, and setting it equal to zero, we obtain

$$\frac{\sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i a_{i,j}}}{\sqrt{\beta}} - 1 = 0,$$

so

$$\beta = \left(\sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i a_{i,j}} \right)^2.$$

Substituting this expression back into (2.24), we obtain (2.22), as required. It can be easily seen

that the primal problem is strictly feasible, so strong duality holds. □

Using the KKT optimality conditions for these problems, one finds that the optimal primal-dual pair (x^*, γ^*) satisfies

$$x_j^* = \frac{\sqrt{\sum_{i \neq i^*} \gamma_i^* a_{i,j}}}{\sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*} \gamma_i^* a_{i,j}}}, \quad j = 1, \ldots, k, \tag{2.25}$$

which allows us to easily convert a dual solution into a primal one. Of course, the index i^* and the inputs $a_{i,j}$ depend on the unknown values μ, so the dual solution still cannot be solved exactly. As in Section 2.3.1, one could potentially resolve it with the true values replaced by estimates, thus converging to the optimal solution over time, but we would like to avoid the cumbersome cost of solving a sequence of convex programs.

At a high level, our approach is based on the following concept. In nonlinear programming, one would typically solve (2.22) using an iterative gradient descent method, which repeats certain computations on the same problem instance until the iterates appear to converge (and one knows from theory that their limit must be a stationary point of the optimization problem). Our sequential procedure resembles such a method, but we now run only one iteration of it in each stage of sampling. In other words, given estimates y^n, we perform a single iteration of the gradient algorithm on problem (2.22) with μ replaced by y^n. This yields a new iterate γ^n, which is likely not optimal for the current dual problem (if we wanted to solve it to optimality, we would need to run many iterations). We then use γ^n to choose an alternative for the next replication, with the precise manner of this choice to be described further down. The results of the replication produce a new vector y^{n+1} of estimates, and the process is repeated. In other words, each iteration of the gradient algorithm is performed on a slightly different problem, namely, the dual with a different

set of estimates. Doing this is much faster than solving each individual problem to optimality, yet we prove that in this way the true solution of (2.22) is still recovered asymptotically.

The formal statement of the procedure is given in Algorithm 2.2, which we refer to as SIGD. We first explain the notation used in the statement, and motivate the steps of the algorithm, before proceeding to the main convergence results. Denote by

$$g(\gamma; y) = -\sum_{j=1}^{k} \sqrt{\sum_{i \neq i^*(y)} \gamma_i a_{i,j}(y)}$$

the objective function of the dual problem (2.22) with a generic vector y in place of μ, and the dependence of $i^*, a_{i,j}$ on y made explicit.

The initialization of the algorithm involves two small constants κ_0 and η, which are used for numerical stability and do not require any additional assumptions on the problem. In the nth iteration, Step 0 updates the inputs of the dual problem using the current estimates y^n. Step $0'$ is included for numerical stability, due to the nondifferentiability of g on the boundary of the dual feasible region; as will be argued in our proof, this step will be invoked at most finitely many times and thus does not play a major role in convergence analysis.

Step 1 selects an index h^n for which $\gamma_{h^n}^{n-1} > 0$, using a small threshold η for numerical stability. This index is needed to determine a descent direction in which to move. It is not necessary to consider all possible directions; Lin et al. [44] has shown that it is sufficient to consider elements of the set

$$D^h(\gamma) = \{e_i - e_h : i \neq h\} \bigcup \{e_h - e_i : i \neq h, \gamma_i > 0\}, \tag{2.26}$$

Algorithm 2.2: Sequential gradient descent algorithm for the dual problem (SIGD).

Input: A small constant κ_0 and $\eta < \frac{1}{k-1}$. Number of initial budgets N_0 for each alternative. Total Budget M.

Output: Sample means y^{M-kN_0+1}.

Initialization: Compute y^1 using a first-stage sample (e.g., by running N_0 replications for each alternative).

Repeat the following steps for $n = 1, 2, \ldots, M - kN_0$:

Step 0: Let $i^{*,n} = \arg\max_i [Qy^n]_i$, compute $a_{i,j}(y^n)$.

Step 0': Check whether $\min_j \sum_{i \neq i^{*,n}} \gamma_i^{n-1} a_{i,j}(y^n) > 0$. If this inequality does not hold, perturb γ^{n-1} to ensure the gradient is well defined.

Step 1: Choose some $h^n \in \left\{ i \neq i^{*,n} : \gamma_i^{n-1} \geq \eta \right\}$ with equal probability.

Step 2: Compute the descent direction d^n

$$\underset{d \in D^{h^n}(\gamma^{n-1})}{\arg\min} \; s^{\max}\left(d, \gamma^{n-1}\right) \nabla g\left(\gamma^{n-1}; y^n\right)^T d,$$

where $D^{h^n}(\gamma^{n-1})$ is as in (2.26), and $s^{\max}(d, \gamma^{n-1})$ is computed using (2.27). Let $V^n = \nabla g(\gamma^{n-1}; y^n)^T d^n$.

Step 3: If V^n does not satisfy (2.28) or (2.29), let $\gamma^n = \gamma^{n-1}$. Otherwise, choose $s^n = \text{LineSearch}\left(d^n, s^{\max}\left(d^n, \gamma^{n-1}\right), \gamma^{n-1}, y^n\right)$ and let $\gamma^n = \gamma^{n-1} + s^n d^n$.

Step 4: Compute x_j^n using (2.25) with inputs γ^n and $a_{i,j}(y^n)$.

Step 5: Choose $j^n = \arg\max_j \frac{x_j^n}{N_j^n}$, set $N_j^{n+1} = N_j^n + \mathbf{1}\{j = j^n\}$. Sample alternative j^n and compute updated sample means y^{n+1}.

for any h satisfying $\gamma_h > 0$. See Section 2.3.4.2 for the details. This reduction of the set of feasible directions to the much smaller set D^h motivates a 2-coordinate descent algorithm. For any direction $d \in D^h(\gamma)$, we define $s^{\max}(d, \gamma)$ to be the largest stepsize s such that $\gamma + s \cdot d$ remains dual-feasible. This quantity can be explicitly computed as

$$s^{\max}(d, \gamma) = \begin{cases} \gamma_h, & \text{if } d = e_j - e_h, j \neq h \text{ and } \nabla g(\gamma; y)^T d < 0, \\ \gamma_j, & \text{if } d = e_h - e_j, j \neq h \text{ and } \nabla g(\gamma; y)^T d < 0, \\ 0, & \text{if } \nabla g(\gamma; y)^T d \geq 0. \end{cases} \tag{2.27}$$

Step 2 performs this computation for all $D^{h^n}(\gamma^{n-1})$ and selects the direction d^n of steepest

descent. In the numerical optimization literature, techniques that minimize a first-order linear approximation of the objective are known as reduced gradient methods [45].

Step 3 then determines how far to go in this direction. First, recall that y^n changes in every iteration, which can be seen as introducing noise into the gradient. For this reason, we prefer not to change the iterate at all if the gradient appears to be small; we only update γ^{n-1} when $V^n = \nabla g(\gamma^{n-1}; y^n)^T d^n$ satisfies

$$V^n \geq \max\left\{-\kappa_0, -\left(\frac{\log n}{n}\right)^{1/4}\right\}, \tag{2.28}$$

and

$$s^{\max}\left(d^n, \gamma^{n-1}\right) V^n \geq \max\left\{-\kappa_0, -\left(\frac{\log n}{n}\right)^{1/2}\right\}, \tag{2.29}$$

where the rate $\frac{\log n}{n}$ is chosen to mitigate the effect of estimation error. When both conditions are satisfied, we update $\gamma^n = \gamma^{n-1} + s^n d^n$, where the stepsize s^n is chosen using Algorithm 2.3. Essentially, this procedure performs a line search to identify s satisfying conditions (2.30)-(2.32). Condition (2.30) ensures that the gradient is well-defined at the new dual solution, while (2.31)-(2.32) are based on the well-known Wolfe conditions [46]. It can be shown that Algorithm 2.3 will terminate in a finite number of iterations.

Step 4 uses (2.25) to convert the updated dual solution γ^n into a primal solution x^n, which represents an approximation of the optimal budget allocation. Finally, Step 5 allocates the next replication to alternative $j^n = \arg\max_j \frac{x_j^n}{N_j^n}$, which again favors under-sampled alternatives.

Although Algorithm 2.2 appears to involve more technical complications than SIOCBA

Algorithm 2.3: LineSearch$(d, s^{\max}, \gamma, y)$

Input: Descent direction d, maximum feasible stepsize $s^{\max}$, dual solution γ, estimated values y, parameters $\alpha, \alpha' > 0$ and $\tau \in (0, 1)$.

Output: Stepsize s.

Let $s = s^{\max}$.

while *Any of the conditions*

$$\min_j \sum_{i \neq i^*} \gamma_i a_{i,j}(y) > 0, \tag{2.30}$$

$$g(\gamma + s \cdot d; y) \leq g(\gamma; y) + \alpha s \nabla g(\gamma; y)^T d, \tag{2.31}$$

$$\nabla g(\gamma + s \cdot d; y)^T d \leq \alpha' \left| \nabla g(\gamma; y)^T d \right|, \quad \textit{if } \nabla g(\gamma + s \cdot d; y)^T d > 0, \tag{2.32}$$

not satisfied **do**

  $s \leftarrow \tau s$.

end

(Algorithm 2.1), in practice it is much more efficient, since it does not require the use of any convex programming solver. Instead of separating estimation and optimization, so that one has to solve a new problem for each set of sample means, we integrate these two aspects into a single iterative algorithm, which provably converges to the optimal solution (x^*, γ^*) of (2.21)-(2.22). The first major result establishes the consistency of the procedure: each alternative will receive a nonzero proportion of the budget asymptotically, ensuring that $y^n \to \mu$.

Theorem 2.5. *Under Algorithm 2.2,* $\liminf_{n\to\infty} x_j^n > 0$ *for all* $j = 1, ..., k$.

Theorem 2.5 simplifies the subsequent analysis, since we will have $i^{*,n} = i^*$ for all sufficiently large n after some transient period of random length. We can then obtain the main convergence result, which is that Algorithm 2.2 learns the true optimal allocation of the budget in the limit.

Theorem 2.6. *Let* $\{\gamma^n\}$ *be the sequence of iterates generated by Algorithm 2.2. Every limit point of this sequence is a stationary point of the true dual problem (2.22).*

Note that Theorem 2.6 is much stronger than Theorem 2.5, as it implies through (2.25)

and strong duality that, for any j, $x_j^n \to x_j^*$, where x_j^* is the optimal solution of the true primal problem (2.21). These results show that the SIGD algorithm learns the asymptotically optimal budget allocation, without tuning, and without the need for a convex programming solver. Recent work in ranking and selection has considered similar goals (e.g., the top-two method of [47] or the tuning-free technique of [48]), but our setting here is much more complicated, because the objective in (2.22) is not separable across alternatives: each y_i affects multiple terms in the sum. The SIGD algorithm has little in common with the aforementioned methods, instead being based on numerical optimization techniques, and thus constitutes a novel contribution in its own right.

2.3.4 Convergence Analysis of SIGD

We now provide the proofs of Theorems 2.5 and 2.6, which require some preliminary results regarding the properties of the dual objective function $g(\cdot\,;\, y^n)$ and Algorithm 2.2.

2.3.4.1 Properties of the Dual Objective Function

First, note that $\bar{y} = \lim_{n\to\infty} y^n$ is always well-defined: if $N_j^n \to \infty$, then $y_j^n \to \mu_j$, whereas if $\sup_n N_j^n < \infty$, the sample mean is only updated finitely many times. Theorem 2.5 will show that $\bar{y} = \mu$, but at the moment this has not yet been established.

Let $i' = \arg\max_j [Q\bar{y}]_j$ be the alternative selected under the limiting estimates $\bar{y}$. Similarly, let $\bar{a}_{i,j} = a_{i,j}(\bar{y})$. From the definition of $a_{i,j}$, it is easy to see that

$$|a_{i,j}(y^n) - \bar{a}_{i,j}| \le K_a \|y^n - \bar{y}\|, \qquad j = 1, ..., k,\ i \ne i'. \tag{2.33}$$

Consequently, we have $a_{i,j}(y^n) \to \bar{a}_{i,j}$, so the sequence $\{a_{i,j}(y^n)\}$ is bounded. We can therefore

let

$$a_{i,j}^{\max} = \sup_n a_{i,j}\left(y^n\right), \qquad a_{i,j}^{\min} = \inf_n a_{i,j}\left(y^n\right),$$

and we will have $a_{i,j}^{\min} > 0$ for any $i \neq i'$ satisfying $Q_{i',j} \neq Q_{i,j}$. This fact leads to the following technical lemma establishing a kind of uniform convergence of g.

Lemma 2.1. *There exists a constant $K_g > 0$ such that, for all n,*

$$\left|g(\gamma; y^n) - g(\gamma; \bar{y})\right| \leq K_g \left\|y^n - \bar{y}\right\|, \qquad \forall \gamma \in \mathcal{G}.$$

Consequently, $g(\gamma; y^n)$ converges uniformly to $g(\gamma; \bar{y})$ on $\mathcal{G}$.

Proof. For any $\gamma \in \mathcal{G}$,

$$\begin{aligned}
\left|g\left(\gamma; y^n\right) - g\left(\gamma; \bar{y}\right)\right| &= \left|\sum_{j=1}^k \sqrt{\sum_{i \neq i'} \gamma_i a_{i,j}\left(y^n\right)} - \sum_{j=1}^k \sqrt{\sum_{i \neq i'} \gamma_i \bar{a}_{i,j}}\right| \\
&\leq \sum_{j=1}^k \frac{\sum_{i \neq i'} \gamma_i \left|a_{i,j}\left(y^n\right) - \bar{a}_{i,j}\right|}{\sqrt{\sum_{i \neq i'} \gamma_i a_{i,j}\left(y^n\right)} + \sqrt{\sum_{i \neq i'} \gamma_i \bar{a}_{i,j}}} \\
&\leq \sum_{j=1}^k \sum_{i: Q_{i',j} \neq Q_{i,j}} \frac{\gamma_i \left|a_{i,j}\left(y^n\right) - \bar{a}_{i,j}\right|}{\sqrt{\gamma_i a_{i,j}\left(y^n\right)} + \sqrt{\gamma_i \bar{a}_{i,j}}} \\
&= \sum_{j=1}^k \sum_{i: Q_{i',j} \neq Q_{i,j}} \frac{\sqrt{\gamma_i} \left|a_{i,j}\left(y^n\right) - \bar{a}_{i,j}\right|}{\sqrt{a_{i,j}\left(y^n\right)} + \sqrt{\bar{a}_{i,j}}} \\
&\leq \sum_{j=1}^k \sum_{i: Q_{i',j} \neq Q_{i,j}} \frac{\left|a_{i,j}\left(y^n\right) - \bar{a}_{i,j}\right|}{\sqrt{\bar{a}_{i,j}}} \qquad (2.34) \\
&\leq \frac{k^2 K_a}{\sqrt{\min_j \min_{i: Q_{i',j} \neq Q_{i,j}} a_{i,j}^{\min}}} \left\|y - \bar{y}\right\|, \qquad (2.35)
\end{aligned}$$

where (2.34) is obtained by $\gamma_i \in [0,1]$ and $a_{i,j}^n \geq 0$, and (2.35) holds from (2.33) and the boundedness of $\{a_{i,j}\left(y^n\right)\}$. □

Lemma 2.2. *There exist constants $b_1 < b_2 < 0$ such that $b_1 \leq g(\gamma; y^n) \leq b_2 < 0$ for all n and all $\gamma \in \mathcal{G}$.*

Proof. By Lemma 2.1, it suffices to show that $g(\gamma; \bar{y})$ is bounded between two negative numbers. It is easy to see that for any $\gamma \in \mathcal{G}$,

$$g(\gamma; \bar{y}) = -\sum_{j=1}^{k} \sqrt{\sum_{i \neq i'} \gamma_i \bar{a}_{i,j}} \geq -\sum_{j=1}^{k} \sqrt{\sum_{i \neq i'} \bar{a}_{i,j}},$$

meaning that $g(\gamma; \bar{y})$ is bounded below by a strictly negative number. For the upper bound, we have

$$g(\gamma; \bar{y}) = -\sum_{j=1}^{k} \sqrt{\sum_{i \neq i'} \gamma_i \bar{a}_{i,j}} \leq -\sqrt{\sum_{i \neq i'} \gamma_i \bar{a}_{i,i'}} \leq -\sqrt{\min_{i \neq i'} \bar{a}_{i,i'}} < 0.$$

The proof is thus complete. □

Lemma 2.3. *Let $\{\gamma^n\}$ be the sequence generated by Algorithm 2.2, and suppose that*

$$\liminf_{n \to \infty} \sum_{i \neq i'} \gamma_i^{n-1} a_{i,j}(y^n) > 0 \tag{2.36}$$

for all j. Then, there exists a constant $K > 0$ such that, for all n,

$$\left\| \nabla g\left(\gamma^{n-1}; y^n\right) - \nabla g\left(\gamma^{n-1}; \mu\right) \right\| \leq K \left\| y^n - \mu \right\|.$$

Proof. From (2.33), for any j and $\gamma^{n-1} \in \mathcal{G}$, we have

$$\left|\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right) - \sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right| = \left|\sum_{i\neq i'} \gamma_i^{n-1}\left(a_{i,j}\left(y^n\right) - a_{i,j}\left(\mu\right)\right)\right| \leq K_a \left\|y^n - \mu\right\|. \tag{2.37}$$

From (2.36), it also follows that

$$\liminf_{n\to\infty} \sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right) > 0.$$

Therefore, there exists a constant $b > 0$ such that

$$\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right) > b, \qquad \sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right) > b$$

for all n and for all j. Then, for $i \neq i'$,

$$\begin{aligned}
& \left|\left[\nabla g\left(\gamma^{n-1}; y^n\right)\right]_i - \left[\nabla g\left(\gamma^{n-1}; \mu\right)\right]_i\right| \\
= & \left|\sum_{j=1}^{k} \frac{a_{i,j}\left(y^n\right)}{\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right)} - \sum_{j=1}^{k} \frac{a_{i,j}\left(\mu\right)}{\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)}\right| \\
\leq & \sum_{j=1}^{k} \frac{\left|a_{i,j}\left(y^n\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right)\right) - a_{i,j}\left(\mu\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right)\right|}{\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right)\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right)} \\
\leq & \frac{1}{b^2} \sum_{j=1}^{k} \left|a_{i,j}\left(y^n\right)\left(\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right)\right)\right) - a_{i,j}\left(\mu\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right)\right| \\
\leq & \frac{1}{b^2} \sum_{j=1}^{k} \left(\left|a_{i,j}\left(y^n\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(y^n\right)\right) - a_{i,j}\left(y^n\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right)\right| + \right. \\
& \left.\left|a_{i,j}\left(y^n\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right) - a_{i,j}\left(\mu\right)\left(\sum_{i\neq i'} \gamma_i^{n-1} a_{i,j}\left(\mu\right)\right)\right|\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{b^2}\sum_{j=1}^{k}\max_{Q_{i',j}\neq Q_{i,j}} a_{i,j}^{\max}\left(\left|\left(\sum_{i\neq i'}\gamma_i^{n-1}a_{i,j}\left(y^n\right)\right)-\left(\sum_{i\neq i'}\gamma_i^{n-1}a_{i,j}\left(\mu\right)\right)\right|+\left|a_{i,j}\left(y^n\right)-a_{i,j}\left(\mu\right)\right|\right)\\
&\leq \frac{2kK_a\max_j\max_{Q_{i',j}\neq Q_{i,j}} a_{i,j}^{\max}}{b^2}\left\|y^n-\mu\right\|,
\end{aligned}
$$

which completes the proof. □

2.3.4.2 Properties of Algorithm 2.2

Below, we discuss several properties of the sequential algorithm that explain various design choices used in its construction. We also prove properties that will be important later for showing the main results.

Nondifferentiability of g. Step $0'$ of Algorithm 2.2 is designed to avoid instability resulting from γ^{n-1} being too close to the boundary of the feasible region

$$
\mathcal{G}=\left\{\gamma\in\mathbb{R}^{k-1}\,:\,\gamma_i\geq 0,\ \mathbf{1}^T\gamma=1\right\}.
$$

This is because $g\left(\gamma;y\right)$ is nondifferentiable only on the boundary, as we will now show. First, from the definition of g it follows that ∇g is not well-defined if and only if there exists $j\in\{1,...,k\}$ such that

$$
\sum_{i\neq i^*(y)}\gamma_i a_{i,j}\left(y\right)=0, \tag{2.38}
$$

where $i^*\left(y\right)=\arg\max_i y_i$.

We first observe that $a_{i,j}\geq 0$ for all j and $i\neq i^*$. Since $Q_{i^*,i^*}-Q_{i,i^*}>0$ by Theorem 2.1, we have $a_{i,i^*}>0$. On the other hand, for $j\neq i^*$, we have $Q_{i^*,j}-Q_{j,j}<0$ by the same result, so $a_{j,j}>0$. Since all $\gamma_i,a_{i,j}\geq 0$, it follows that (2.38) can hold if and only if $\gamma_j=0$.

This motivates Step $0'$, in which we perturb γ^{n-1} to avoid this situation. Nonetheless, the issue of nondifferentiability is not critical to the long-term performance of the algorithm, as shown in the following result.

Lemma 2.4. *Let $\gamma \in \mathcal{G}$ and suppose that $\sum_{i \neq i^*} \gamma_i a_{i,j}(y) > 0$ for all j. Then, $\sum_{i \neq i^*} \gamma_i a_{i,j}(y') > 0$ for all j and y'.*

Proof. From the premise, we know that for each j, there exists at least one i with both $\gamma_i, a_{i,j}(y) > 0$. Note that $a_{i,j}(y) = \frac{\sigma_j^2 P_{i,j}^2}{\left(P_i^T y\right)^2}$, so changing y to y' will not change whether or not $a_{i,j}$ is zero or nonzero. The conclusion then follows. □

Since $i^{*,n}$ converges to i', for all sufficiently large n, the sum $\sum_{i \neq i^*} \gamma_i a_{i,j}(y)$ will be computed for a fixed i'. Once this happens, Lemma 2.4 ensures that Step $0'$ will no longer be invoked after some finite number of iterations.

Sufficiency of reduced direction set. Because the dual problem (2.22) has a fairly simple feasible region, verifying the optimality of a dual solution can also be simplified. Given $\gamma \in \mathcal{G}$, the set of all feasible directions is given by

$$\mathcal{D}(\gamma) = \left\{ d \in \mathbb{R}^{k-1} \,:\, \mathbf{1}^T d = 0, d_i \geq 0 \text{ if } \gamma_i = 0 \right\}.$$

However, Lin et al. [44] showed that $\mathcal{D}(\gamma)$ can be generated by a much smaller set

$$D^h(\gamma) = \{e_i - e_h \,:\, i \neq h\} \bigcup \{e_h - e_i \,:\, i \neq h, \gamma_i > 0\}$$

for any h with $\gamma_h > 0$. In other words, $\text{Cone}\left(D^h(\gamma)\right) = \mathcal{D}(\gamma)$. Consequently, we obtain the following simplified optimality condition.

Lemma 2.5. *[44] Fix any h with $\gamma_h > 0$. A dual solution $\gamma \in \mathcal{G}$ is a stationary point of (2.22) if and only if $\nabla g(\gamma; y)$ is well-defined and*

$$\nabla g(\gamma; y)^T d \geq 0, \qquad \forall d \in D^h(\gamma).$$

Lemma 2.5 motivates the use of a reduced direction set in Step 2 of Algorithm 2.2, as well as the arbitrary selection of h in Step 1. Note that, for η small enough, there will always be at least one h with $\gamma_h \geq \eta$ due to the equality constraint in $\mathcal{G}$.

Modified line search. Step 3 in Algorithm 2.2 uses the modified line search procedure in Algorithm 2.3. We now prove that this procedure terminates in finite time. The proof is based on Proposition 4.1 in [44] and Lemma 3.1 in [46], and also uses Lemmas 2.3 and 2.5 from Section 2.3.4.1.

Note that, in Step 3 of Algorithm 2.2, we may not call the line search procedure in every time stage n. Thus, technically, the following result should be applied to the subsequence of time stages at which line search is called; however, to avoid overcomplicating the notation, we work with the sequence of *distinct* iterates obtained from line search and relabel their indices as $\{1, 2, ...\}$.

Lemma 2.6. *Let γ^n be the solution obtained from the nth time that line search is called (thus, γ^{n-1} denotes the dual solution obtained from the previous call, y^n is the vector of most recent sample means, and d^n is the most recent descent direction). Then:*

(i) At step n, a stepsize s^n can always be found in a finite number of iterations such that conditions (2.30), (2.31) and (2.32) are satisfied with $\gamma = \gamma^{n-1}$ and $y = y^n$.

(ii) Suppose that $\gamma^n \to \bar{\gamma}$ *and*

$$\lim_{n\to\infty} g\left(\gamma^{n-1};\mu\right) - g\left(\gamma^{n-1} + s^n d^n;\mu\right) = 0.$$

Then,

$$\lim_{n\to\infty} s^{\max}\left(d^n,\gamma^{n-1}\right)\nabla g\left(\gamma^{n-1};\mu\right)^T d^n = 0 \tag{2.39}$$

$$\lim_{n\to\infty} s^{\max}\left(d^n,\gamma^{n-1}\right)\nabla g\left(\gamma^{n-1};y^n\right)^T d^n = 0. \tag{2.40}$$

Proof. We know that $\nabla g\left(\gamma^{n-1};y^n\right)^T d^n < 0$ by Lemma 2.5. Since $\min_j \sum_{i\neq i^{*,n}} \gamma_i a_{i,j}\left(y^n\right) = 0$ can only occur on the boundary of $\mathcal{G}$, condition (2.30) must be satisfied for any $0 < s < s^{\max}\left(d^n,\gamma^{n-1}\right)$. Since $g\left(\cdot\,;\,y^n\right)$ is smooth on $\mathcal{G}$, there must exist some $s_0 > 0$ such that $\nabla g(\gamma^{n-1} + sd^n\,;\,y^n)^T d^n < 0$ for any $s \in (0,s_0)$. In other words, when the stepsize is small enough, condition (2.32) can be neglected. Since $g\left(\cdot;y^n\right)$ is bounded below by Lemma 2.2, while the mapping $\ell(s) = g\left(\gamma^{n-1};y^n\right) + \alpha s\nabla g\left(\gamma^{n-1};y^n\right)^T d^n$ is unbounded below, it follows that $\ell(s)$ must intersect $g\left(\gamma^{n-1}+sd^n;y^n\right)$. Let s_1 be the smallest such intersection point; then, for any $s \in (0,s_1)$, condition (2.31) must be satisfied. It is easy to see that there exists some positive integer m such that

$$\tau^m s^{\max}\left(d^n,\gamma^{n-1}\right) < \min\left\{s^{\max}\left(d^n,\gamma^{n-1}\right), s_0, s_1\right\},$$

which is sufficient to establish (i).

With regard to (ii), (2.39) follows directly from Proposition 4.1(ii) in [44], while (2.40)

follows from Lemma 2.3. □

2.3.4.3 Proof of Theorem 2.5

Proof. To establish the main result, it is sufficient to show

$$\liminf_{n\to\infty} \sum_{i\neq i'} \gamma_i^n a_{i,j}\left(y^n\right) > 0, \tag{2.41}$$

and apply (2.25) together with Lemma 2.2. Note that we have replaced $i^{*,n}$ in (2.41) by $i' = \lim_{n\to\infty} i^{*,n}$ as in Section 2.3.4.1. In the following, we assume without loss of generality that $i^{*,n} = i'$, as this will be true for sufficiently large n, and we are interested in the asymptotic behaviour of the algorithm.

When $j = i'$, (2.41) follows from the boundedness of $\{a_{i,j}\left(y^n\right)\}$, established previously. Thus, it is sufficient to handle the case $j \neq i'$. We proceed by contradiction: suppose that there exists $j \neq i'$ with

$$\liminf_{n\to\infty} \sum_{i\neq i'} \gamma_i^n a_{i,j}\left(y^n\right) = 0.$$

Then, there exists a subsequence $\{n_t\}$ such that $\sum_{i\neq i'} \gamma_i^{n_t} a_{i,j}\left(y^{n_t}\right) \to 0$. Because $\{a_{i,j}\left(y^n\right)\}$ is uniformly bounded, it must be the case that $\gamma_i^{n_t} \to 0$ for all i satisfying $Q_{i',j} \neq Q_{i,j}$.

Define

$$\begin{aligned} C &= \max_{i\neq i'} \sum_{j'=1}^{k} \sqrt{a_{i,j'}^{\max}}, \\ C_1 &= \frac{\min_{i:Q_{i',j}\neq Q_{i,j}} \left(a_{i,j}^{\min}\right)^2}{\sum_{i:Q_{i',j}\neq Q_{i,j}} a_{i,j}^{\max}}, \\ c_0 &= \frac{C_1}{C^2}\eta. \end{aligned}$$

Note that $\frac{C_1}{C^2} \leq 1$ and $c_0 \leq \eta$. Letting b_1, b_2 be the values in Lemma 2.2, choose ε satisfying

$$0 < \varepsilon < \min \left\{ \frac{C_1}{C^2 \left(\frac{1+\alpha'}{\sqrt{\eta}} + \frac{\alpha'}{\sqrt{c_0}} \right)^2}, \frac{C_1}{\left(\sqrt{\frac{3}{2}} \frac{C}{\eta} + \frac{3\alpha'(b_2 - b_1)}{\alpha \eta} \right)^2}, c_0, \frac{\eta}{3} \right\}. \tag{2.42}$$

There exists t_0 such that, for any $t > t_0$, $\gamma_i^{n_t} \leq \varepsilon$. Since the direction is in the form of $e_h - e_{h'}$, $h \neq h'$, $h, h' \neq i'$, the value of $\gamma^{n_t - 1}$ must be described by one and only one of the following three cases:

1) For all i satisfying $Q_{i',j} \neq Q_{i,j}$, $\gamma_i^{n_t - 1} \leq c_0$.

2) There exists i satisfying $Q_{i',j} \neq Q_{i,j}$ such that $\gamma_i^{n_t - 1} \in (c_0, \eta)$ and $\gamma_{j'}^{n_t - 1} \leq \varepsilon$ for $j' \neq i$ satisfying $Q_{i',j} \neq Q_{j',j}$.

3) There exists i satisfying $Q_{i',j} \neq Q_{i,j}$ such that $\gamma_i^{n_t - 1} \geq \eta$, and $\gamma_{j'}^{n_t - 1} \leq \varepsilon$ for $j' \neq i$ satisfying $Q_{i',j} \neq Q_{j',j}$.

We consider each case separately.

In case 1),

$$\sum_{i \neq i'} \gamma_i^{n_t - 1} a_{i,j}^{n_t} \leq c_0 \sum_{Q_{i',j} \neq Q_{i,j}} a_{i,j}^{n_t} \leq c_0 \sum_{Q_{i',j} \neq Q_{i,j}} a_{i,j}^{\max}.$$

By assumption, $\gamma_i^{n_t - 1} \leq c_0 \leq \eta$ for all i satisfying $Q_{i',j} \neq Q_{i,j}$, and therefore the index h^{n_t} selected in Step 1 of Algorithm 2.2 cannot be such an i. The corresponding component of the gradient $\nabla g(\gamma^{n_t - 1}; y^{n_t})$ must then be bounded below, as can be seen from

$$\left[\nabla g(\gamma^{n_t - 1}; y^{n_t}) \right]_{h^{n_t}} = -\frac{1}{2} \sum_{i=1}^{k} \frac{a_{h^{n_t},i}(y^{n_t})}{\sqrt{\sum_{j' \neq i'} \gamma_{j'}^{n_t - 1} a_{j',i}(y^{n_t})}}$$

$$
\begin{aligned}
&\geq -\frac{1}{2}\sum_{i=1}^{k}\frac{a_{h^{n_t},i}\left(y^{n_t}\right)}{\sqrt{\gamma_{h^{n_t}}^{n_t-1}a_{h^{n_t},i}\left(y^{n_t}\right)}}\\
&= -\frac{1}{2}\sum_{i=1}^{k}\sqrt{\frac{a_{h^{n_t},i}\left(y^{n_t}\right)}{\gamma_{h^{n_t}}^{n_t-1}}}\\
&\geq -\frac{1}{2\sqrt{\eta}}\sum_{i=1}^{k}\sqrt{a_{h^{n_t},i}\left(y^{n_t}\right)}\\
&\geq -\frac{C}{2\sqrt{\eta}}.
\end{aligned}
\tag{2.43}
$$

On the other hand, for i satisfying $Q_{i',j} \neq Q_{i,j}$, the corresponding component of the gradient satisfies

$$
\begin{aligned}
\left[\nabla g(\gamma^{n_t-1};y^{n_t})\right]_i &= -\frac{1}{2}\sum_{j'=1}^{k}\frac{a_{i,j'}\left(y^{n_t}\right)}{\sqrt{\sum_{j''\neq i'}\gamma_{j''}^{n_t-1}a_{j'',j'}\left(y^{n_t}\right)}}\\
&\leq -\frac{a_{i,j}\left(y^{n_t}\right)}{2\sqrt{\sum_{j''\neq i'}\gamma_{j''}^{n_t-1}a_{j'',j}\left(y^{n_t}\right)}}\\
&\leq -\frac{a_{i,j}^{\min}}{2\sqrt{c_0\sum_{j'':Q_{i',j}\neq Q_{j'',j}}a_{j'',j}^{\max}}}\\
&\leq -\frac{Ca_{i,j}^{\min}}{2\sqrt{\eta\min_{j'':Q_{i',j}\neq Q_{j'',j}}a_{j'',j}^{\min}}}\\
&\leq -\frac{C}{2\sqrt{\eta}}\\
&\leq \left[\nabla g(\gamma^{n_t-1};y^{n_t})\right]_{h^{n_t}}.
\end{aligned}
$$

Therefore, the direction d^{n_t} chosen in Step 2 cannot be $e_{h^{n_t}} - e_i$. Consequently, $\gamma_i^{n_t} \geq \gamma_i^{n_t-1}$, i.e., if $\gamma_i^{n_t-1} \leq c_0$ and $Q_{i',j} \neq Q_{i,j}$, then the ith component of γ^{n_t-1} cannot decrease at step n_t. Therefore, the only possible way to have $\sum_{i\neq i'}\gamma_i^{n_t}a_{i,j}\left(y^{n_t}\right) \to 0$ is $\gamma_i^{n_t} \equiv 0$ for all i satisfying $Q_{i',j} \neq Q_{i,j}$, but this is prevented by Step $0'$ of Algorithm 2.2 and condition (2.30).

In case 2), similarly, we have $h^{n_t} \neq i$ for any i satisfying $Q_{i',j} \neq Q_{i,j}$. To have $\gamma_i^{n_t} \leq \varepsilon$ for all such i, the direction d^{n_t} chosen in Step 2 must be $e_{h^{n_t}} - e_i$ with maximum feasible stepsize $s^{\max}\left(d^{n_t}, \gamma^{n_t-1}\right) = \gamma_i^{n_t-1}$ along this direction, so that the Algorithm can enter into the line search step at n_t. Furthermore, the selected stepsize s^{n_t} must satisfy $s^{n_t} \geq c_0 - \varepsilon$ and the modified Wolfe condition

$$\nabla g\left(\gamma^{n_t}; y^{n_t}\right)^T d^{n_t} > 0 \qquad \Rightarrow \qquad \nabla g\left(\gamma^{n_t}; y^{n_t}\right)^T d^{n_t} \leq \alpha' \left|\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)^T d^{n_t}\right|, \tag{2.44}$$

where $\gamma^{n_t} = \gamma^{n_t-1} + s^{n_t} d^{n_t}$. Since $d^{n_t} = e_{h^{n_t}} - e_i$, we have $\gamma_{h^{n_t}}^{n_t} \geq \gamma_{h^{n_t}}^{n_t-1} \geq \eta$. Similarly to (2.43), we can obtain

$$\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_{h^{n_t}} \in \left[-\frac{C}{2\sqrt{\eta}}, 0\right], \qquad \left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_{h^{n_t}} \in \left[-\frac{C}{2\sqrt{\eta}}, 0\right].$$

Similarly, since $\gamma_i^{n_t-1} > c_0$, we also have

$$\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i \in \left[-\frac{C}{2\sqrt{c_0}}, 0\right].$$

Then (2.44) yields

$$\begin{aligned}
\left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_i &\geq \left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_{h^{n_t}} - \alpha' \left|\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_{h^{n_t}} - \left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i\right| \\
&\geq \left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_{h^{n_t}} - \alpha' \left(\left|\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_{h^{n_t}}\right| + \left|\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i\right|\right) \\
&\geq -\frac{C}{2\sqrt{\eta}} - \alpha' \left(\frac{C}{2\sqrt{\eta}} + \frac{C}{2\sqrt{c_0}}\right) \\
&= -\frac{C}{2}\left(\frac{1+\alpha'}{\sqrt{\eta}} + \frac{\alpha'}{\sqrt{c_0}}\right).
\end{aligned} \tag{2.45}$$

Since, by assumption, we have $\gamma_i^{n_t} \leq \varepsilon$ for any i satisfying $Q_{i',j} \neq Q_{i,j}$, it follows that

$$\sum_{i \neq i'} \gamma_i^{n_t} a_{i,j}\left(y^{n_t}\right) \leq \varepsilon \sum_{i: Q_{i',j} \neq Q_{i,j}} a_{i,j}^{\max},$$

and, consequently,

$$\left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_i \leq -\frac{a_{i,j}^{\min}}{2\sqrt{\varepsilon \sum_{i: Q_{i',j} \neq Q_{i,j}} a_{i,j}^{\max}}} < -\frac{C}{2}\left(\frac{1+\alpha'}{\sqrt{\eta}} + \frac{\alpha'}{\sqrt{c_0}}\right), \tag{2.46}$$

which contradicts (2.45). Therefore, for case 2), it is not possible to have $\gamma_i^{n_t} \leq \varepsilon$ for all i satisfying $Q_{i',j} \neq Q_{i,j}$.

In case 3), there are two subcases: either (i) $Q_{i',j} = Q_{h^{n_t},j}$, or (ii) $i = h^{n_t}$. Subcase (i) is handled similarly to case 2, resulting in the desired contradiction. Therefore, we can focus on the second subcase $i = h^{n_t}$, and we have

$$\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i \geq -\frac{C}{2\sqrt{\eta}},$$

since $\gamma_i^{n_t-1} > \eta$. It is necessary for the value of $\gamma_i^{n_t-1}$ to decrease such that $\gamma_i^{n_t} \leq \varepsilon$, which implies that $d^{n_t} = e_{i''} - e_i$ for some $i'' \neq i, i'$ and

$$s^{n_t} \geq \eta - \varepsilon \geq \frac{2\eta}{3} > \varepsilon.$$

First of all, notice that i'' cannot satisfy $Q_{i',j} \neq Q_{i'',j}$; otherwise we would have $\gamma_{i''}^{n_t} = \gamma_{i''}^{n_t-1} +$

$s^{n_t} > \varepsilon$. Then, condition (2.31) yields

$$\begin{aligned} g\left(\gamma^{n_t}; y^{n_t}\right) &\leq g\left(\gamma^{n_t-1}; y^{n_t}\right) + \alpha s^{n_t} \nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)^T d^{n_t} \\ &= g\left(\gamma^{n_t-1}; y^{n_t}\right) + \alpha s^{n_t} \left(\nabla \left[g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_{i''} - \left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i\right). \end{aligned}$$

Then,

$$\begin{aligned} 0 &> \nabla \left[g(\gamma^{n_t-1}; y^{n_t})\right]_{i''} - \left[\nabla g(\gamma^{n_t-1}; y^{n_t})\right]_i \\ &\geq \frac{1}{\alpha s^{n_t}} \left(g\left(\gamma^{n_t}; y^{n_t}\right) - g\left(\gamma^{n_t-1}; y^{n_t}\right)\right) \\ &\geq -\frac{b_2 - b_1}{\alpha s^{n_t}} \\ &\geq -\frac{3\left(b_2 - b_1\right)}{2\alpha\eta}, \end{aligned}$$

where the third line is obtained by Lemma 2.2, and the fourth line is due to the fact that $s^{n_t} \geq \frac{2\eta}{3}$. Since $\gamma_{i''}^{n_t} = \gamma_{i''}^{n_t-1} + s^{n_t} \geq \frac{2\eta}{3}$, we have

$$\left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_{i''} \geq -\frac{\sqrt{3}C}{2\sqrt{2\eta}}.$$

Then, repeating the arguments of (2.44) and (2.45), we have

$$\begin{aligned} \left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_i &\geq \left[\nabla g\left(\gamma^{n_t}; y^{n_t}\right)\right]_{i''} - \alpha' \left|\left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_{i''} - \left[\nabla g\left(\gamma^{n_t-1}; y^{n_t}\right)\right]_i\right| \\ &\geq -\frac{\sqrt{3}C}{2\sqrt{2\eta}} - \alpha' \left(\frac{3\left(b_2 - b_1\right)}{2\alpha\eta}\right). \end{aligned} \tag{2.47}$$

By the definition of ε, similar to (2.46), we find that (2.47) contradicts the assumption that $\gamma_i^{n_t} \leq \varepsilon$

for any i satisfying $Q_{i',j} \neq Q_{i,j}$. This completes the proof. □

2.3.4.4 Proof of Theorem 2.6

With Theorem 2.5, we can now replace $i^{*,n}$ by i^* in all calculations made by the algorithm, since we now know that $\lim_{n\to\infty} y^n = \mu$. We first prove a technical lemma showing that ∇g is Lipschitz continuous in γ (previously Lemma 2.3 considered Lipschitz continuity in the estimated value y).

Lemma 2.7. *Let $\{\gamma^n\}$ be the sequence generated by Algorithm 2.2, and define*

$$\mathcal{B}^n = Conv\left(\left\{\gamma^{n-1}, \gamma^n\right\}\right), \quad n = 1, 2, ...$$

There exists a constant $K_\gamma > 0$, such that for all n,

$$\left\|\nabla g\left(\gamma; y^n\right) - \nabla g\left(\gamma'; y^n\right)\right\| \leq K_\gamma \left\|\gamma - \gamma'\right\|, \quad \forall \gamma, \gamma' \in \mathcal{B}^n.$$

Proof. From the boundedness of $\{a_{i,j}\left(y^n\right)\}$, we can see that, for any j and any $\gamma, \gamma' \in \mathcal{G}$,

$$\left|\sum_{i\neq i^*} \left(\gamma_i - \gamma_i'\right) a_{i,j}^n\right| \leq \max_{i,j\,:\,Q_{i^*,j}\neq Q_{i,j}} a_{i,j}^{\max} \left\|\gamma - \gamma'\right\|_\infty \leq K_r \left\|\gamma - \gamma'\right\|, \tag{2.48}$$

for some constant $K_r > 0$. Combining (2.41) with (2.48), we have

$$\liminf_{n\to\infty} \sum_{i\neq i^*} \gamma_i^n a_{i,j}\left(y^n\right) > 0.$$

Therefore, there exists a constant $b_r > 0$ such that

$$\sum_{i \neq i^*} \gamma_i^{n-1} a_{i,j}\left(y^n\right) \geq b_r, \qquad \sum_{i \neq i^*} \gamma_i^{n} a_{i,j}\left(y^n\right) \geq b_r$$

for all j and all n. Since $\gamma \mapsto \sum_{i \neq i^*} \gamma_i a_{i,j}\left(y^n\right)$ is an affine function, we have

$$\sum_{i \neq i^*} \gamma_i a_{i,j}\left(y^n\right) \geq b_r$$

for any $\gamma \in \mathcal{B}^n$. Now, for any $i \neq i^*$ and any $\gamma, \gamma' \in \mathcal{B}^n$, we obtain

$$\begin{aligned}
\left|\left[\nabla g\left(\gamma; y^n\right)\right]_i - \left[\nabla g\left(\gamma'; y^n\right)\right]_i\right| &= \left|\sum_{j=1}^{k} \frac{a_{i,j}\left(y^n\right)}{\sum_{i \neq i^*} \gamma_i a_{i,j}\left(y^n\right)} - \sum_{j=1}^{k} \frac{a_{i,j}\left(y^n\right)}{\sum_{i \neq i^*} \gamma_i' a_{i,j}\left(y^n\right)}\right| \\
&\leq \sum_{j=1}^{k} a_{i,j}\left(y^n\right) \frac{\left|\sum_{i \neq i^*} \gamma_i' a_{i,j}\left(y^n\right) - \sum_{i \neq i^*} \gamma_i a_{i,j}\left(y^n\right)\right|}{\left(\sum_{i \neq i^*} \gamma_i a_{i,j}\left(y^n\right)\right)\left(\sum_{i \neq i^*} \gamma_i' a_{i,j}\left(y^n\right)\right)} \\
&\leq \frac{k K_r \max_{i,j: Q_{i^*,j} \neq Q_{i,j}} a_{i,j}^{\max}}{b_r^2} \left\|\gamma - \gamma'\right\|,
\end{aligned}$$

which completes the proof. □

We now prove Theorem 2.6. Let $\bar{\gamma} \in \mathcal{G}$ be a limit point of $\{\gamma^{n-1}\}$, so that there exists a subsequence $\gamma^{n_t - 1} \to \bar{\gamma}$. We can pick h such that $\bar{\gamma}_h \geq \eta$ and $\mathcal{D}\left(\bar{\gamma}\right) = \text{Cone}\left(D^h\left(\bar{\gamma}\right)\right)$. Furthermore, there exists t_0 such that, for all $t > t_0$, this same h satisfies $\gamma^{n_t - 1} \geq \eta$. Since h^{n_t} is chosen from among all such h with equal probability, we can further extract a subsequence $\{n_u\} \subseteq \{n_t\}$ such that $h^{n_u} \equiv h$ for all sufficiently large u.

Suppose that $\bar{\gamma}$ is not a stationary point of (2.22) under the true values μ. Then there is a

feasible $\bar{d} \in D^h(\bar{\gamma})$ such that

$$\nabla g(\bar{\gamma}; \mu)^T \bar{d} < 0. \tag{2.49}$$

By the convergence of $\{\gamma^{n_u-1}\}$, we have $\bar{d} \in D^h(\gamma^{n_u-1})$ for all sufficiently large u. Since $y^n \to \mu$, we can find $c_1 > 0$ such that

$$\nabla g\left(\gamma^{n_u-1}; y^{n_u}\right)^T \bar{d} \leq -c_1 < 0 \tag{2.50}$$

for all sufficiently large u. Combining the convergence of $\{\gamma^{n_u-1}\}$ with Proposition A.1 in [44], we obtain a constant $c_2 > 0$ such that $s^{\max}\left(\bar{d}, \gamma^{n_u-1}\right) \geq c_2$ for all sufficiently large u. Then, at iteration n_u, Step 2 of Algorithm 2.2 yields

$$s^{\max}\left(d^{n_u}, \gamma^{n_u-1}\right) \nabla g\left(\gamma^{n_u-1}; y^{n_u}\right)^T d^{n_u} \leq s^{\max}\left(\bar{d}, \gamma^{n_u-1}\right) \nabla g\left(\gamma^{n_u-1}; y^{n_u}\right)^T \bar{d} \leq -c_1 c_2 < 0. \tag{2.51}$$

Define three subsequences

$$\begin{aligned}
E_1 &= \left\{ n : V^n \geq \max\left\{ -\kappa_0, -\left(\frac{\log n}{n}\right)^{1/4} \right\} \right\}, \\
E_2 &= \left\{ n : s^{\max}\left(d^n, \gamma^{n-1}\right) V^n \geq \max\left\{ -\kappa_0, -\left(\frac{\log n}{n}\right)^{1/2} \right\} \right\}, \\
E_3 &= \{1, 2, ...\} \setminus (E_1 \cup E_2).
\end{aligned}$$

Notice that the line search method is invoked, and thus γ^n is updated, only along E_3. We also

observe that

$$\liminf_{n\in E_1} s^{\max}\left(d^n,\gamma^{n-1}\right)V^n\geq 0,\qquad \liminf_{n\in E_2} s^{\max}\left(d^n,\gamma^{n-1}\right)V^n\geq 0. \tag{2.52}$$

Along E_3, we do line search, which terminates when s^n satisfies (2.31)-(2.32).

Using Lemma 2.7 and the convexity of $g\left(\cdot;y^n\right)$, we obtain

$$g\left(\gamma^{n-1}+sd^n;y^n\right)\leq g\left(\gamma^{n-1};y^n\right)+s\nabla g\left(\gamma^{n-1};y^n\right)^T d^n+\frac{s^2K_\gamma}{2}\left\|d^n\right\|_2^2.$$

Then, (2.31) will be satisfied if we choose s such that

$$g\left(\gamma^{n-1};y^n\right)+s\nabla g\left(\gamma^{n-1};y^n\right)^T d^n+\frac{\lambda^2K_\gamma}{2}\left\|d^n\right\|_2^2\leq g\left(\gamma^{n-1};y^n\right)+\alpha s\nabla g\left(\gamma^{n-1};y^n\right)^T d^n,$$

so

$$s\leq\frac{(\alpha-1)\nabla g\left(\gamma^{n-1};y^n\right)^T d^n}{K_\gamma}=\frac{(\alpha-1)V^n}{K_\gamma}.$$

Define $s^{K,n}=\frac{(\alpha-1)V^n}{K_\gamma}$ and $s^{*,n}=\arg\min_s g\left(\gamma^{n-1}+sd^n\right)$. Let $s^{W,n}$ be the largest s that satisfies (2.32). By the convexity of $g\left(\cdot;y^n\right)$, we have $s^{W,n}\geq s^{*,n}$.

Now suppose that $s^{\max}\left(d^n,\gamma^{n-1}\right)\leq\min\left\{s^{K,n},s^{W,n}\right\}$. It follows that $s^n=s^{\max}\left(d^n,\gamma^{n-1}\right)$ and

$$g\left(\gamma^{n-1};y^n\right)-g\left(\gamma^n;y^n\right)\geq-\alpha s^{\max}\left(d^n\right)V^n. \tag{2.53}$$

On the other hand, if $s^{\max}\left(d^n, \gamma^{n-1}\right) > \min\left\{s^{K,n}, s^{W,n}\right\}$, we consider two cases. In the first case, $s^{W,n} \geq s^{K,n}$, so $s^n \geq \tau s^{K,n}$, and

$$g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^n; y^n\right) \geq \frac{\alpha\tau\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma}. \tag{2.54}$$

In the second case, $s^{K,n} \geq s^{W,n}$. Then $s^n \in \left[\tau s^{*,n}, s^{W,n}\right]$. Since $g(\cdot; y^n)$ is convex, we have

$$g\left(\gamma^{n-1} + s^n d^n; y^n\right) \leq \max\left\{g\left(\gamma^{n-1} + \tau s^{*,n} d^n; y^n\right), g\left(\gamma^{n-1} + s^{W,n} d^n; y^n\right)\right\}.$$

Since

$$\begin{aligned}
& g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^{n-1} + \tau s^{*,n} d^n; y^n\right) \\
\geq\ & g\left(\gamma^{n-1}; y^n\right) - \left(\left(1-\tau\right) g\left(\gamma^{n-1}; y^n\right) + \tau g\left(\gamma^{n-1} + s^{*,n} d^n; y^n\right)\right) \\
=\ & \tau\left(g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^{n-1} + s^{*,n} d^n; y^n\right)\right) \\
\geq\ & \tau\left(g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^{n-1} + s^{K,n} d^n; y^n\right)\right) \\
\geq\ & \frac{\alpha\tau\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma},
\end{aligned}$$

and, similarly,

$$\begin{aligned}
g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^{n-1} + s^{W,n} d^n; y^n\right) &\geq g\left(\gamma^{n-1}; y^n\right) - g\left(\gamma^{n-1} + s^{K,n} d^n; y^n\right) \\
&\geq \frac{\alpha\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma},
\end{aligned}$$

we again obtain (2.54). Combining (2.53) and (2.54), we arrive at

$$g\left(\gamma^{n-1};y^n\right)-g\left(\gamma^n;y^n\right)\geq\min\left\{\frac{\alpha\tau\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma},-\alpha\tau s^{\max}\left(d^n,\gamma^{n-1}\right)V^n\right\}.$$

Combining Lemma 2.1 with the law of the iterated logarithm, we have

$$\begin{aligned}
& g\left(\gamma^{n-1};\mu\right)-g\left(\gamma^n;\mu\right)\\
= \;& g\left(\gamma^{n-1};\mu\right)-g\left(\gamma^{n-1};y^n\right)+g\left(\gamma^{n-1};y^n\right)-g\left(\gamma^n;y^n\right)+g\left(\gamma^n;y^n\right)-g\left(\gamma^n;\mu\right)\\
\geq \;& \min\left\{\frac{\alpha\tau\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma},-\alpha\tau s^{\max}\left(d^n,\gamma^{n-1}\right)V^n\right\}+\left(g\left(\gamma^{n-1};\mu\right)-g\left(\gamma^{n-1};y^n\right)\right)\\
+ \;& \left(g\left(\gamma^n;y^n\right)-g\left(\gamma^n;\mu\right)\right)\\
= \;& \min\left\{\frac{\alpha\tau\left(1-\alpha\right)\left(V^n\right)^2}{K_\gamma},-\alpha\tau s^{\max}\left(d^n,\gamma^{n-1}\right)V^n\right\}+O\left(\sqrt{\frac{\log\log n}{n}}\right).
\end{aligned}$$

By the definition of E_3, we have

$$g\left(\gamma^{n-1};\mu\right)-g\left(\gamma^n;\mu\right)\geq\Omega\left(\sqrt{\frac{\log n}{n}}\right)+O\left(\sqrt{\frac{\log\log n}{n}}\right),\qquad n\in E_3,$$

which implies that, for sufficiently large $n\in E_3$, we have $g\left(\gamma^{n-1};\mu\right)\geq g\left(\gamma^n;\mu\right)$. In other words, the effect of estimation error is removed along E_3, and the objective value becomes monotonically decreasing.

Let $w\left(n\right)=\max\left\{w_0\in E_3:w_0<n\right\}$. Since γ^n is updated only along E_3, we have

$$g\left(\gamma^{w(n)};\mu\right)=g\left(\gamma^{w(n)+1};\mu\right)=\cdots=g\left(\gamma^{n-1};\mu\right)\geq g\left(\gamma^n;\mu\right).\tag{2.55}$$

Since $g(\cdot;\mu)$ is bounded below, the sequence $\{g(\gamma^n;\mu) : n \in E_3\}$ converges. Since $g(\gamma;\mu)$ is continuous in $\gamma \in \mathcal{G}$, it follows that $\{\gamma^n : n \in E_3\}$ also converges to some limit $\hat{\gamma} \in \mathcal{G}$. Taking limits of both sides of (2.55) yields

$$\lim_{n\in E_3} g\left(\gamma^{n-1};\mu\right) - g\left(\gamma^{n-1} + s^n d^n;\mu\right) = 0.$$

Applying Lemma 2.6(ii), we have

$$\lim_{n\in E_3} s^{\max}\left(d^n, \gamma^{n-1}\right) V^n = 0. \tag{2.56}$$

From (2.52) and (2.56), we have $\liminf_{n\to\infty} s^{\max}(d^n, \gamma^{n-1}) V^n \geq 0$, which contradicts (2.51). Therefore, $\bar{\gamma}$ must be a stationary point. □

2.4 Myopic Allocation Under S-index Selection (SIMA)

In this section, we offer a different strategy for allocating the simulation budget, based on Bayesian value-of-information methodology. Since the S-index is itself a new selection criterion, there are no existing allocation methods that can serve as natural benchmarks for the procedure proposed in Section 2.3.3. The myopic Bayesian procedure developed here provides such a benchmark, but is itself a very promising method.

2.4.1 Formulation of An Myopic Allocation Policy

Let $\{W_i^n\}$ be a sequence of i.i.d. samples from the distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, representing the output of repeated replications with alternative i. This time, however, we use the Bayesian

model $\mu_i \sim \mathcal{N}\left(y_i^0, (\sigma_i^0)^2\right)$, with μ_i and μ_j independent for all $i \neq j$. Let $\{j^n\}_{n=0}^{\infty}$ be a sequence of alternatives chosen for simulation, and denote by $\mathcal{F}^n$ the σ-algebra generated by $j^0, W_{j^0}^1, ..., j^{n-1}, W_{j^{n-1}}^n$. We allow $j^n \in \mathcal{F}^n$, meaning that decisions are made adaptively. It is well known [49] that the conditional distribution of μ_i given $\mathcal{F}^n$ is $\mathcal{N}\left(y_i^n, (\sigma_i^n)^2\right)$, and the parameters can be updated recursively using

$$
\begin{aligned}
y_i^{n+1} &= \begin{cases} \frac{\left(\sigma_i^n\right)^{-2} y^n + \sigma_i^{-2} W_i^{n+1}}{\left(\sigma_i^n\right)^{-2} + \sigma_i^{-2}} & i = j^n, \\ y_i^n & i \neq j^n. \end{cases} \\
\left(\sigma_i^{n+1}\right)^2 &= \begin{cases} \left(\left(\sigma_i^n\right)^{-2} + \sigma_i^{-2}\right)^{-1} & i = j^n, \\ \left(\sigma_i^n\right)^2 & i \neq j^n. \end{cases}
\end{aligned}
$$

It can also be shown that the predictive distribution of y_i^{n+1}, conditional on $\mathcal{F}^n$, is $\mathcal{N}\left(\mu_i^n, (\tilde{\sigma}_i^n)^2\right)$, where $(\tilde{\sigma}_i^n)^2 = (\sigma_i^n)^2 - \left(\sigma_i^{n+1}\right)^2$.

As before, let $z^n = Qy^n$ be the S-indices computed from our point estimates at time n, with $i^{*,n} = \arg\max_i z_i^n$ being the alternative with the largest S-index. Define

$$
\nu_j^n = \mathbb{E}\left[\mu_{i^{*,n+1}}^{n+1} - \mu_{i^{*,n+1}}^n \mid \mathcal{F}^n, j^n = j\right], \tag{2.57}
$$

to be the expected improvement in the estimated value of alternative $i^{*,n+1}$ as a result of allocating one additional replication to alternative j. If $S = 0$ and $Q = I$, the criterion in (2.57) becomes very similar to the knowledge gradient criterion of [2], with the difference that we use $i^{*,n+1}$ instead of $i^{*,n}$ in the second term to simplify the computation. Since we are trying to identify $\arg\max_i \mu_i$, we use μ to express the value of the selected alternative, but the index of that

alternative is now determined using S-indices. Under the normality assumption, (2.57) can be computed in closed form as

$$\nu_j^n = \tilde{\sigma}_j^n \phi\left(\max_{i \neq j} \frac{z_i^n - z_j^n}{\tilde{\sigma}_j^n \left(Q_{j,j} - Q_{j,i}\right)}\right), \tag{2.58}$$

where ϕ is the standard normal density. This formula is obtained by direct computation combined with the fact, shown in Theorem 2.1, that $Q_{j,j} > Q_{j,i}$ for $j \neq i$. A myopic allocation policy under S-index selection policy can thus be proposed, which allocates the next replication to alternative $j^n = \arg\max_j \nu_j^n$. We call this policy S-index Myopic Allocation (SIMA).

Specifically, when an aligned graph is available, it is possible to characterize the performance of SIMA in terms of the asymptotic sampling ratios $\lim_{n\to\infty} \frac{N_i^n}{N_j^n}$, which describe the limiting proportions $\lim_{n\to\infty} \frac{N_j^n}{n}$ of the budget allocated to each j. These proportions do not match the optimal proportions studied in Section 2.3, in keeping with previous work on ranking and selection [50], where methods based purely on expected improvement criteria also generally do not achieve optimal allocations without additional modifications [48, 51].

2.4.2 Convergence Analysis of SIMA

In this section, we give the explicit asymptotic sampling ratios of SIMA.

Lemma 2.8. *Suppose that $k \geq 3$ with $\mu_1 > \ldots > \mu_k$. Given any fixed similarity matrix S with $S_{i,j} \geq 0$ and $S_{i,j} = S_{j,i}$, $i, j = 1, \ldots, k$, the sampling ratios achieved by SIMA are given by*

$$\lim_{n\to\infty} \frac{N_i^n}{N_j^n} = \frac{c_j}{c_i},$$

where

$$c_j = \left| \max_{u \neq i} \frac{z_u - z_j}{\sigma_j (Q_{j,j} - Q_{j,u})} \right|, \tag{2.59}$$

and $z = (I + \lambda L(S))^{-1}\mu = Q(S)\mu$.

To prove Lemma 2.8, we basically follow [50]. First consider a modified version of SIMA policy, which is given by

$$\bar{\nu}_j^n = \tilde{\sigma}_j^n \phi \left(- \max_{u \neq j} \frac{z_u - z_j}{\tilde{\sigma}_j^n (Q_{j,j} - Q_{j,u})} \right), \tag{2.60}$$

where $z = Q(S)\mu$. Suppose N_j^n is the number of measurements of alternative j up to step n. Let $O_j^n = N_j^n(N_j^n + 1)$, and then we have $\tilde{\sigma}_j^n = \frac{\sigma_j}{\sqrt{O_j^n}}$. By the definition of c_j in (2.59), $\bar{\nu}_j^n$ can be written as:

$$\bar{\nu}_j^n = \frac{\sigma_j}{\sqrt{O_j^n}} \phi \left(-c_j \sqrt{O_j^n} \right). \tag{2.61}$$

Construct a continuous function

$$\bar{\nu}_j(w) = \frac{\sigma_j}{\sqrt{w}} \phi \left(-c_j \sqrt{w} \right), \quad w \geq 0. \tag{2.62}$$

When w takes value O_j^n, (2.62) becomes (2.61). Notice that the standard normal p.d.f. has the following property:

Proposition 2.3. *Fix $c_1, c_2 \geq 0$. Then*

$$\lim_{x\to\infty} \frac{\phi(-c_1\sqrt{x})}{\phi(-c_2\sqrt{x})} = \begin{cases} \infty, & c_1 < c_2 \\ 1, & c_1 = c_2 \\ 0, & c_1 > c_2 \end{cases}.$$

Proof. Since

$$\lim_{x\to\infty} \frac{\phi(-c_1\sqrt{x})}{\phi(-c_2\sqrt{x})} = \lim_{x\to\infty} \frac{e^{-c_1^2 x}}{e^{-c_2^2 x}} = \lim_{x\to\infty} e^{-(c_1^2-c_2^2)x},$$

the results follow easily. □

The remainder of the proof is very similar to [50], as we only need to replace $f(x) = x\Phi(x) + \phi(x)$ in that paper by $\phi(x)$.

Lemma 2.8 gives the asymptotic sampling ratios using SIMA for general similarity graph. Specifically, when an aligned graph is used, c_j defined in (2.59) can be computed explicitly since the maximum can be found. To obtain the closed-form results, we require the following lemma.

Lemma 2.9. *Let $\mathbb{S}^k$ be the set of all real possible similarity matrices $S(= [S_{i,j}]_{k\times k})$ for aligned graphs, $\mathbb{Y}^k = \{y \in \mathbb{R}^k | y_1 \geq y_2 \geq ... \geq y_k\}$ and $\Omega^k = \{2, ..., k\}$. Given $\lambda > 0$, $\mathcal{D}^k(S, y, \zeta, j) : \mathbb{S}^k \times \mathbb{Y}^k \times (\mathbb{R}^+)^k \times \Omega^k \mapsto \mathbb{R}$ is defined by*

$$\mathcal{D}^k(S, y, \zeta, j) = \det\left([y, W_2, ..., W_{j-1}, \mathbf{1}, W_{j+1}, ..., W_k]\right), \tag{2.63}$$

where $W_i = [-\lambda S_{i,1}, ..., -\lambda S_{i,i-1}, 1 + \lambda D_{i,i} + \lambda\zeta_i, -\lambda S_{i,i+1}, ..., -\lambda S_{i,k}]^T$ ($i \in \Omega^k \setminus \{j\}$) with

$D_{i,i} = \sum_{l \neq i} S_{i,l}$, *and* $\zeta_1 = \zeta_j = 0$. *Then for finite integer* $k \geq 3$, $\mathcal{D}^k(S, y, \zeta, j) \geq 0$ *for any* S, y, ζ, j *in its domain.*

Remark 2.1. *To clarify the slight abuse of notation, we specifically discuss the following two cases: If* $j = 2$, $\mathbf{1}$ *is right behind* y*, i.e., there is no* W_2 *in the matrix in (2.63). If* $j = k$, $\mathbf{1}$ *is the last column and there is no* W_k.

Remark 2.2. *Let* $B = I + \lambda L$*, then* $W_i = B_i + \lambda\zeta_i e_i$*, where* B_i *is the* i*th column of* B*, meaning that* W_i *is obtained by simply adding some non-negative number to the* i*th component of* B_i.

Proof. Without loss of generality, we may assume $\lambda = 1$, since for $\lambda \neq 1$, we can scale S and ζ by λ to get the results. We prove the lemma by induction. For $k = 3$, $j = 2$,

$$\mathcal{D}^3(S, y, \zeta, 2) = \begin{vmatrix} y_1 & 1 & -S_{3,1} \\ y_2 & 1 & -S_{3,2} \\ y_3 & 1 & 1 + D_{3,3} + \zeta_3 \end{vmatrix}$$

$$= (y_1 - y_3)(1 + D_{3,3} + \zeta_3 + S_{3,2}) - (y_2 - y_3)(1 + D_{3,3} + \zeta_3 + S_{3,1}) \geq 0.$$

The last inequality holds because $S_{3,2} \geq S_{3,1}$ and $y_1 \geq y_2$. For $k = 3$, $j = 3$,

$$\mathcal{D}^3(S, y, \zeta, 3) = \begin{vmatrix} y_1 & -S_{2,1} & 1 \\ y_2 & 1 + D_{2,2} + \zeta_2 & 1 \\ y_3 & -S_{2,3} & 1 \end{vmatrix}$$

$$= (y_1 - y_3)(1 + D_{2,2} + \zeta_2 + S_{2,3}) - (y_2 - y_3)(-S_{2,1} + S_{2,3}) \geq 0.$$

Suppose for $k = n-1$, the statement is true. Denote the matrix $[y, W_2, ..., W_{j-1}, \mathbf{1}, W_{j+1}, ..., W_n]$

by $A^{(n)}$. We have:

$$A^{(1)} = [y, B_2, ..., B_{j-1}, \mathbf{1}, B_{j+1}, ..., B_n],$$
$$A^{(t)} = A^{(t-1)} + \zeta_t e_t e_t^T, \quad t = 2, 3, ..., n,$$

where B_i, $i = 2, .., j-1, j+1, ..., n$, is the ith column of $B = I + L$ associated with similarity matrix S. From Theorem 2.2, we know there exists $z \in \mathbb{Y}^n$ such that

$$A^{(1)} = B[z, e_2, ..., e_{j-1}, \mathbf{1}, e_{j+1}, ..., e_n].$$

Therefore $\det(A^{(1)}) = \det(B)(z_1 - z_j) \geq 0$. By the matrix determinant lemma [52], for $t = 2, ..., j-1, j+1, ..., n$,

$$\det\left(A^{(t)}\right) = \det\left(A^{(t-1)}\right) + \zeta_t e_t^T \left(\text{adj}\left(A^{(t-1)}\right)\right) e_t = \det\left(A^{(t-1)}\right) + \zeta_t \left(\text{adj}\left(A^{(t-1)}\right)\right)_{t,t}, \tag{2.64}$$

where $\left(\text{adj}\left(A^{(t-1)}\right)\right)_{t,t}$ is the (t,t)th component of the adjoint matrix of $A^{(t-1)}$. Notice that the (t,t)th component is on the diagonal of adj $\left(A^{(t-1)}\right)$, therefore $\left(\text{adj}\left(A^{(t-1)}\right)\right)_{t,t} = \det\left(C^{(t)}\right)$, where

$$C^{(t)} = \left[y^{(t)}, W_2^{(t)}, ..., W_{t-1}^{(t)}, B_{t+1}^{(t)}, ..., B_{j-1}^{(t)}, \mathbf{1}, B_{j+1}^{(t)}, ..., B_k^{(t)}\right] \quad \text{for } t \leq j-2,$$
$$C^{(t)} = \left[y^{(t)}, W_2^{(t)}, ..., W_{t-1}^{(t)}, \mathbf{1}, B_{j+1}^{(t)}, ..., B_k^{(t)}\right] \quad \text{for } t = j-1,$$
$$C^{(t)} = \left[y^{(t)}, W_2^{(t)}, ..., W_{j-1}^{(t)}, \mathbf{1}, B_{t+1}^{(t)}, ..., B_k^{(t)}\right] \quad \text{for } t = j+1,$$
$$C^{(t)} = \left[y^{(t)}, W_2^{(t)}, ..., W_{j-1}^{(t)}, \mathbf{1}, W_{j+1}^{(t)}, ..., W_{t-1}^{(t)}, B_{t+1}^{(t)}, ..., B_k^{(t)}\right] \quad \text{for } t \geq j+2,$$

and $y^{(t)}$, $W_i^{(t)}$, $i \leq t-1$, and $B_i^{(t)}$, $i \geq t+1$, are obtained by deleting the tth component of y, W_i and B_i, respectively. Then,

$$\begin{aligned} W_{i,i}^{(t)} =& W_{i,i} = 1 + \sum_{l \neq i}^{n} |W_{i,l}| + \zeta_i = 1 + \sum_{l \neq i}^{n-1} |W_{i,l}^{(t)}| + S_{i,t} + \zeta_i, \quad i = 2, ..., t-1, \\ B_{i,i-1}^{(t)} =& B_{i,i} = 1 + \sum_{l \neq i}^{n} |B_{i,l}| = 1 + \sum_{l \neq i-1}^{n-1} |B_{i,l}^{(t)}| + S_{i,t}, \quad i = t+1, ..., k. \end{aligned} \tag{2.65}$$

Here $W_{i,i}^{(t)}$ and $B_{i,i-1}^{(t)}$ are on the diagonal of $C^{(t)}$. Denote the matrix obtained by deleting the tth column and row of S by $S_{(n-1)\times(n-1)}^{(t)}$. And let $\zeta^{(t)}$ be a $n-1$ dimensional vector with $\zeta_1^{(t)} = 0$. If $t \leq j-1$, let $\zeta_{j-1}^{(t)} = 0$; otherwise, let $\zeta_j^{(t)} = 0$. All the other components of $\zeta^{(t)}$ are set to be $C_{i,i}^{(t)} - \left(1 + \sum_{l \neq i}^{n-1} C_{i,l}^{(t)}\right)$, which are always non-negative from (2.65). Define $j^{(t)} = j-1$ if $t \leq j-1$; $j^{(t)} = j$, otherwise. It's easy to see $(S_{(n-1)\times(n-1)}^{(t)}, y^{(t)}, \zeta^{(t)}, j^{(t)}) \in dom(\mathcal{D}^{n-1})$ and $\det(C^{(t)}) = \mathcal{D}^{n-1}\left(S_{(n-1)\times(n-1)}^{(t)}, y^{(t)}, \zeta^{(t)}, j^{(t)}\right)$. By our assumption that the statement is true for $k = n-1$, we have

$$\left(\text{adj}\left(A^{(t-1)}\right)\right)_{t,t} = \det(C^{(t)}) \geq 0.$$

From (2.64), we know that if $\left(\text{adj}\left(A^{(t-1)}\right)\right)_{t,t} \geq 0$ and $\det\left(A^{(t-1)}\right) \geq 0$, then $\det\left(A^{(t)}\right) \geq 0$. Since $\det\left(A^{(1)}\right) \geq 0$, by induction, we can get that for $k = n$ and any $(S, y, \zeta, j) \in dom(\mathcal{D}^n)$,

$$\mathcal{D}^n(S, y, \zeta, j) = \det\left(A^{(n)}\right) \geq 0.$$

The proof is complete. □

With Lemma 2.9, we can further simplify (2.59), and give the following result.

Lemma 2.10. *Given an aligned graph, suppose* $k = 3, 4, 5, ...,$ *then*

$$\arg\max_{u \neq 1} \frac{z_u - z_1}{Q_{1,1} - Q_{1,u}} = 2,$$
$$\arg\max_{u \neq j} \frac{z_u - z_j}{Q_{j,j} - Q_{j,u}} = 1, \quad j = 2, ..., k.$$

Proof. For $j \neq 1$, if $u > j$, since $z_1 \geq z_j \geq z_u$ and $Q_{j,j} > \max\{Q_{j,1}, Q_{j,u}\}$, we can easily obtain $\frac{z_1 - z_j}{Q_{j,j} - Q_{j,1}} \geq \frac{z_u - z_j}{Q_{j,j} - Q_{j,u}}$. For $1 < u < j$, let $A = [z, e_2, ..., e_{u-1}, \mathbf{1}, e_{u+1}, ...e_{j-1}, Q_j, e_{j+1}, ..., e_k]$. Then $\frac{z_1 - z_j}{Q_{j,j} - Q_{j,1}} \geq \frac{z_u - z_j}{Q_{j,j} - Q_{j,u}}$ is equivalent to $\det(A) \geq 0$. Let $B = I + \lambda L$ and construct Υ by

$$\Upsilon = [\mu, B_2, ..., B_{u-1}, \mathbf{1}, B_{u+1}, ..., B_{j-1}, e_j, B_{j+1}, ..., B_k].$$

Since $z = Q\mu$, $\mathbf{1} = Q\mathbf{1}$ and $e_i = QB_i$, we have $A = (I + \lambda L)^{-1}\Upsilon = Q\Upsilon$. Suppose $\Upsilon^{(j)}_{(k-1)\times(k-1)} = [\mu^{(j)}, B_2^{(j)}, ..., B_{u-1}^{(j)}, \mathbf{1}, B_{u+1}^{(j)}, ..., B_{j-1}^{(j)}, B_{j+1}^{(j)}, ..., B_k^{(j)}]$ is the matrix obtained by deleting the jth column and row of Υ, then $\det(\Upsilon) = \det(\Upsilon^{(j)})$. By Lemma 2.9, we have $\det(\Upsilon^{(j)}) \geq 0$. Furthermore, $\det(A) = \det(Q)\det(\Upsilon) = \det(Q)\det(\Upsilon^{(j)}) \geq 0$.

For $j = 1$, let $A = [Q_1, z, e_3, ..., e_{u-1}, \mathbf{1}, e_{u+1}, ..., e_k]$. Then $\frac{z_2 - z_1}{Q_{1,1} - Q_{1,2}} \geq \frac{z_u - z_1}{Q_{1,1} - Q_{1,u}}$ is equivalent to $\det(A) \geq 0$. Now construct Υ by

$$\Upsilon = [e_1, \mu, B_3, ..., B_{u-1}, \mathbf{1}, B_{u+1}, ..., B_k],$$

Delete the first row and column of Υ, we obtain

$$\Upsilon^{(1)}_{(k-1)\times(k-1)} = [\mu^{(1)}, B_3^{(1)}, \ldots, B_{u-1}^{(1)}, \mathbf{1}, B_{u+1}^{(1)}, \ldots, B_k^{(1)}],$$

then $\det(\Upsilon) = \det(\Upsilon^{(1)}) \geq 0$ by Lemma 2.9 again. Finally, we have

$$\det(A) = \det(Q)\det(\Upsilon) \geq 0.$$

The proof is complete. □

With Lemmas 2.8 and 2.10, the explicit asymptotic sampling ratios of SIMA with an aligned graph can be derived.

Theorem 2.7. *Suppose that $k \geq 3$ with $\mu_1 > ... > \mu_k$ for simplicity. Given an aligned graph, the sampling ratios achieved by SIMA are given by*

$$\lim_{n\to\infty} \frac{N_i^n}{N_j^n} = \frac{\sigma_i\left(\frac{z_1 - z_j}{Q_{j,j} - Q_{j,1}}\right)}{\sigma_j\left(\frac{z_1 - z_i}{Q_{i,i} - Q_{i,1}}\right)}, \quad i, j \neq 1, \tag{2.66}$$

$$\lim_{n\to\infty} \frac{N_1^n}{N_2^n} = \frac{\sigma_1\left(Q_{1,1} - Q_{1,2}\right)}{\sigma_2\left(Q_{2,2} - Q_{2,1}\right)}, \tag{2.67}$$

where $z = Q\mu$ are the S-indices computed from the true values.

2.5 Dynamic Updating of the Similarity Graph

Up to this point, we have assumed the availability of a suitable similarity graph (e.g., an aligned graph). In practice, one might construct a graph S^0 from relevant past data before conducting any new replications; see Section 2.7 for examples of how this might be done in specific applications. In any case, however, such a graph is unlikely to be aligned with respect to the true, unknown values μ. To avoid misdirecting the allocation procedure with inaccurate similarity information, it is desirable to update S^0 over time as new information is acquired. At

the same time, the computational cost of the update is a concern, because our algorithms rely heavily on the inverse matrix Q, and repeated computation of the inverse would compromise the computational efficiency of SIGD and SIMA.

We propose a simple updating scheme based on the observation that, given *any* S^0, it is possible to create an aligned graph S by simply permuting the rows and columns of S^0. This directly follows from Definition 2.1. Moreover, the matrix $Q = Q(S;\lambda)$ can also be computed by applying the same permutation operations on $Q^0 = Q(S^0;\lambda)$ instead of recomputing the inverse. Thus, the appeal of our scheme is twofold: it makes the similarity graph aligned (with sufficiently many samples) in a computationally efficient manner. Although we have mentioned before that an aligned graph does not guarantee $\text{PCS}(z) \geq \text{PCS}(y)$, it does ensure that S-index selection will reliably identify the best alternative asymptotically.

Formally, let r be an ordering of the alternatives, i.e., a permutation of the integers $1, ..., k$ with r_i indicating the position of alternative i in the ordering. Given two such orderings r, r', let $H(r, r')$ be a $k \times k$ matrix whose (j, j')th components are equal to 1 if there exists $i \in \{1, \ldots, k\}$ satisfying $r_j = r'_{j'} = i$, and zero otherwise. Such an H is a permutation matrix. Given the estimated values y^n at time n, let $\hat{r}^n$ be a ranking of the values y^n in descending order, i.e, a larger value of $\hat{r}^n_i$ means alternative i has a smaller estimated mean and is ranked lower. Suppose that we then sample alternative j^n. Then, its estimated value will be updated to $y^{n+1}_{j^n}$, and its position in the ranking will change to $\hat{r}^{n+1}_{j^n}$. If $\hat{r}^{n+1}_{j^n} \geq \hat{r}^n_{j^n}$, the other alternatives $i \neq j^n$ will be

reranked according to the rule

$$\hat{r}_i^{n+1} = \begin{cases} \hat{r}_i^n, & \text{if } \hat{r}_i^n > \hat{r}_{j^n}^{n+1} \text{or } \hat{r}_i^n < \hat{r}_{j^n}^n, \\ \hat{r}_i^n - 1, & \text{otherwise.} \end{cases} \tag{2.68}$$

In other words, the ranking can be updated without a sorting algorithm. A similar updating rule can be derived when $\hat{r}_{j^n}^{n+1} \leq \hat{r}_{j^n}^n$. Once $\hat{r}^{n+1}$ is computed, a permutation matrix $\hat{H}^{n+1} = H\left(\hat{r}^0, \hat{r}^{n+1}\right)$ can be created. It is straightforward to show that the similarity matrix $\hat{S}^{n+1} = \left(\hat{H}^{n+1}\right)^\top S^0 \hat{H}^{n+1}$ is aligned with respect to y^{n+1}. Therefore, as n becomes large, $\hat{S}^n$ will converge to a graph that is aligned with respect to the true means μ, provided that each alternative is sampled infinitely often.

In practice, we may not wish to update the similarity matrix at every iteration, especially in the early stages where our estimated values are subject to high uncertainty. Furthermore, even if S^0 is not aligned, it may still contain some valuable information about the similarities between different values. Thus, we may wish to hold off on permuting S^0 until there is strong evidence in favor of doing so. To that end, we define another ranking $\tilde{r}^n$ which is only updated when there is a "significant" change in the ordering of the values y^n. Once $\tilde{r}^n$ is created, the similarity matrix is determined. We initialize $\tilde{r}^n$ as $\tilde{r}^0 = \hat{r}^0$. The change in $\hat{r}^n$ gives guidance on how we should change $\tilde{r}^n$, but as we discussed earlier, we only follow this guidance when we have high confidence. Therefore, we propose the following updating scheme of $\tilde{r}^n$. Again, suppose that we sample j^n at time n and find that $\hat{r}_{j^n}^{n+1} > \hat{r}_{j^n}^n$, suggesting that j^n should possibly be ranked lower. At this point we need to decide whether we would like to trust this information and increase $\tilde{r}_{j^n}^n$

accordingly. Define two sets

$$\begin{aligned}\tilde{\Omega}^{-,n} &= \{i = 1, \ldots, k : \tilde{r}_i^n > \tilde{r}_{j^n}^n\}, \\ \hat{\Omega}^{+,n+1} &= \{i = 1, \ldots, k : y_i^{n+1} > y_{j^n}^{n+1}\} = \{i = 1, \ldots, k : \hat{r}_i^{n+1} < \hat{r}_{j^n}^{n+1}\}.\end{aligned} \tag{2.69}$$

The set $\tilde{\Omega}^{-,n}$ contains alternatives that were ranked lower than j^n according to $\tilde{r}^n$, and $\hat{\Omega}^{+,n+1}$ contains alternatives whose sample means are above j^n after the update. For $i \in \tilde{\Omega}^{-,n} \cap \hat{\Omega}^{+,n+1}$, the probability that $\mu_i < \mu_{j^n}$ can be approximated using a normal distribution, and if this probability is smaller than some small tolerance parameter $\kappa^{tol} < 0.5$, we have sufficient confidence to rank i more highly than j^n. We can reduce κ^{tol} if we wish to update the ranking more conservatively. Formally, we define

$$\tilde{\Omega}^{n+1} = \left\{ i \in \tilde{\Omega}^{-,n} \cap \hat{\Omega}^{+,n+1} : \Phi\left(\frac{y_{j^n}^{n+1} - y_i^{n+1}}{\sqrt{\frac{\sigma_i^2}{N_i^{n+1}} + \frac{\sigma_{j^n}^2}{N_{j^n}^{n+1}}}} \right) < \kappa^{tol} \right\}, \tag{2.70}$$

where $\Phi(\cdot)$ is the standard normal cdf. Then, the new rank of alternative j^n is set according to

$$\tilde{r}_{j^n}^{n+1} = \max_{i \in \tilde{\Omega}^{n+1}} \hat{r}_i^{n+1}, \tag{2.71}$$

so alternative j^n is moved to the end of those alternatives that are sufficiently likely to be better based on both the updated sample means and the previous ranking. For alternatives $i \neq j^n$, their belief ranks can be updated following (2.68) by replacing $\hat{r}$ with $\tilde{r}$.

Algorithm 2.4: Sequential update of Q

Input: Q^0 associated with prior similarity matrix S^0, estimated means y^n, alternative j^n sampled at step n and its new estimated mean $y_{j^n}^{n+1}$, previous orders $\hat{r}^n$ and $\tilde{r}^n$, tolerance parameter κ^{tol}.

Output: updated orders $\hat{r}^{n+1}$ and $\tilde{r}^{n+1}$, updated matrix $\tilde{Q}^{n+1}$ at step $n+1$.

Find $\hat{r}_{j^n}^{n+1}$ by a binary search of y^{n+1};
Update $\hat{r}^{n+1}$ by (2.68);
$\tilde{r}_{j^n}^{n+1} \leftarrow \tilde{r}_{j^n}^{n}$;
if $\hat{r}_{j^n}^{n+1} > \hat{r}_{j^n}^{n}$ *or* $(\hat{r}_{j^n}^{n+1} = \hat{r}_{j^n}^{n}$ *and* $\hat{r}_{j^n}^{n+1} > \tilde{r}_{j^n}^{n})$ **then**
  Find $\tilde{\Omega}^{n+1}$ by (2.69) and (2.70);
  $\tilde{r}_{j^n}^{n+1} \leftarrow \max_{i \in \tilde{\Omega}^{n+1}} \hat{r}_i^{n+1}$ if $\tilde{\Omega}^{n+1}$ is not empty;
else
  if $\hat{r}_{j^n}^{n+1} < \hat{r}_{j^n}^{n}$ *or* $(\hat{r}_{j^n}^{n+1} = \hat{r}_{j^n}^{n}$ *and* $\hat{r}_{j^n}^{n+1} < \tilde{r}_{j^n}^{n})$ **then**
    Find $\tilde{\Omega}^{n+1}$ by (2.72);
    $\tilde{r}_{j^n}^{n+1} \leftarrow \min_{i \in \tilde{\Omega}^{n+1}} \hat{r}_i^{n+1}$ if $\tilde{\Omega}^{n+1}$ is not empty;
  end
end
Update $\tilde{r}^{n+1}$ similar to (2.68);
$\tilde{H}^{n+1} \leftarrow H\left(\tilde{r}^0, \tilde{r}^{n+1}\right)$;
$\tilde{Q}^{n+1} \leftarrow \left(\tilde{H}^{n+1}\right)^\top Q^0 \tilde{H}^{n+1}$.

For the case where $\hat{r}_{j^n}^{n+1} < \hat{r}_{j^n}^{n}$, we can use symmetric computations based on the definitions

$$
\begin{aligned}
&\tilde{\Omega}^{+,n} = \{i = 1, \ldots, k : \tilde{r}_i^n < \tilde{r}_{j^n}^n\}, \\
&\hat{\Omega}^{-,n+1} = \{i = 1, \ldots, k : y_i^{n+1} < y_{j^n}^{n+1}\} = \{i = 1, \ldots, k : \hat{r}_i^{n+1} > \hat{r}_{j^n}^{n+1}\}, \\
&\tilde{\Omega}^{n+1} = \left\{ i \in \tilde{\Omega}^{+,n} \cap \hat{\Omega}^{-,n+1} : \Phi\left(\frac{y_i^{n+1} - y_{j^n}^{n+1}}{\sqrt{\frac{\sigma_i^2}{N_i^{n+1}} + \frac{\sigma_{j^n}^2}{N_{j^n}^{n+1}}}} \right) < \kappa^{tol} \right\}.
\end{aligned}
\tag{2.72}
$$

For complete details, refer to Algorithm 2.4, which can be called after the allocation and sampling step in either of our proposed algorithms.

Our theoretical guarantees for both SIGD and SIMA are preserved when either procedure is combined with Algorithm 2.4. This is a consequence of the following result.

Proposition 2.4. *Under any allocation method, Algorithm 2.4 will update the similarity graph finitely many times; in other words, there is a random but a.s. finite n_0 such that $\tilde{Q}^n = \tilde{Q}^{n_0}$ for $n \geq n_0$. Furthermore, if every alternative is measured infinitely often, then $\tilde{Q}^{n_0}$ must represent an aligned graph on that same sample path.*

Proof. We first argue that, under any allocation method, both $\hat{r}^n$ and $\tilde{r}^n$ will be updated finitely many times. First, we note that, under any policy, the estimated means y^n converge to a limit $\hat{y}$: if alternative i is measured infinitely often, then $\hat{y}_i = \mu_i$, otherwise y_i^n will be updated finitely many times and thus converges trivially. Then, $\hat{r}^n$ will converge to a limit $\hat{r}$ that corresponds to a ranking of the limiting values of y^n. We have $\hat{y}_i \neq \hat{y}_j$ for $i \neq j$ because the true values are distinct, and for alternatives i that are measured finitely many times, the limiting values $\hat{y}_i$ depend on normally distributed noise. Then, there exists $n_0 > 0$ such that $\hat{r}^n = \hat{r}$ for all $n > n_0$.

Next, we argue that $\tilde{r}^n$ converges. Let Ω be the set of alternatives that are sampled infinitely often, and take $i \in \Omega$ (we know that Ω must be non-empty). Let $\{n_m\}$ be the subsequence of time stages at which i is sampled.

We first consider the case $\hat{r}_i^{n_m+1} > \tilde{r}_i^{n_m}$, which is equivalent to the inequality $\hat{r}_i > \tilde{r}_i^{n_m}$ for large enough m. Because we can further take m to be large enough that no $j \in \Omega^c$ is sampled at any time $n > n_m$, the set

$$\bar{\Omega}^{n_m+1} := \left\{ j \in \hat{\Omega}^{+,n_m+1} : \Phi \left(\frac{y_j^{n_m+1} - y_i^{n_m+1}}{\sqrt{\frac{\sigma_j^2}{N_j^{n_m+1}} + \frac{\sigma_i^2}{N_i^{n_m+1}}}} \right) < \kappa^{tol} \right\}$$

will cease to be updated (i.e., will contain the same elements) after some sufficiently large m. Moreover, if $\tilde{\Omega}^{n_m+1} \neq \emptyset$, $\tilde{\Omega}^{-,n_m+1} \subsetneq \tilde{\Omega}^{-,n_m}$ by the rank updating step (2.71). Therefore,

eventually, $\tilde{\Omega}^{n_m+1}$ will become empty, meaning that $\tilde{r}_i$ will not be updated. The case where $\hat{r}_i < \tilde{r}_i^{n_m}$ can be handled symmetrically. Since the rank of alternatives in Ω^c can only change when there is a change in the rank of alternatives in Ω, it follows that $\tilde{r}_i^n$ will also be updated finitely many times for $i \in \Omega^c$. Therefore, for all i, the ranking $\tilde{r}_i^n$ will be updated only finitely many times.

Now suppose that all of the alternatives are sampled infinitely often. Clearly, $y^n \to \mu$ and $\hat{r}^n \to r$, where r is the true ranking of the alternatives based on the values μ. We also know that there exists $\tilde{r}$ and some n_1 such that, for any $n > n_1$, $\tilde{r}^n = \tilde{r}$. Consider an arbitrary alternative j and suppose that $\tilde{r}_j < \hat{r}_j$. Then there must exist an alternative i, such that $\tilde{r}_j < \tilde{r}_i$ and $\hat{r}_j > \hat{r}_i$. The latter inequality implies $\mu_j < \mu_i$. For all large enough $n > n_1$, we then have

$$\Phi\left(\frac{y_j^{n+1} - y_i^{n+1}}{\sqrt{\frac{\sigma_i^2}{N_i^{n+1}} + \frac{\sigma_j^2}{N_j^{n+1}}}}\right) < \kappa^{tol}.$$

By the updating step (2.71), we should then change the ranking so that $\tilde{r}_j^n > \tilde{r}_i^n$, contradicting the fact that $\tilde{r}^n$ is no longer updated for $n > n_1$. Therefore, the assumption that $\tilde{r}_j < \hat{r}_j$ cannot hold. Similar arguments can be made for the case where $\tilde{r}_j > \hat{r}_j$. Therefore, $\tilde{r} = r$, which means that the limit of $\tilde{S}^n$ must be aligned with respect to the true means μ. □

By Proposition 2.4, there is always a finite time n_0 after which the allocation method runs under the same matrix $\tilde{Q}^{n_0}$. The behavior of SIGD and SIMA after time n_0 is identical to the situation where these algorithms are initialized with y^{n_0} at time 0 and run under the fixed matrix $\tilde{Q}^{n_0}$. Both methods are guaranteed to sample each alternative infinitely often under any fixed Q, so they will be consistent. Therefore, $\tilde{Q}^{n_0}$ must be aligned, from which the results of Theorems

2.6 and 2.7 are easily recovered. As a result, SIGD and SIMA can perform well even in settings where the initial similarity structure S^0 is uninformative or misleading, as will be demonstrated in our numerical experiments.

2.6 Handling Alternatives with Equal Values

We provide some additional discussions about the case where there exist alternatives sharing the same value, i.e., $\mu_i = \mu_j$ for some $i \neq j$. For this special case, the definition of an aligned graph is adapted as follows.

Definition 2.2. *A similarity graph S is aligned if, for any $i = 1, \ldots, k$, we have $S_{i,j} \leq S_{i,m}$ for $j < m < \min\{l \leq i : \mu_l = \mu_i\}$ as well as all $j > m > \max\{l \geq i : \mu_l = \mu_i\}$.*

Essentially, Definition 2.2 ensures that the similarity relationships satisfy the monotonicity requirement of Definition 2.1 for all alternatives whose values are not tied. When there are multiple alternatives with the same value, their similarity relationships can have any arbitrary ordering. The result of Theorem 2.2 is preserved under Definition 2.2 if we neglect the relative order of S-indices between alternatives that have the same true mean. Thus, when y^n converges to μ, S-index selection will choose an alternative i^* satisfying $\mu_{i^*} = \max_{i=1,\ldots,k} \mu_i$. There may be multiple alternatives satisfying this condition.

Moreover, in the dynamic updating strategy of the similarity matrix S^n, to prevent permuting S infinitely many times, we can set $\hat{r}_i^n < \hat{r}_j^n$ only when $y_i^n - y_j^n > C_d\sqrt{\frac{\log n}{n}}$ for some fixed $C_d > 0$. By the law of the iterated logarithm, for any sample path, the similarity matrix will be updated finitely many times. Consequently, we will obtain an aligned graph in the sense of Definition 2.2 after finitely many time steps, on the condition that each alternative is sampled

infinitely often on that same sample path.

2.7 Numerical Experiments

In this section, we present numerical comparisons on four test problems. The following allocation methods were implemented:

- *SIGD*. We use Algorithm 2.2 to learn the optimal allocation. At the beginning of the sequential allocation process, we run additional iterations of the coordinate descent algorithm to obtain a good starting point.

- *SIMA*. The nth replication is allocated to alternative $\arg\max_j \nu_j^n$, where ν_j^n is given by (2.58).

- *Correlated knowledge gradient (CKG)* of Frazier et al. [3]. This method, a standard benchmark in the literature, models similarity between alternatives using correlated Bayesian beliefs.

- *Optimal computing budget allocation (OCBA)*. The fully sequential OCBA method of Chen et al. [1] is another standard benchmark from the ranking and selection literature.

Each method is tested under two different selection criteria, namely S-indices and either sample means (for SIGD and OCBA) or posterior means (for SIMA and CKG). In this way, CKG can benefit from the correlated Bayesian learning model described in [3], which is not used by any other method. Comparing these criteria to S-indices allows us to assess the value of using the S-index for selection separately from any one particular choice of allocation method. Performance is measured in two ways, namely PCS and the expected opportunity cost (EOC),

defined as EOC $= \max_j \mu_j - \mathbb{E}\mu_{i^M}$, where i^M is the selected alternative when the total budget is M. In general, allocation procedures use one of the two metrics in their derivation: in the above list, SIGD and OCBA are derived to optimize PCS, while SIMA and CKG are specialized for EOC. Empirically, however, we find that in most of the experiments, each policy performs similarly (relative to the others) under either criterion. Thus, though there is a version of OCBA specialized for EOC [21], we did not implement it in these experiments. Both performance metrics are estimated by averaging over 10000 independent replications of each method. To estimate the variance of each metric and calculate confidence intervals, we divide the replications into 40 sets ("macro-replications"), each containing 250 "micro-replications."

In all of the experiments, the standard deviations σ_i of the simulation output were revealed to all competing methods. We used $\lambda = 2$ as the regularization parameter and $\kappa^{tol} = 0.05$ as the tolerance parameter; in Section 2.7.5, we present additional numerical results showing that performance is relatively insensitive to both parameters.

2.7.1 Experiment 1: A Servo System Selection Problem

A servo system is an automatic control system where the controller provides commands to activate motions [53]. A critical characteristic of a servo system is the rise time, which is the time required for the signal to change to a higher position. This time depends on the mechanical configuration of the system, and may be estimated from simulation. Thus, we may consider a ranking and selection problem where different configurations are alternatives, and performance is represented by rise time.

Our test problem is adapted from a real-world dataset originating from the UCI Machine

Learning Repository [54]. The dataset contains 167 instances, each of which has a distinct configuration of 4 settings (two gain settings and two choices of mechanical linkages). Two of these settings are categorical variables, which can be handled by one-hot encoding. We randomly choose 40% of these instances to serve as the alternatives in our test problem, i.e., $k = 67$. Their recorded rise times are treated as their true μ_i values, and normally distributed noise with $\sigma_i = 0.1\mu_i + 0.1$ is added during replications.

We construct an initial similarity graph by running Gaussian process (GP) regression [55] with the remaining 60% of instances used as training data. For each instance, the values of its settings are treated as a point in multi-dimensional Euclidean space. GP regression uses the spatial proximity between the two instances to model the similarity in their rise times. Specifically, our GP prior uses the Matérn covariance kernel with parameters 5 and 2. The prior mean is initialized (before any training data are applied) to zero. These prior beliefs are then updated using the training data; for maximum accuracy, we use the `fitgpr` package in MATLAB to optimize the value of the noise parameter used in GP regression. Through the covariance kernel, the update induces posterior means for the test alternatives, which we then use as the initial estimates y^0 in our ranking and selection problem. No samples of the test alternatives were used in the computation of y^0. The posterior covariances between test alternatives obtained from GP regression are used as the prior covariances for the CKG method. For S-index selection, we construct initial similarities $S^0_{i,j} = e^{-\left|\hat{y}^0_i - \hat{y}^0_j\right|}$, where $\hat{y}^0_i = \frac{y^0_i - \frac{1}{k}\sum_{j=1}^k y^0_j}{\sqrt{\frac{\sum_{m=1}^k \left(y^0_m - \frac{1}{k}\sum_{j=1}^k y^0_j\right)^2}{k-1}}}$, essentially normalizing the distances between estimated values. This initial similarity graph is updated over time using Algorithm 2.4.

Since the alternatives used for training are chosen randomly, this setup can produce different

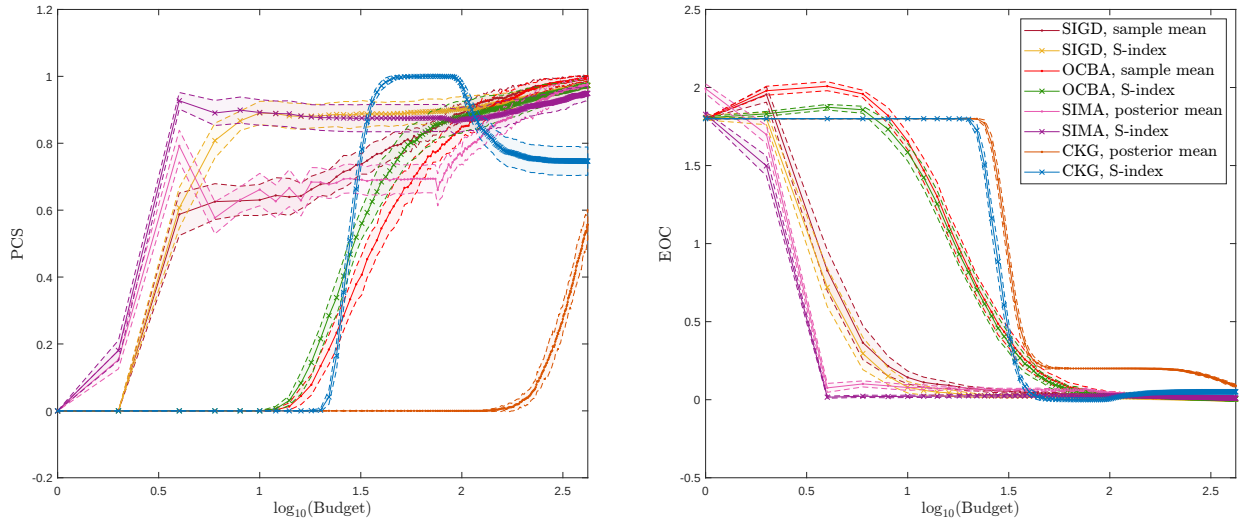

Figure 2.2: PCS and EOC results with standard error bounds obtained by 10000 independent runs (40 macro-replications, 250 micro-replications) with an unaligned prior similarity matrix.

values of y^0, which may affect performance in the subsequent ranking and selection problem. Here, we show one situation where $\arg\max_i y_i^0 \neq \arg\max_i \mu_i$, as would be the case in a difficult instance of ranking and selection. Figure 2.2 presents empirical results (averaged over 40 macro-replications) showing that, even with a simple initial similarity structure, dynamic updating via Algorithm 2.4 enables S-index selection to outperform standard selection based on sample/posterior means, regardless of which method is used to allocate the simulation budget (that is, CKG and OCBA also perform better under S-index selection). For a portion of the time horizon, CKG achieves the best performance, but experiences some degradation in later iterations. In general, PCS may not be monotonically increasing in the number of samples[56]. The EOC results suggest that CKG is getting stuck on the second-best alternative, which makes PCS drop significantly, because that metric only counts correct selection of the best. On the other hand, SIGD and SIMA consistently perform well and improve over time.

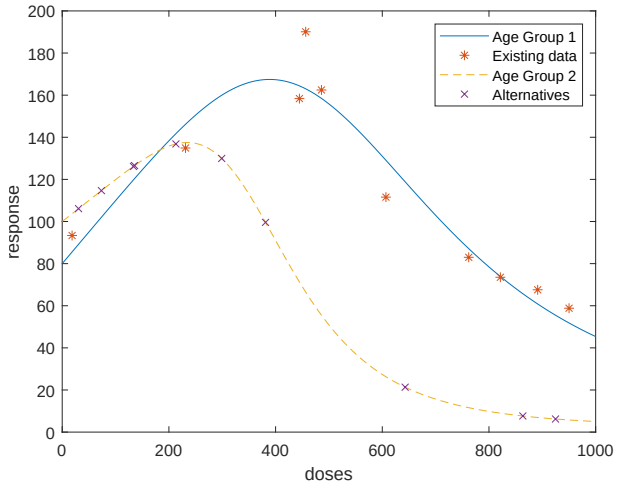

Figure 2.3: Response curves for the dose selection problem. $\chi_1(v) = \chi(v; [2, 80, 0.3, 600, 4])$, $\chi_2(v) = \chi(v; [2, 100, 0.2, 400, 5])$.

2.7.2 Experiment 2: A Dose Selection Problem

The objective of a dose-finding clinical trial is to identify the maximal response to a drug dose [57]. The literature has shown that different factors, such as gender and age, often exert great impact on the dose-response curve. Often, two such curves for different patient categories may be accurately represented using different parameter values within the same model [27].

For the purposes of this example, suppose that there are two age groups. For Age Group 1, clinical trials have been conducted for 10 doses, and noisy samples of their expected responses are available. We now wish to find the dose v_{i^*} that achieves the maximal expected response for Age Group 2. The underlying dose-response curves for both groups are assumed to follow the Brain-Cousens model $\chi(v; c) = c_1 + \frac{c_2 - c_1 + c_3 v}{\exp(1 + c_5(\log(v) - \log(c_4)))}$ with different parameters [58], as shown in Figure 2.3. We randomly choose $k = 10$ doses from $\mathcal{U}[0, 1000]$ as alternatives. Each of them has mean $\mu_i = \chi_2(v_i)$ and standard deviation $\sigma_i = 0.1\mu_i$, $i = 1, \ldots, k$.

As in Experiment 1, we use GP regression to construct initial estimates y_i^0 of v_i, $i =$

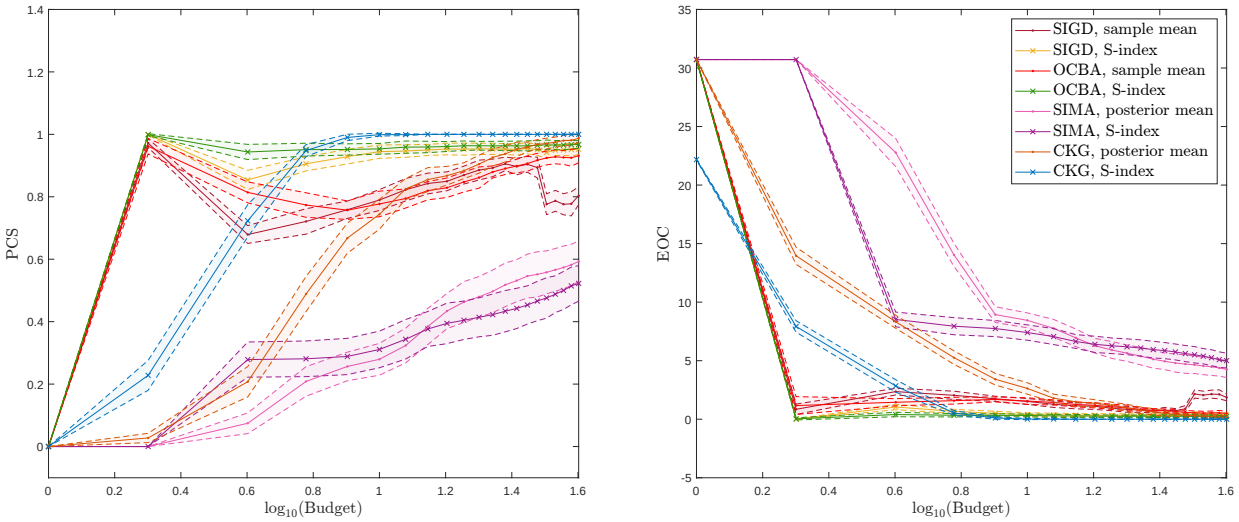

Figure 2.4: PCS and EOC results with standard error bounds for Experiment 2 obtained by 10000 independent runs (40 macro-replications, 250 micro-replications)

$1, \ldots, 10$, as well as the initial S^0 and prior covariance used by CKG. Each alternative is represented by a two-dimensional vector, where the first element is the dose and the second element is a binary value indicating the age group. The prior covariance uses a squared exponential kernel function with a separate length-scale parameter for each dimension; since we have no training data for Age Group 2, the initial length-scale parameter for the categorical feature cannot be determined and is simply set to 100. This belief model is updated using 10 observations from Age Group 1, inducing posterior beliefs about the new alternatives for Age Group 2, which are then used to initialize the ranking and selection problem as described previously.

As seen in Figure 2.4, the best performance is obtained from OCBA, SIGD, and CKG, all with S-index selection, which produces significant improvement over standard selection (sample or posterior mean) for all three methods. CKG achieves the highest PCS for moderate to large budgets, but has a longer learning curve in the beginning. OCBA has a slightly higher PCS than SIGD early on, but all three of the top-performing methods have very similar EOC in the

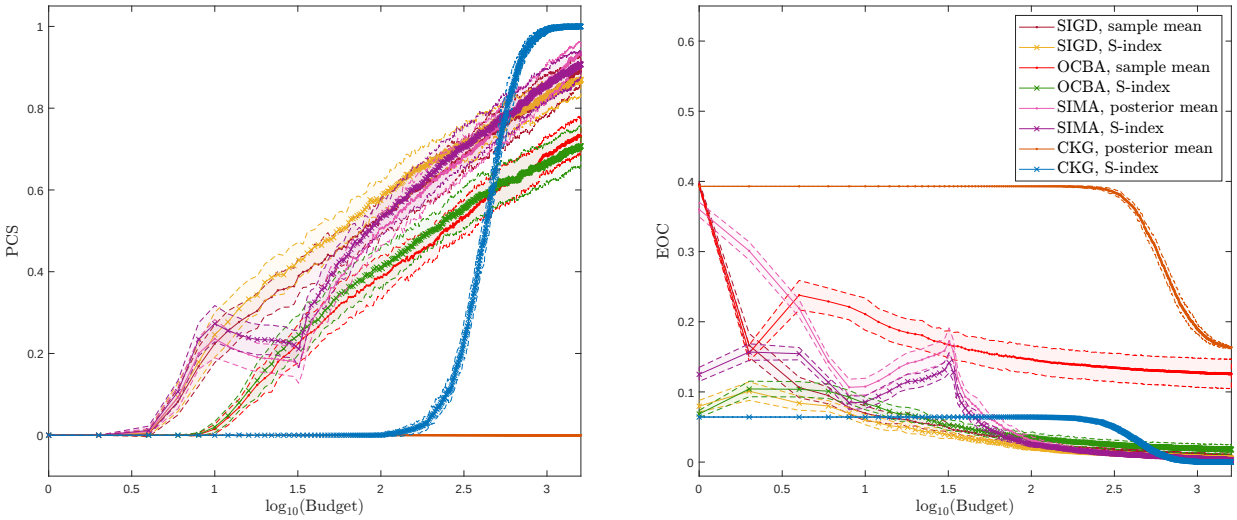

Figure 2.5: PCS and EOC results with standard error bounds for Experiment 3 obtained by 10000 independent runs (40 macro-replications, 250 micro-replications)

large-budget regime.

2.7.3 Experiment 3: M/G/1 Queues

Consider the problem of selecting, from a set of first-come-first-serve (FCFS) M/G/1 queues, the one with the largest system time for the *second* customer. The service time of each queue follows a generalized Pareto distribution, with parameter values l^P (location), σ^P (scale), and ξ^P (shape). Thus, the ith queue is characterized by a 4-dimensional vector $u_i = [\lambda_i^A, l_i^P, \sigma_i^P, \xi_i^P]^T$, where λ_i^A is an arrival rate. We consider all combinations $u_i \in \{0.3, 0.31, 0.32\} \times \{2.8, 2.9, 3.0\} \times \{0.3, 0.4\} \times \{0.3, 0.4\}$, which yields 36 alternatives. We then randomly choose 5 of them for training, leaving $k = 31$, and fit a GP regression model with a squared exponential kernel in the manner described previously. During allocation, when one unit of the sampling budget is assigned to a queue, we simulate 50 independent copies of the system time for the second customer and return their average.

Figure 2.5 presents numerical results for this problem. For very large budgets, CKG

achieves the best performance, but only when S-indices are used for selection. At the same time, the learning curve for CKG is even steeper here than in the previous problem: it has the worst performance early on, and eventually undergoes a period of very rapid improvement. By contrast, SIGD exhibits steady, consistent improvement throughout the entire time horizon, and outperforms the other methods for small- to moderate-sized budgets.

2.7.4 Experiment 4: Rosenbrock Test Function

The Rosenbrock function is a well-known benchmark in the continuous optimization literature. The function is defined on $\mathbb{R}^2$ as $f(u) = (1 - u_1)^2 + 100(u_2 - u_1^2)^2$. For the purposes of this experiment, suppose that our alternatives correspond to pairs of values taken from the set $\{0, 0.25, 0.5, 0.75, 1\}$. Thus, there are $k = 25$ alternatives; the mean of alternative i is the negative of the corresponding Rosenbrock function value, and the standard deviation of the simulation noise is given by $\sigma_i = 0.1|\mu_i| + 0.1$. We set y^0 and S^0 by running GP regression, in the same manner as before, on 9 training points randomly chosen in $[0, 1] \times [0, 1]$.

Figure 2.6 presents numerical results for this problem. In this setting, nearly every method performs very poorly at first, then improves very rapidly and settles on the best solution. The reason for this "jump" in performance is that the noise levels σ_i are relatively small compared to the differences between μ_{i^*} and μ_i for $i \neq i^*$. Initially, the procedures sample various suboptimal alternatives, and at some point (which differs by procedure) sample i^*. After this, PCS remains high because of the separation between the optimal value and suboptimal values. We observe that the jump always happens sooner when S-indices are used for selection, for any allocation strategy. SIGD and OCBA are the first to find the best, followed by SIMA and CKG.

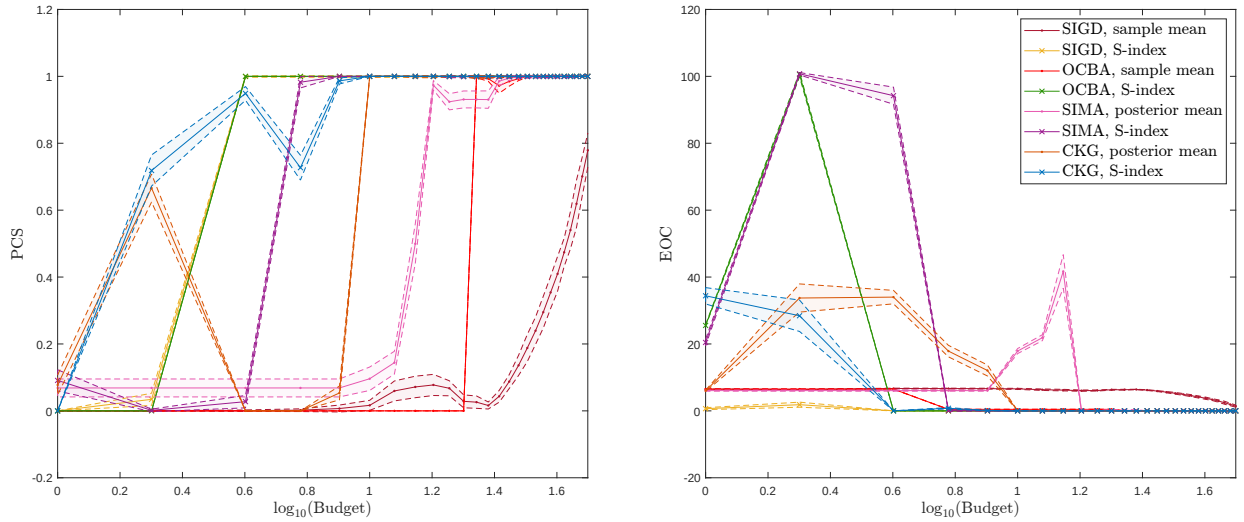

Figure 2.6: PCS and EOC results with standard error bounds for Experiment 4 based on 10000 independent replications (40 macro-replications, 250 micro-replications.)

2.7.5 Experiments on Sensitivity to Hyperparameters

We conduct experiments to test the sensitivity to λ and κ^{tol} of our proposed allocation algorithms combined with dynamic updating of the similarity structure. Since the sensitivity results for both parameters are similar for all four experiments, we show results using different values of λ for the servo system problem, and results using different values of κ^{tol} for the M/G/1 queue selection problem, as examples.

2.7.5.1 Experiments on Sensitivity to λ for the Servo System Problem

Results testing different values of λ are shown in Figures 2.7 and 2.8. From those results, we can see that performances are not very sensitive to λ, especially for large budgets.

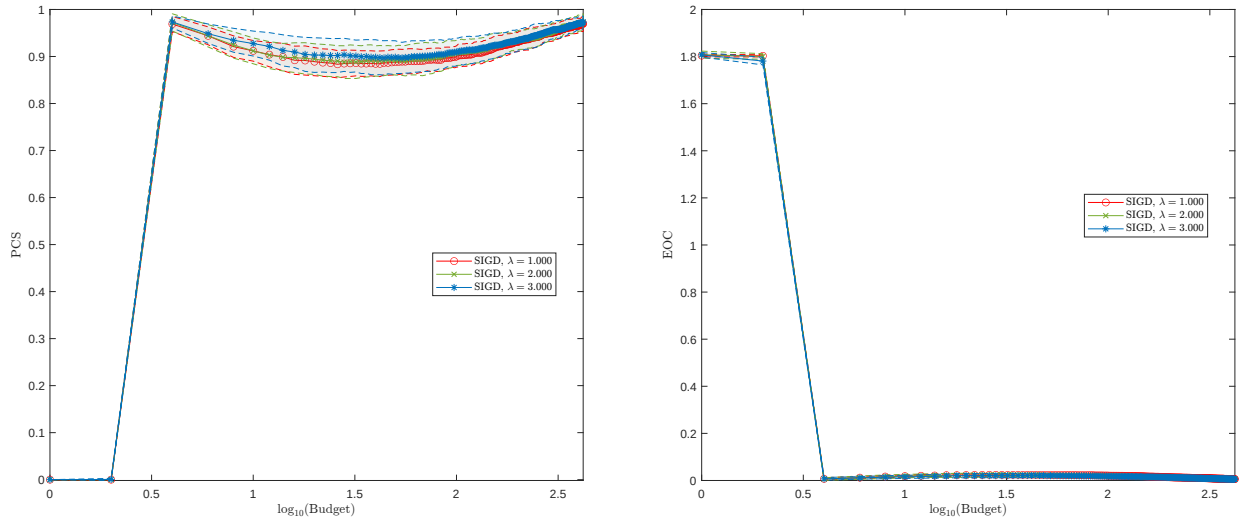

Figure 2.7: Servo System. Average PCS and EOC results for different values of λ using SIGD, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors.

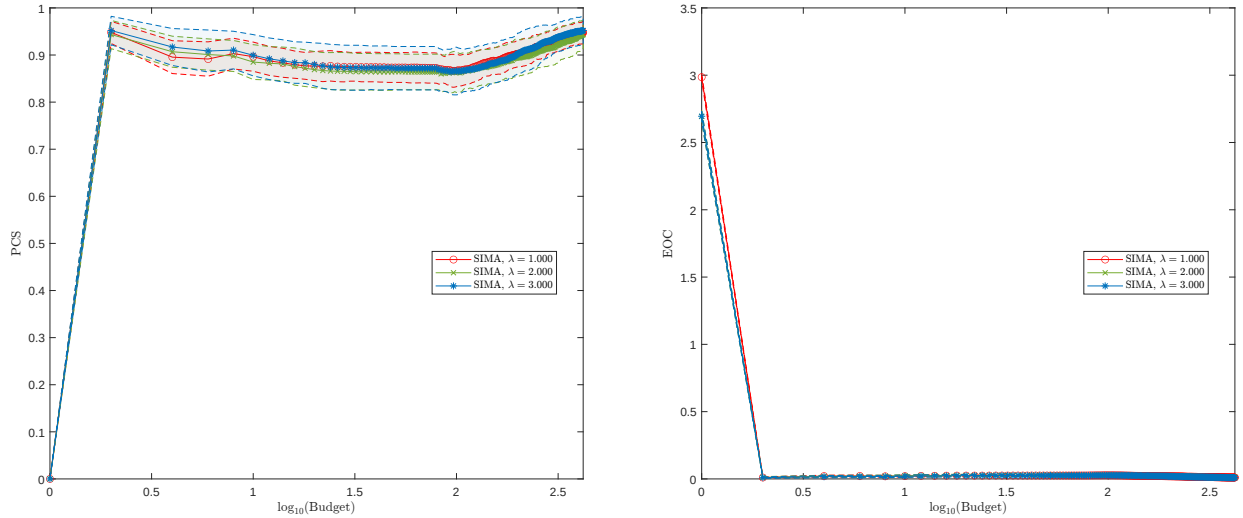

Figure 2.8: Servo System. Average PCS and EOC results for different values of λ using SIMA, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors.

2.7.5.2 Experiments on Sensitivity to κ^{tol} for the M/G/1 Queue Selection Problem

Results testing different values of κ^{tol} are shown in Figures 2.9 and 2.10. From those results, we can see that performances are not very sensitive to κ^{tol} as well.

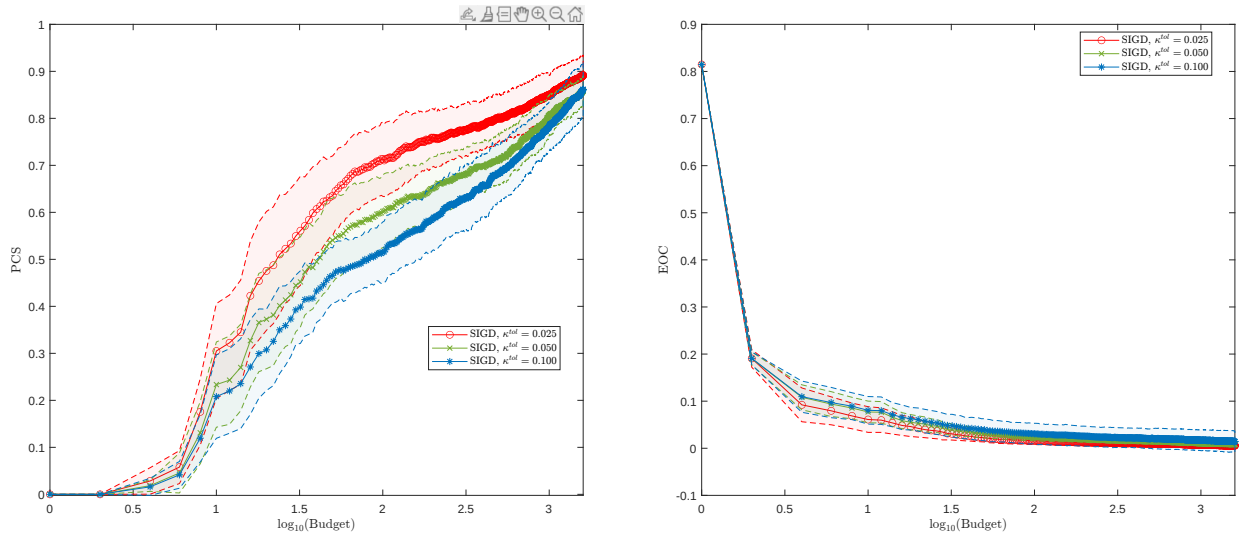

Figure 2.9: M/G/1 queue selection problem. Average PCS and EOC results for different values of κ^{tol} using SIGD, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors.

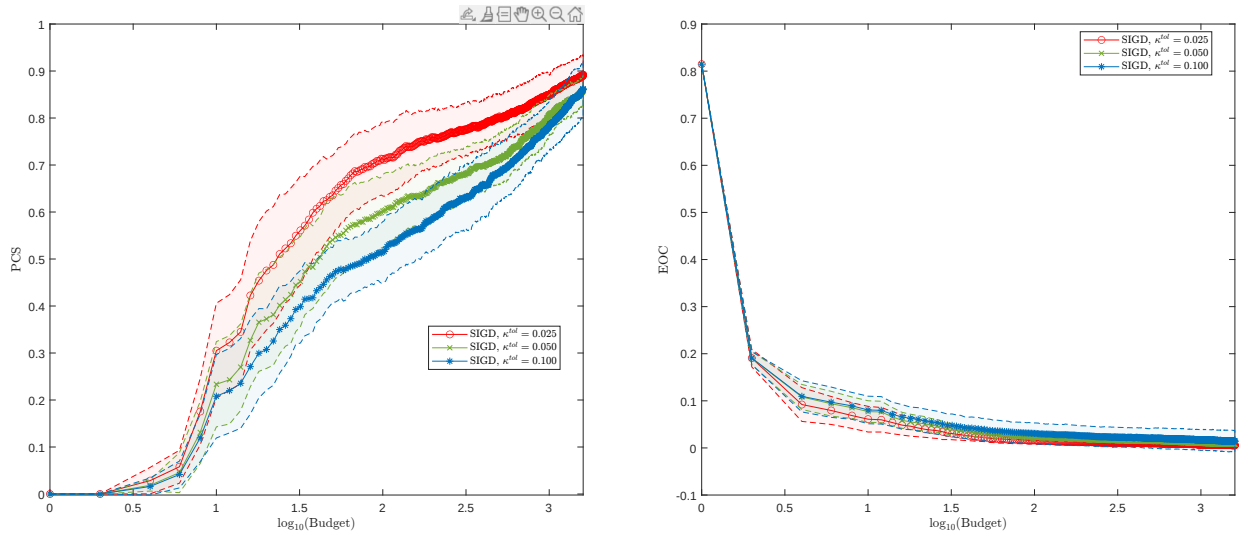

Figure 2.10: M/G/1 queues selection problem. Average PCS and EOC results for different values of κ^{tol} using SIMA, obtained by 2000 independent runs (20 macro-replications, 100 micro-replications). Dotted curves represent standard errors.

2.7.6 Summary of Results and Computation Time

Performance trajectories vary between the four test problems, with significant differences in how quickly a good solution can be obtained, but several trends can nonetheless be observed.

Table 2.1: Average computation time (ms) (including simulation) over 100 independent macro-replications.

	k	SIGD	SIMA	OCBA	CKG
Experiment 1	67	0.49	0.13	0.25	3.83
Experiment 2	10	0.22	0.10	0.11	0.69
Experiment 3	31	0.25	0.15	0.20	2.01
Experiment 4	25	0.27	0.15	0.14	1.48

First, S-index selection virtually always improves on selection based on sample/posterior means, even for those methods that were designed only for the latter criterion. Among the benchmarks, only CKG has the ability to consider similarity information, but it only uses this information for sampling, not selection. The performance of CKG can be further improved by using S-indices to incorporate similarities into selection.

SIGD is consistently among the top-performing methods, and SIMA also often performs well. Both are more computationally efficient than CKG, which uses a time-consuming sorting step with complexity $O(k \log k)$ when calculating the expected value of information in each iteration. See Table 2.1 for average computation times in each experiment, conducted on a 3.6 GHz Intel Core i7 PC with 16 GB of RAM.

2.8 Conclusion

We have presented a new framework for using similarity between the performance output of different alternatives in ranking and selection to enable more efficient identification of the best alternative. Because the S-index similarity measure is built into the selection step, we retain the computational efficiency of traditional methods that model each alternative independently of the

others. S-indices can be used when there are natural spatial similarities between the performance output of different alternatives (our numerical experiments have also focused on such problems), but are not limited to such situations, and could potentially be applied to problems where similarities come, e.g., from network structure ([59]).

We have developed two allocation methods adapted to the new selection criterion. The first is based on a mathematical program that optimizes the probability of correct selection. To avoid having to repeatedly solve the mathematical program to optimality, we have developed a novel sequential algorithm based on reduced gradient methods from numerical optimization. The second approach uses a simple myopic calculation based on the Bayesian value of information. Both approaches can be integrated with a scheme for dynamically updating the similarity structure based on new information, thus avoiding the need to rely on inaccurate or misleading prior information. Numerical experiments show that both procedures perform well with this additional subroutine.

In practice, the quality of the similarity information used to calculate S-indices is critical. Our dynamic updating scheme was effective in utilizing this information to learn the similarity (or Q) matrix, but developing ways to tailor the similarity structure to specific problem settings is key to practical implementation of the algorithms developed here. One possible avenue for future research is to investigate clustering [60] or machine learning [61] techniques for identifying similarities between alternatives.

Chapter 3: Estimating a Conditional Expectation With the Generalized Likelihood Ratio Method

3.1 Introduction

Many real-world problems, such as energy storage [62] and option pricing [63], can be modeled using stochastic processes, with parameters estimated from real-world data. When the underlying stochastic process evolves, new observations are collected, which usually enables us to make forecasts about the process in the future. One can interpret the forecasting problem as estimating the expected performance of the random system in the future, conditional on information collected up to the current period. This often involves a complicated integration. One approach to estimate the conditional expectation is to use Monte Carlo methods, i.e., simulate a number of sample paths starting from a given information set and calculate the sample average of the performance function as the estimator. However, this approach would require additional simulations, which may be time-consuming. Such an approach would be poorly suited for applications that require quick responses, such as portfolio risk management in a rapid-changing market [64]. In this chapter, we consider estimating a conditional expectation under the setting where a collection of pre-simulated sample paths is given to the user and no additional simulations should be performed. Many approaches have been proposed for this purpose. One line of research

employs least squares Monte Carlo (LSM) to estimate the conditional expectation from cross-sectional information [12]. Besides LSM, Fournié et al. [65] applied the density method (DM) and Malliavin calculus to estimate the Greeks of European options. Daveloose et al. [66] further extends their results by considering a jump-diffusion setting with both the conditional density method (CDM) and Malliavin calculus. Our focus in this chapter is to generalize the CDM method by noticing that by definition, the conditional expectation can be represented as a ratio of two derivatives that involve indicator functions. With this formulation, we are able to employ existing stochastic gradient estimation (SGE) techniques.

There are many well-known SGE techniques, among which the finite difference (FD) method may be the most straightforward and easiest to implement; however, FD estimators are biased. Two widely-used approaches for deriving an unbiased gradient estimator are infinitesimal perturbation analysis (IPA) [7] and the likelihood ratio (LR) method [8], which is also known as the Score Function (SF) method [67]. However, IPA cannot handle a discontinuous sample performance, while LR fails if the parameters of interest appear explicitly (structural parameters) in the sample performance. In our formulation, the performance functions in the corresponding SGE problems are discontinuous with structural parameters. Therefore, IPA and LR are not valid for our purpose. To overcome this shortcoming, a push-out LR technique [68] that pushes the structural parameter into the probability measure, can be applied for some problems under appropriate conditions. Recently, Peng et al. [9] developed a generalized likelihood ratio (GLR) method that extends IPA and LR, allowing discontinuities in the sample performance with the presence of structural parameters. As shown in Peng et al. [9], GLR is a generalization of the push-out LR method.

Our work shows that CDM is essentially equivalent to push-out LR. By applying GLR,

we generalize CDM under some assumptions, one of which requires the output function whose value is conditioned on to be globally invertible. We address this limitation by locally isolating the stationary points of the output function. Since our derived ratio estimator of the conditional expectation involves an indicator function, we then propose a simple way to reduce the variance of the estimator, which is particularly effective for rare events

The rest of this chapter is organized as follows. In Section 3.2, we review the problem of estimating a conditional expectation with the conditional density method under some assumptions, and introduce the stochastic gradient estimation formulation. In Section 3.3, we derive a ratio representation with GLR for two cases. We then present some numerical experiments in Section 3.4 and conclude in Section 3.5.

3.2 Estimating a Conditional Mean

In this section, we review the problem of estimating the conditional expectation with the conditional density method. We then show that the conditional expectation can be expressed as a ratio of two derivatives, which motivates the application of the generalized likelihood ratio method.

3.2.1 Problem Formulation

Consider the problem of estimating the conditional expectation:

$$\mathbb{E}\left[g_1(X,Y)|g_2(U,V)=\alpha\right], \tag{3.1}$$

where X, Y, U, V are random variables, $g_i : \mathbb{R}^2 \to \mathbb{R}$, $i = 1, 2$, are Borel-measurable functions, and α is a fixed real number. Our goal is to derive an expression for (3.1) using ordinary expectations.

This problem has been studied by Daveloose et al. [66] using conditional density methods. Here, we state their major assumptions and results, and we will show later that our approach extends their result for more complicated settings.

Assumption 3.1.

(i) (X, U) *is independent of* (Y, V)*;*

(ii) (X, U) *has joint density* $f_{X,U}(x, u)$ *with support* $\mathbb{R}^2$*;*

(iii) $\ln f_{X,U}(x, \cdot)$ *is differentiable for all* $x \in \mathbb{R}$*;*

(iv) $\mathbb{E}\left[(\pi_{X,U}(X, U))^2\right] < \infty$*, where* $\pi_{X,U}(x, u) = -\partial_u \ln f_{X,U}(x, u)$.

Assumption 3.2. *There exist a Borel-measurable function* g^* *and a strictly increasing differentiable function* h *such that* $g_2(u, v) = h^{-1}(u + g^*(v))$.

By first rewriting (3.1) as

$$\frac{\mathbb{E}\left[g_1(X, Y)\delta(g_2(U, V) - \alpha)\right]}{\mathbb{E}\left[\delta(g_2(U, V) - \alpha)\right]},$$

where δ is the Dirac delta function, and then applying the co-area formula [69] and integration by parts, Daveloose et al. [66] derived the following representation of the conditional expectation.

Theorem 3.1. *[66] Under Assumptions 3.1 and 3.2, suppose* $\mathbb{E}\left[|g_1(X, Y)|^2\right] < \infty$*, then for any*

$\alpha \in Dom(h)$,

$$\mathbb{E}\left[g_1(X,Y)|g_2(U,V)=\alpha\right]=\frac{\mathbb{E}\left[g_1(X,Y)\mathbf{1}\left\{g_2(U,V)\geq\alpha\right\}\pi_{X,U}(X,U)\right]}{\mathbb{E}\left[\mathbf{1}\left\{g_2(U,V)\geq\alpha\right\}\pi_{X,U}(X,U)\right]},$$

where $\mathbf{1}\{\cdot\}$ *is the indicator function.*

From Theorem 3.1, the conditional expectation (3.1) can be estimated by a ratio of two sample averages. However, the existence of Assumption 3.2 limits the application of the above result. In the next section, we propose an approach to overcome this limitation by applying the generalized likelihood ratio method [9].

3.2.2 Stochastic Gradient Estimation

First, we show that (3.1) can be expressed as a ratio of two derivatives. By the definition of conditional expectation, under mild regularity conditions, we have

$$\begin{aligned}
&\mathbb{E}\left[g_1(X,Y)|g_2(U,V)=\alpha\right]\\
=&\lim_{\varepsilon\to 0}\mathbb{E}\left[g_1(X,Y)|g_2(U,V)\in[\alpha-\varepsilon,\alpha-\varepsilon]\right]\\
=&\lim_{\varepsilon\to 0}\frac{\mathbb{E}\left[g_1(X,Y)\mathbf{1}\left\{g_2(U,V)\in[\alpha-\varepsilon,\alpha+\varepsilon]\right\}\right]}{\mathbb{E}\left[\mathbf{1}\left\{g_2(U,V)\in[\alpha-\varepsilon,\alpha+\varepsilon]\right\}\right]}\\
=&\frac{\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\mathbb{E}\left[g_1(X,Y)\mathbf{1}\left\{g_2(U,V)\in[\alpha-\varepsilon,\alpha+\varepsilon]\right\}\right]}{\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\mathbb{E}\left[\mathbf{1}\left\{g_2(U,V)\in[\alpha-\varepsilon,\alpha+\varepsilon]\right\}\right]}\\
=&\frac{\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\mathbb{E}\left[g_1(X,Y)\left(\mathbf{1}\left\{g_2(U,V)\leq\alpha+\varepsilon\right\}-\mathbf{1}\left\{g_2(U,V)\leq\alpha-\varepsilon\right\}\right)\right]}{\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\mathbb{E}\left[\mathbf{1}\left\{g_2(U,V)\leq\alpha+\varepsilon\right\}-\mathbf{1}\left\{g_2(U,V)\leq\alpha-\varepsilon\right\}\right]}\\
=&\frac{\frac{d}{d\theta}\mathbb{E}\left[g_1(X,Y)\mathbf{1}\left\{g_2(U,V)\leq\theta\right\}\right]\big|_{\theta=\alpha}}{\frac{d}{d\theta}\mathbb{E}\left[\mathbf{1}\left\{g_2(U,V)\leq\theta\right\}\right]\big|_{\theta=\alpha}}.
\end{aligned}\tag{3.2}$$

Using equation (3.2), we can employ SGE techniques to derive estimators of both the numerator and the denominator. The finite difference method is easy to implement, but often results in large variances. Since the performance functions in both parts are discontinuous with respect to θ, and θ appears explicitly in the indicator functions, infinitesimal perturbation analysis and the likelihood ratio method are not valid. However, for a particular g_2, a push-out LR technique can be applied to push θ into the density function by a change of variable. For example, if g_2 satisfies Assumption 3.2, $\mathbf{1}\{g_2(U,V) \leq \alpha\}$ is equivalent to $\mathbf{1}\{U \leq h(\alpha) - g^*(V)\}$. Letting $Z_\theta = \frac{U}{\theta}$, under appropriate regularity conditions, we can first derive a single-run unbiased estimator of $\frac{d}{d\theta}\mathbb{E}\left[\mathbf{1}\{Z_\theta \leq 1\}\right]\big|_{\theta=h(\alpha)-g^*(V)}$ by conditioning on V using LR and then taking expect- ation w.r.t to V. As discussed before, such methodology only works when g_2 enjoys certain properties. To deal with more complicated g_2, we apply the idea of generalized likelihood ratio (GLR) method developed by Peng et al. [9], which is shown to be a generalization of the push-out LR when the latter can be applied.

3.3 Main Results

Before we introduce the main results, we clarify some notation. When many random variables are involved in an expectation, we use a subscript to indicate with respect to which variables the expectation is taken. If a subscript is not present, the expectation is taken with respect to all present random variables. In addition, we sometimes use $\chi(\cdot)$ in place of the indicator function, i.e., $\chi(z) = \mathbf{1}\{z \leq 0\}$.

3.3.1 Globally Invertible g_2

The main idea of GLR involves three ingredients [9]: approximating $\chi(\cdot)$ by a smooth sequence, applying integration by parts, and taking limits. These steps can be justified under some regularity conditions, as presented in Assumption 3.3.

To summarize, (i) guarantees that the existence of a certain smooth sequence for approximation, (ii) justifies the application of integration by parts and the resulted surface integral vanishes from (iii), and the convergence of the approximated sequences can be shown from (iv).

With these assumptions, a generalized result of Theorem 3.1 can be obtained, as shown in Theorem 3.2, whose proof follows Peng et al. [9] by additionally conditioning on the σ-algebra generated by random variables Y and V.

Assumption 3.3. *Let $\Theta(\alpha) = [\alpha - \tau, \alpha + \tau]$ for some $\tau > 0$ be the interval around α.*

(i) There exists a sequence of smooth functions $\{\chi_\epsilon\}$ such that for some $p \in [1, \infty]$,

$$\lim_{\epsilon \to 0} \sup_{\theta \in \Theta(\alpha)} \left(\int_{\mathbb{R}} |\chi(g_2(u, V) - \theta)) - \chi_\epsilon(g_2(u, V) - \theta))|^p \, du \right)^{1/p} = 0 \quad a.s.,$$

and for q satisfying $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathbb{E}_{X,U} \left[|g_1(X, Y) w(X, U, V)|^q \right] < \infty, \quad \mathbb{E}_{X,U} \left[|w(X, U, V)|^q \right] < \infty \quad a.s.$$

(ii) The function $g_2(u, V(\omega))$ is twice continuously differentiable with respect to u a.s. and $\partial_u g_2(u, V(\omega))^{-1}$ is differentiable w.r.t $u \in \mathbb{R}$ a.s.

(iii) $\lim_{u \to \pm\infty} g_1(x, Y(\omega))) \partial_u g_2(u, V(\omega))^{-1} f_{X,U}(x, u) = 0$ for all $x \in \mathbb{R}$ and $\theta \in \Theta(\alpha)$,

$\lim_{u\to\pm\infty} \partial_u g_2(u, V(\omega))^{-1} f_{X,U}(x,u) = 0$ *for all* $x \in \mathbb{R}$ *and* $\theta \in \Theta(\alpha)$.

(iv) $\mathbb{E}_{X,U}\left[\sup_{\theta\in\Theta(\alpha)} |g_1(X,Y)\chi(g_2(u,V)-\theta)w(X,U,V)|\right] < \infty$ *a.s.,*

$\mathbb{E}_{X,U}\left[\sup_{\theta\in\Theta(\alpha)} |\chi(g_2(u,V)-\theta)w(X,U,V)|\right] < \infty$ *a.s.*

Theorem 3.2. *Under Assumption 3.1 (i)-(iii) and Assumption 3.3, the conditional expectation (3.1) can be expressed as*

$$\mathbb{E}\left[g_1(X,Y)|g_2(U,V)=\alpha\right] = \frac{\mathbb{E}\left[g_1(X,Y)\chi(g_2(U,V)-\alpha)w(X,U,V)\right]}{\mathbb{E}\left[\chi(g_2(U,V)-\alpha)w(X,U,V)\right]},$$

where

$$w(x,u,v) = -\frac{\partial_u^2 g_2(u,v)}{\left(\partial_u g_2(u,v)\right)^2} + \frac{\partial_u \ln f_{X,U}(x,u)}{\partial_u g_2(u,v)}. \tag{3.3}$$

Proof. Under Assumption 3.1 (ii), by the definition of the conditional expectation, (3.1) can be formulated as (3.2). Now we first consider the numerator of (3.2).

Let $\sigma(Y,V)$ be the σ-algebra generated by random variables Y and V and define

$$\tilde{g}_2(u,v;\theta) = g_2(u,v) - \theta.$$

By Assumption 3.1 (ii), we have

$$\begin{aligned}
&\frac{d}{d\theta}\mathbb{E}\left[\mathbb{E}\left[g_1(X,Y)\chi(g_2(U,V)-\theta)\right)|\sigma(Y,V)]\right] \\
=&\frac{d}{d\theta}\mathbb{E}\left[\int_{\mathbb{R}^2} g_1(x,Y)\chi(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)dxdu\right] \\
=&\mathbb{E}\left[\frac{d}{d\theta}\int_{\mathbb{R}^2} g_1(x,Y)\chi(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)dxdu\right],
\end{aligned} \tag{3.4}$$

The justification of the second equality will be given later in this proof. Consider the argument inside the expectation operator in (3.4):

$$
\begin{aligned}
&\frac{d}{d\theta}\int_{\mathbb{R}^2} g_1(x,Y)\chi_\epsilon(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)dxdu \\
=&\int_{\mathbb{R}^2} g_1(x,Y)\frac{d}{d\theta}\chi_\epsilon(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)dxdu \qquad (3.5)\\
=&-\int_{\mathbb{R}^2} g_1(x,Y)\partial_u\tilde{g}_2(u,V;\theta)^{-1}\frac{d}{du}\chi_\epsilon(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)dxdu \\
=&\int_{\mathbb{R}}\Big(-g_1(x,Y)\partial_u\tilde{g}_2(u,V;\theta)^{-1}\chi_\epsilon(\tilde{g}_2(u,V;\theta))f_{X,U}(x,u)|_{-\infty}^{\infty} \\
&+\int_{\mathbb{R}} g_1(x,Y)\chi_\epsilon(\tilde{g}_2(u,V;\theta))\partial_u\left(\partial_u\tilde{g}_2(u,V;\theta)^{-1}f_{X,U}(x,u)\right)du\Big)dx \qquad (3.6)\\
=&\int_{\mathbb{R}^2} g_1(x,Y)\chi_\epsilon(\tilde{g}_2(u,V;\theta))w(x,u,V)f_{X,U}(x,u)dudx \\
=&\mathbb{E}\left[g_1(X,Y)\chi_\epsilon(\tilde{g}_2(U,V;\theta))w(X,U,V)\right].
\end{aligned}
$$

The interchange of integration and differentiation will also be justified later in this proof. The second equality is obtained by $\frac{d}{du}\chi_\epsilon(\tilde{g}_2(u,V;\theta)) = \frac{d}{dz}\chi_\epsilon(z)|_{z=\tilde{g}_2(u,V;\theta)}\cdot\partial_u\tilde{g}_2(u,V;\theta)$ and $\frac{d}{d\theta}\chi_\epsilon(\tilde{g}_2(u,V;\theta)) = \frac{d}{dz}\chi_\epsilon(z)|_{z=\tilde{g}_2(u,V;\theta)}\cdot\partial_\theta\tilde{g}_2(u,V;\theta) = -\frac{d}{dz}\chi_\epsilon(z)|_{z=\tilde{g}_2(u,V;\theta)}$. The third equality is obtained by integration by parts under Assumption 3.3 (ii). The last equality holds since the first term in (3.6) vanishes by Assumption 3.3 (iii). Then we prove the following uniform convergence result:

$$
\begin{aligned}
&\lim_{\epsilon\to 0}\sup_{\theta\in\Theta(\alpha)}\Big|\mathbb{E}_{X,U}\left[g_1(X,Y)\chi(\tilde{g}_2(U,V;\theta))w(X,U,V)\right] \\
&-\mathbb{E}_{X,U}\left[g_1(X,Y)\chi_\epsilon(\tilde{g}_2(U,V;\theta))w(X,U,V)\right]\Big| \\
\leq&\lim_{\epsilon\to 0}\sup_{\theta\in\Theta(\alpha)}\mathbb{E}_{X,U}\left[|g_1(X,Y)w(X,U,V)|\,|\chi(\tilde{g}_2(U,V;\theta))-\chi_\epsilon(\tilde{g}_2(U,V;\theta))|\right]
\end{aligned}
$$

$$\leq \lim_{\epsilon \to 0} \sup_{\theta \in \Theta(\alpha)} \left(\mathbb{E}_{X,U} \left[|g_1(X,Y) w(X,U,V)|^q \right] \right)^{1/q} \left(\int_{\mathbb{R}} |\chi(\tilde{g}_2(u,V;\theta)) - \chi_\epsilon(\tilde{g}_2(u,V;\theta))|^p \, du \right)^{1/p}$$

$$\leq C_1 \lim_{\epsilon \to 0} \sup_{\theta \in \Theta(\alpha)} \left(\int_{\mathbb{R}} |\chi(\tilde{g}_2(u,V;\theta)) - \chi_\epsilon(\tilde{g}_2(u,V;\theta))|^p \, du \right)^{1/p} = 0,$$

for some constant $C_1 \geq 0$. The second inequality is obtained by Holder's inequality. The third inequality and the last equality are obtained by Assumption 3.3 (i).

Then by Assumption 3.3 (iv), we have

$$\mathbb{E}_{X,U} \left[\sup_{\theta \in \Theta(\alpha)} |g_1(X,Y) \chi_\epsilon(\tilde{g}_2(U,V;\theta)) w(X,U,V)| \right] < \infty,$$

which can be used together with the mean value theorem to justify the interchange of integration and differentiation to obtain (3.5).

Similarly, we have

$$\lim_{\epsilon \to 0} \sup_{\theta \in \Theta(\alpha)} |\mathbb{E}_{X,U} \left[g_1(X,Y) \chi(\tilde{g}_2(U,V;\theta)) \right] - \mathbb{E}_{X,U} \left[g_1(X,Y) \chi_\epsilon(\tilde{g}_2(U,V;\theta)) \right]| = 0.$$

Therefore, by the uniform convergence of $\frac{d}{d\theta} \mathbb{E}_{X,U} \left[g_1(X,Y) \chi_\epsilon(\tilde{g}_2(U,V;\theta)) \right]$ and the convergence of the expectation, we have

$$\frac{d}{d\theta} \mathbb{E}_{X,U} \left[g_1(X,Y) \chi(\tilde{g}_2(U,V;\theta)) \right] = \mathbb{E}_{X,U} \left[g_1(X,Y) \chi(\tilde{g}_2(U,V;\theta)) w(X,U,V) \right].$$

By Assumption 3.3 (iv), the interchange of integration and differentiation to obtain (3.4)

can be justified, then

$$\frac{d}{d\theta}\mathbb{E}\left[g_1(X,Y)\chi(\tilde{g}_2(U,V;\theta))\right] = \mathbb{E}\left[g_1(X,Y)\chi(\tilde{g}_2(U,V;\theta))w(X,U,V)\right].$$

Repeating the above steps for the denominator in (3.2), we have

$$\frac{d}{d\theta}\mathbb{E}\left[\chi(\tilde{g}_2(U,V;\theta))\right] = \mathbb{E}\left[\chi(\tilde{g}_2(U,V;\theta))w(X,U,V)\right].$$

The proof is completed by letting $\theta = \alpha$. □

The verification of Assumption 3.3 is usually not very straightforward, especially condition (i). One possible choice of the approximated sequence of smooth functions $\{\chi_\epsilon\}$, as suggested in Peng et al. [9], can be given by $\chi_\epsilon(z) = \chi_\epsilon * \phi_\epsilon(z)$, where $*$ stands for the convolution operator, ϕ_ϵ is the density function of a normal distribution $\mathcal{N}(0,\epsilon^2)$, and χ_ϵ is defined by

$$\chi_\epsilon(z) = \begin{cases} 1 & z < -\epsilon \\ -\frac{1}{2\epsilon}z + \frac{1}{2} & -\epsilon \le z \le \epsilon \\ 0 & z > \epsilon \end{cases}.$$

Then condition (i) can be checked by taking the function g_2 into account. In the following, we provide some simplified conditions that can imply Assumption 3.3 (i), (iv), but are easier to check for some cases.

Corollary 3.1. *Under Assumption 3.1 (i)-(iii) and Assumption 3.3 (ii)-(iii), suppose that*

$$\mathbb{E}\left[|g_1(X,Y)|^2\right] < \infty, \quad \mathbb{E}\left[|w(X,U,V)|^2\right] < \infty,$$

the representation of the conditional expectation (3.1) given in Theorem 3.2 holds.

Proof. By Holder's inequality, we have

$$\mathbb{E}\left[|g_1(X,Y)w(X,U,V)|\right] \leq \left(\mathbb{E}\left[|g_1(X,Y)|^2\right]\right)^{1/2}\left(\mathbb{E}\left[|w(X,U,V)|^2\right]\right)^{1/2} < \infty.$$

By Jensen's inequality, we have $\mathbb{E}\left[|w(X,U,V)|\right] \leq \left(\mathbb{E}\left[|w(X,U,V)|^2\right]\right)^{1/2} < \infty$. Taking $p = \infty$ and $q = 1$, we can see Assumption 3.3 (i) holds. Assumption 3.3 (iv) can also be verified by noticing that $\chi(\cdot)$ is bounded. □

Comparing Corollary 3.1 and Theorem 3.1, we can clearly see that our approach is a generaliza- tion of CDM for estimating conditional expectations, since Assumption 3.2 is no longer required. However, to apply our results, some integrability conditions have to be verified.

3.3.2 g_2 with Stationary Points

One limitation of Theorem 3.2 is that Assumption 3.3 (ii) requires $(\partial_u g_2(u,V))^{-1}$ to be differentiable with respect to u a.s. Generally speaking, this implies that for a fixed sample path, g_2 does not have a stationary point and is a strictly monotonic function of u, which is however weaker than Assumption 3.2. For instance, $g_2(u,v) := e^{uv} + v$, $v > 0$, satisfies Assumption 3.3 (ii), but not Assumption 3.2.

Next, we consider a simplified setting where random variables Y and V are not included

and g_2 has stationary points, i.e., $g_2'(u) = 0$ has real roots. The major problem in this setting is that the function $w(x,u) = \partial_u \left(g_2'(u)^{-1}\right) + g_2'(u)^{-1}\partial_u \ln f_{X,U}(x,u)$ defined similar to (3.3) may not be integrable w.r.t $f_{X,U}(x,u)$. The idea to handle this problem is to consider isolated intervals around the stationary points.

Suppose that $g_2'(u) = 0$ has a finite number of stationary points p_i, $i = 1, ..., n$, and $p_1 < p_2 < \cdots < p_n$, with $p_0 = -\infty$ and $p_{n+1} = \infty$. Then $g_2(u) = \alpha$ can have at most $n+1$ roots, since it is continuous. We denote the roots by u_i, $i = 0, ..., k-1$, where $u_0 < u_1 < \cdots < u_{k-1}$ and k is the number of roots. We further assume for any u_i, $g_2(u)$ is locally invertible in a neighborhood around it. Then there exist points a_i and b_i, such that $p_j < a_i \leq u_i \leq b_i < p_{j+1}$. A simple example is depicted in Figure 3.1.

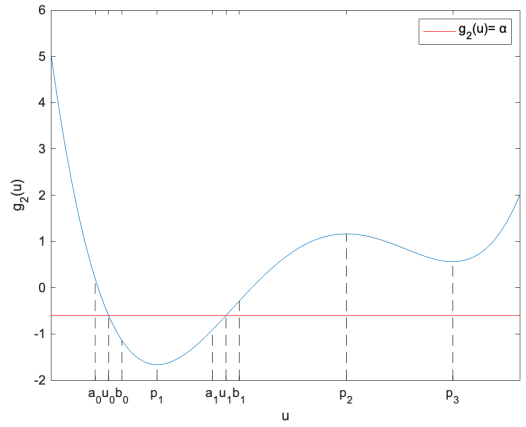

Figure 3.1: Plot of $g_2(u)$ with stationary points.

Since the conditional expectation can be formulated as a ratio of two SGE problems, as seen in (3.2), and the integration around the stationary points p_i does not affect the SGE results around $\theta = \alpha$, it is sufficient to consider the SGE problems for $U \in \bigcup_{i=0}^{k-1}[a_i, b_i]$.

Next, we consider the SGE problems around a particular u_i where g_2 is locally invertible. Without loss of generality, we assume that $g_2(a_i) > \alpha$ and $g_2(b_i) \leq \alpha$. Here, we omit the

technical details, since they are similar to Theorem 3.2. By the three ingredients of GLR, under appropriate conditions, for the numerator of the SGE formulation, we have

$$
\begin{aligned}
&\frac{d}{d\theta}\Big(\int g_1(x)\int_{a_i}^{b_i}\mathbf{1}\left\{g_2(u)\le\theta\right\}f_{X,U}(x,u)dudx\Big)|_{\theta=\alpha}\\
=&-\frac{1}{g_2'(b_i)}\int g_1(x)f_{X,U}(x,b_i)dx+\int\int_{a_i}^{b_i}g_1(x)\mathbf{1}\left\{g_2(u)\le\alpha\right\}w(x,u)f_{X,U}(x,u)dudx\\
=&-\frac{1}{g_2'(b_i)}\mathbb{E}[g_1(X)\delta(U-b_i)]]+\mathbb{E}[g_1(X)\mathbf{1}\left\{g_2(U)\le\alpha\right\}w(X,U)\mathbf{1}\left\{a_i\le U\le b_i\right\}]\\
=&-\frac{1}{g_2'(b_i)}\mathbb{E}[g_1(X)\mathbf{1}\left\{U\le b_i\right\}\partial_u\ln f_{X,U}(X,U)]]\\
&+\mathbb{E}[g_1(X)\mathbf{1}\left\{g_2(U)\le\alpha\right\}w(X,U)\mathbf{1}\left\{a_i\le U\le b_i\right\}],
\end{aligned}
$$

where the first equality is obtained by noticing that $\mathbf{1}\left\{g_2(a_i)\le\alpha\right\}=0$ and $\mathbf{1}\left\{g_2(b_i)\le\alpha\right\}=1$, and the last equality is obtained by GLR again. Thus, by considering piecewise SGE problems, we can represent the conditional expectation by ordinary expectations when g_2 has stationary points.

In most cases, the application of the above results requires a rough knowledge of locations of p_i, $i=0,1,...,n+1$, and u_i, $i=0,...,k-1$. For the special case where $g_2(u)$ has a unique stationary point, a simplified representation can be obtained, an example of which is given in the numerical experiment section (Experiment 3, setting (i)).

3.3.3 Confidence Interval of the Ratio Estimator

Theorem 3.2 indicates that we can construct a ratio of two sample averages as an estimator of the conditional expectation from N independent sample paths. Suppose the samples of the numerator and denominator are denoted by A_i and B_i, $i=1,...,n$, respectively. The sample

means $\hat{\mu}_A$, $\hat{\mu}_B$, the sample variances S_A^2, S_B^2, and the sample covariance S_{AB}^2 can be computed correspondingly. Then, the mean and the variance of the ratio estimator

$$R = \frac{\frac{1}{N}\sum_{i=1}^{N} A_i}{\frac{1}{N}\sum_{i=1}^{N} B_i},$$

can be estimated by the delta method [70] as

$$\begin{aligned}
\mathbb{E}[R] &\approx \hat{r} = \frac{\hat{\mu}_A}{\hat{\mu}_B}, \\
\mathbb{V}[R] &\approx \hat{\sigma}_R^2 = \frac{1}{N\hat{\mu}_B^2}\left(S_A^2 - 2S_{AB}^2\hat{r} + \hat{r}^2 S_B^2\right).
\end{aligned}$$

For a large N, a $(1-a)\%$ confidence interval for the ratio estimator is given by

$$\left[\hat{r} - z_{\frac{a}{2}}\hat{\sigma}_R,\ \hat{r} + z_{\frac{a}{2}}\hat{\sigma}_R\right],$$

where $z_{\frac{a}{2}}$ is the $(1-a/2)\%$ quantile of the standard normal distribution.

3.3.4 Variance Reduction

When formulating the associated stochastic gradient estimation problems, we can use either $\mathbf{1}\{g_2(U,V) \geq \theta\}$ or $\mathbf{1}\{g_2(U,V) \leq \theta\}$. Theoretically, they are both correct, but the sign inside the indicator function can have a significant impact on the variance of the estimator. Suppose that $P(g_2(U,V) \leq \alpha)$ is very small; then, the event $\mathbf{1}\{(g_2(U,V) \leq \alpha)\}$ is a rare event. Thus, if the number of independent replications is relatively limited, it is highly possible that the variances are large. Moreover, for the finite difference method, the denominator may be zero, which results

in an invalid ratio estimator. Therefore, a preliminary estimate of the distribution of $g_2(U, V)$ around the given α can be used to guide the appropriate formulation.

3.4 Simulation Experiments

In this section, we test our approach on three estimation problems. In the first experiment, we consider a financial option problem and demonstrate the usefulness of our simple variance reduction technique. In the second experiment, we verify our approach by applying it to a toy problem of estimating a conditional expectation where g_2 is globally invertible. In the last experiment, we consider the case where g_2 is locally invertible. We show that our approach can yield better estimations and lower variances, and can outperform the finite difference method.

3.4.1 Experiment 1: Financial Option

We consider the European option pricing problem for which Daveloose et al. [66] applied the conditional density method to estimate the option price at maturity T, given the stock price at time t, $0 < t < T < \infty$:

$$\mathbb{E}\left[\psi(S_T)|S_t = \alpha\right], \tag{3.7}$$

where S_t is the stock price at time t, $t \geq 0$. S_t is assumed to follow the Merton model [63], i.e.,

$$S_t = S_0 \exp\left(\left(r - \frac{\sigma^2}{2}\right)t + \sigma W_t + \tilde{N}_t\right),$$

where $\{W_t\}$ is the standard Brownian motion, $r > 0$ is the risk-free interest rate and σ is the volatility, $\tilde{N}_t = \sum_{i=1}^{N(t)} Z_i$, $\{N(t)\}$ is a Poisson process with arrival rate λ, and Z_i are i.i.d. random variables following $\mathcal{N}\left(-\frac{\rho^2}{2}, \frac{\rho^2}{2}\right)$ with a constant ρ. Thus, $\{\tilde{N}_t\}$ is a compound Poisson process. For a put option, $\psi(z) = e^{-r(T-t)}(K - z)^+$, where K is the strike price, and the true value of (3.7) can be computed using an analytic formula which is in the form of an infinite sum [63].

Let $X = \beta W_T$, $U = \beta W_t$, $Y = \mu T + \tilde{N}_T$, $V = \mu t + \tilde{N}_t$, $g_1(x, y) = \psi(S_0 e^{x+y})$, $g_2(u, v) = u + v$. The joint density $f_{X,U}(x, u)$ has an analytic formula. Since the exponential function is monotonically increasing, (3.7) can be rewritten as

$$\mathbb{E}\left[g_1(X, Y) | g_2(U, V) = \ln(\alpha/S_0)\right].$$

For this problem, Assumption 3.2 holds, therefore, CDM is applicable. Since the GLR method is a generalization of CDM, both methods yield the same result:

$$\mathbb{E}\left[\psi(S_T)|S_t = \alpha\right] = \frac{\mathbb{E}\left[\psi(S_T)\mathbf{1}\left\{S_t \le \alpha\right\}\frac{TW_t - tW_T}{\sigma t(T-t)}\right]}{\mathbb{E}\left[\mathbf{1}\left\{S_t \le \alpha\right\}\frac{TW_t - tW_T}{\sigma t(T-t)}\right]}. \tag{3.8}$$

Daveloose et al. [66] considered variance reduction techniques such as localization and control variates. Here, we reduce the estimation variances by appropriately choosing the inequality sign inside the indicator function in (3.8) based on the value of α, as discussed in Section 3.3.4.

We use the same parameter values as in Daveloose et al. [66]: $S_0 = 40$, $K = 45$, $r = 0.08$, $\sigma^2 = \rho^2 = 0.05$, $\lambda = 5$, $T = 1$. For CDM, we apply the localization techniques for variance reduction. For GLR, we first estimate $P(S_t \le \alpha)$ using existing sample paths. If the probability

Table 3.1: Put option price estimates in Experiment 1 ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation).

t	CDM	GLR	FD ($\varepsilon = 0.05$)	TRUE
0.1	28.48(2.01)	32.14(0.98)	27.92(1.49)	31.89
0.2	31.49(0.96)	32.13(0.39)	31.47(1.30)	32.22
0.3	32.54(0.59)	32.84(0.19)	33.06(0.53)	32.56
0.4	32.45(0.37)	32.86(0.13)	32.66(0.43)	32.90
0.5	33.07(0.28)	33.44(0.08)	33.04(0.26)	33.24
0.6	33.08(0.21)	33.65(0.07)	33.49(0.19)	33.58
0.7	33.55(0.19)	33.96(0.05)	33.91(0.14)	33.93
0.8	34.06(0.17)	34.31(0.05)	34.20(0.09)	34.29
0.9	34.53(0.18)	34.66(0.05)	34.68(0.05)	34.64

exceeds 0.5, we use $\mathbf{1}\{S_t \leq \alpha\}$; otherwise, we use $\mathbf{1}\{S_t \geq \alpha\}$. We run the experiments for $\alpha = 10$ using $N = 5 \times 10^6$ independent runs. The numerical results are shown in Table 3.1, where the true values are approximated by truncating the infinite series after 25 terms. From the results, we can see that an appropriate choice of the inequality sign inside the indicator function can have a great impact on the estimation quality. For example, since $P(S_{0.1} \leq \alpha)$ is very small, if we use $\mathbf{1}\{S_{0.1} \leq \alpha\}$ in the estimator even with the localization technique, the estimator can have a large relative error, whereas using $\mathbf{1}\{S_{0.1} \geq \alpha\}$ as in the GLR estimators leads to much better results.

3.4.2 Experiment 2: g_2 without Explicit Inverse

Consider estimating

$$\mathbb{E}\left[g_1(X)|g_2(U,V) = \alpha\right], \tag{3.9}$$

where U is a random variable following a Student's t-distribution with degrees of freedom $\nu = 4$, the conditional distribution of $X|U$ is $\mathcal{N}(U, 1)$, V follows a Bernoulli distribution, which takes value 1 with probability $3/4$ and value 2 with probability $1/4$, and $g_1(x) = e^x$, $g_2(u, v) = e^{uv} + u$. For a given $\alpha = 0$, the conditional expectation (3.9) can be computed exactly by applying the properties of the Dirac function.

Since g_2 does not satisfy Assumption 3.2, we cannot directly apply the conditional density method. Before applying our approach, we need to check the conditions. The joint density function of (X, U) is given by

$$f_{X,U}(x, u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-u)^2}{2}} \cdot \frac{3}{8\left(1 + \frac{u^2}{4}\right)^{5/2}}, \quad -\infty < x, u < \infty.$$

It is easy to see that Assumption 3.1 and Assumption 3.3 (iii) hold. Since $\chi(\cdot)$ is bounded and the decay rate of joint density $f_{X,U}(x, u)$ when x and u go to $\pm\infty$ suppresses $g_1(x)$ and $\partial_u g_2(u, V)^{-1}$ for any x a.s., Assumption 3.3 (ii) holds. The other conditions in Corollary 3.1 can also be verified. Therefore, (3.9) can be represented by

$$\frac{\mathbb{E}[g_1(X)\mathbf{1}\{g_2(U, V) \leq \alpha\} w(X, U, V)]}{\mathbb{E}[\mathbf{1}\{g_2(U, V) \leq \alpha\} w(X, U, V)]},$$

where

$$w(x, u, v) = -\frac{v^2 e^{uv}}{(ve^{uv} + 1)^2} + \frac{-u + x - \frac{5u}{u^2+4}}{ve^{uv} + 1}.$$

Given $\alpha = 2$, the results are shown in Table 3.2. When the number of independent runs is small, e.g., $N = 1000$, the samples of the denominator may be all zero if using the FD method with a

Table 3.2: Estimations in Experiment 2 ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation).

	$N = 10^3$		$N = 10^6$		
method	GLR	FD($\varepsilon = 0.1$)	GLR	FD($\varepsilon = 0.01$)	TRUE
estimation	2.45(0.42)	2.91(0.72)	2.48(0.01)	2.55(0.03)	2.50

small perturbation ($\varepsilon = 0.01$). Thus, for a small number of sample paths, FD requires a larger perturbation ($\varepsilon = 0.1$). The results show that GLR yields more accurate estimations, with lower variances than FD.

3.4.3 Experiment 3: g_2 with Stationary Points

Here, we consider the problem of estimating $\mathbb{E}\left[g_1(X)|g_2(U) = 0\right]$ where $g_2(u)$ has stationary points for two settings: (i) $g_1(x) = x^2e^{-x}$ and $g_2(u) = 1 - u^2$, (ii) $g_1(x) = x$ and $g_2(u) = (u-1)(u-2)(u-3)$. In both settings, we assume that $U \sim \mathcal{N}(2,1)$ and $X|U \sim \mathcal{N}(U,1)$, and the true values of the conditional expectation can be computed analytically. The joint density function is given by

$$f_{X,U}(x,u) = \frac{1}{2\pi} e^{-\frac{(x-u)^2}{2} - \frac{(u-2)^2}{2}}, \quad -\infty < x, u < \infty.$$

Setting (i). The associated function $w(x,u)$ is given by $w(x,u) = \frac{1}{2u^2} + \frac{2u-x-2}{2u}$.

Since $\mathbf{1}\left\{g_2(u) \leq 0\right\} = 0$ around a small neighborhood of the unique stationary point $p_1 = 0$ of $g_2(u)$, $\mathbb{E}\left[\mathbf{1}\left\{g_2(U) \leq 0\right\} w(X,U)\right] < \infty$ and $\mathbb{E}\left[g_1(X)\mathbf{1}\left\{g_2(U) \leq 0\right\} w(X,U)\right] < \infty$. Therefore, we have

$$\mathbb{E}\left[g_1(X)|g_2(U) = 0\right] = \frac{\mathbb{E}\left[g_1(X)\mathbf{1}\left\{g_2(U) \leq 0\right\} w(X,U)\right]}{\mathbb{E}\left[\mathbf{1}\left\{g_2(U) \leq 0\right\} w(X,U)\right]}.$$

Setting (ii). The associated function $w(x,u)$ is given by $w(x,u) = -\frac{6u-12}{(3u^2-12u+11)^2} - \frac{2u-x}{3u^2-12u+11}$. $g_2(u)$ has two stationary points $p_1 = 1.423$ and $p_2 = 2.577$ and the roots of $g_2(u) = 0$ are $u_0 = 1$, $u_1 = 2$ and $u_3 = 3$. Applying the method introduced in Section 3.3.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left[g_1(X)|g_2(U)=0\right] \\
=&\frac{\mathbb{E}\left[g_1(X)\left(\left(\mathbf{1}\left\{g_2(U)\le 0\right\}+\mathbf{1}\left\{U\le b\right\}-1\right)w(X,U)-\frac{\mathbf{1}\{U\le b\}\partial_u \ln f_{X,U}(X,U)}{g_2'(b)}\right)\right]}{\mathbb{E}\left[\left(\mathbf{1}\left\{g_2(U)\le 0\right\}+\mathbf{1}\left\{U\le b\right\}-1\right)w(X,U)-\frac{\mathbf{1}\{U\le b\}\partial_u \ln f_{X,U}(X,U)}{g_2'(b)}\right]},
\end{aligned}$$

where b can be any value in $(2, 2.577)$. The numerical results for the two settings are given in Table 3.3, showing that the GLR estimator is more accurate and has a smaller variance compared to the FD estimator.

Table 3.3: Estimation results in Experiment 3 with 10^6 independent runs ($\hat{\sigma}_R$ in parentheses, ε is FD perturbation, $b = 2.3$ for Setting (ii)).

	GLR	FD ($\varepsilon = 0.01$)	TRUE
Setting (i)	1.02(0.02)	0.82(0.12)	1.00
Setting (ii)	2.00(0.004)	1.98(0.01)	2.00

3.5 Conclusion

Expressing the conditional expectation as a ratio of two derivatives, we apply the generalized likelihood ratio method to both SGE problems (numerator and denominator). Our method generalizes the conditional density method and thus can handle more complicated measurement functions. For the case where the measurement function has stationary points, we consider the simplified setting without the random variable V. However, our methodology can still work if the random variable V does not affect the locations of the stationary points of $g_2(u,V)$ and the roots of

$g_2(u, V) = \alpha$. Future work is needed for the more general setting when this is not the case. We also provide a simple way to reduce the variance of the estimator, which may be incorporated with jackknifing to reduce the ratio bias and importance sampling to achieve further variance reduction.

Chapter 4: Policy Evaluation With Stochastic Gradient Estimation Techniques

4.1 Introduction

Reinforcement learning (RL) techniques have been successful in decision-making problems that can be modeled using finite-horizon stochastic processes. For a general introduction to RL techniques, one can refer to Sutton and Barto [71]. In this chapter, we focus on the problem of policy evaluation, where the task is to estimate a value function for a given action policy. The action policy can then be iteratively improved by finding the optimal action based on the estimated value function. This procedure is known as the policy improvement step.

Two widely-used approaches for policy evaluation are the Monte Carlo (MC) method and the temporal difference (TD) method [17]. The TD label encompasses a broader class of methods, such as TD(λ), introduced by Tsitsiklis and Van Roy [72]. The value of a given state can be represented by a conditional expectation, and both MC and TD estimate it by simulating the trajectory of the process starting from that same state. The main difference between them is that MC requires the simulation to continue until termination, while TD simulates only a single transition and calculates an estimate based on the Bellman equation. In this chapter, we investigate ways to estimate the value of a state using sample paths that start in other states.

In Chapter 3, we have shown that under appropriate assumptions, a conditional expectation can be represented as a ratio of the gradients of two expectations. The finite difference (FD)

method is a simple but effective method, but a major drawback is that it generally produces a biased estimator, which could impair the convergence rate of policy evaluation. Another important disadvantage of FD is that when we use a smaller perturbation parameter to decrease the bias of the estimator, a larger number of sample paths are needed to reduce the variance. It is possible to obtain an unbiased estimator using the generalized likelihood ratio (GLR) method [9], but this requires an explicit form for the distribution of the initial state. Though we have the flexibility to specify this density function, it is not clear how to do this "well". Moreover, the application of GLR generally requires differentiability of the value function, thus has many limitations.

With a FD-based or GLR-based estimator of the value function, we can use stochastic gradient descent to minimize the Bellman error, analogously to TD. The asymptotic convergence of such a procedure can be obtained by extending the available theory for TD methods, which is mainly based on the ODE method for stochastic approximation [73]. This approach, however, does not provide finite-time performance guarantees. Bhandari et al. (2018) investigated the finite-time performance of TD in infinite-horizon, discrete-state problems with linear function approximation. However, for the algorithms associated with our proposed new FD and GLR estimators, the finite-time analysis could become more complicated, since there are many parameters that should be determined by the user. For example, the number of sample paths used to obtain the FD estimator could have a great impact on the performance of the algorithm, and a large number of sample paths would lead to high simulation costs, which is important when we consider budget-dependent convergence rate results [74] for these algorithms.

The chapter is organized as follows. In Section 4.2, we review the problem of policy evaluation in a finite horizon setting along with two classical methods, MC and TD. In Section

4.3, we propose new estimators for the value function using stochastic gradient estimation techniques. In Section 4.4, we describe the detailed algorithms and give preliminary insights into the convergence analysis. Numerical experiments are conducted in Section 4.5.

4.2 Problem Formulation

We consider the policy evaluation problem for a Markov decision process (MDP) in a finite-horizon setting where the state variables take continuous values. Let T be the termination time and $\mathcal{X}$ be the state space. Suppose at time step $t = 0, \ldots, T$, the state variable is $x_t \in \mathcal{X}$, and the next state x_{t+1} is generated by the transition

$$x_{t+1} = h_t(x_t, \pi(x_t, t), Z_{t+1}), \tag{4.1}$$

where π is the given action policy that we aim to evaluate, i.e., $\pi(x_t, t)$ gives the action taken at t, and the quantity Z_{t+1} represents the randomness of the transition, assumed to be independent of the state and the action. Let $R_{t+1}(x_t, x_{t+1})$ be a deterministic function representing the one-step reward obtained between steps t and $t+1$. The value function associated with the MDP, denoted by V^π is then given by

$$V^\pi(x, t) = \mathbb{E}\left[\sum_{i=t}^{T-1} \gamma^{i-t} R_{i+1}(x_i, x_{i+1}) | x_t = x, t\right],$$

where $\gamma \in (0, 1)$ is a discount factor. It is well known that V^π satisfies the Bellman equations $V^\pi(x, T) = 0$, and $V^\pi(x, t) = \mathcal{L}V^\pi(x, t+1)$ for $t = 0, \ldots, T-1$, where the operator $\mathcal{L}$ is

defined by

$$\mathcal{L}V(x,t+1) = \mathbb{E}\left[R_{t+1}(X_t, X_{t+1}) + \gamma V(X_{t+1}, t+1) | X_t = x, t\right], \quad t = 0, \ldots, T-1.$$

Given a fixed initial state x_0 and the fixed policy π, the distribution $F_t(x)$ of the state variable X_t is fixed. Since the state variables are assumed to take continuous values, we suppose that the value function V^π is approximated by a parametrized model $V_\theta(x,t)$ for parameter $\theta \in \Theta$, where Θ is a closed and convex parameter space. In general, the parametrized model could take different forms for different t. For each $t = 0, \ldots, T$, it is desirable to find θ_t^* such that

$$\theta_t^* = \arg\min_{\theta_t \in \Theta} \frac{1}{2} \left\| V_{\theta_t}(x,t) - V^\pi(x,t) \right\|_{F_t}^2 \tag{4.2}$$

where $\|\cdot\|_{F_t}^2$ is defined by

$$\|V(x,t)\|_{F_t}^2 := \mathbb{E}_{X \sim F_t}\left[|V(X,t)|^2\right].$$

Let Π_t represent the projection operator that gives the best approximation of a given function with respect to F_t. Then the solution to (4.2) can be written as $\Pi_t V^\pi(x,t)$. For simplicity, we assume that θ_t^* exists and is unique for each t.

4.2.1 Monte Carlo (MC) and Temporal Difference (TD) Methods

In the following, we give a brief review of the Monte Carlo (MC) and temporal difference (TD) methods for fitting the value of a fixed policy π.

Suppose that at iteration $n+1$, we simulate a sample path $\{x_0, x_1, R_1, x_2, R_2, \ldots, x_T, R_T\}$. Then, the MC and the TD estimators for $V^\pi(x,t)$ are given by $\hat{V}^{MC}(x,t) = \sum_{i=t}^{T-1} \gamma^{i-t} R_{i+1}$ and $\hat{V}^{TD}(x,t) = R_{t+1} + \gamma V_{\theta^n_{t+1}}(x_{t+1}, t+1)$, respectively, where θ^n is the parameter obtained in iteration n. In practice, for both methods, estimators are calculated in a backward direction. The MC estimator at iteration $n+1$ does not depend on θ^n, whereas the TD estimator is affected by θ^n. Therefore, in TD, instead of solving (4.2) as in the MC method, we are essentially solving a different problem for θ_t,

$$\min_{\theta_t \in \Theta} \frac{1}{2} \left\| V_{\theta_t}(x,t) - \mathcal{L} V_{\theta_{t+1}}(x,t+1) \right\|^2_{F_t}. \tag{4.3}$$

Consequently, the error in the approximation of $V^\pi(x,t)$ by $V_{\theta^*_t}(x,t)$ depends on the approximation error induced by θ^*_{t+1}. If $V_{\theta^*_{t+1}} = \Pi_{t+1} V^\pi$, then the approximation error of $V^\pi(x,t)$ by $V_{\theta^*_t}(x,t)$ is simply a projection error, which can be seen from

$$\begin{aligned}
&\left\| V_{\theta^*_t}(x,t) - V^\pi(x,t) \right\|_{F_t} \\
&= \left\| \Pi_t \mathcal{L} V_{\theta^*_{t+1}}(x,t+1) - V^\pi(x,t+1) \right\|_{F_t} \\
&\leq \left\| \Pi_t \mathcal{L} V_{\theta^*_{t+1}}(x,t+1) - \Pi_t \mathcal{L} V^\pi(x,t+1) \right\|_{F_t} + \left\| \Pi_t V^\pi(x,t) - V^\pi(x,t) \right\|_{F_t} \\
&\leq \gamma \left\| V_{\theta^*_{t+1}}(x,t+1) - V^\pi(x,t+1) \right\|_{F_{t+1}} + \left\| \Pi_t V^\pi(x,t) - V^\pi(x,t) \right\|_{F_t},
\end{aligned}$$

where the first inequality is obtained from the Bellman equation, and the second equality follows because the projection operator Π_t is non-expansive and the Bellman operator $\mathcal{L}$ is a contraction mapping.

The optimal parameters θ are searched using a stochastic gradient descent algorithm for the

objectives (4.2) and (4.3), where the gradient estimator can be written in a general form:

$$g_t(\theta_t, \theta_{t+1}) = (\hat{V}^\pi(x,t) - V_{\theta_t}(x,t))\nabla_{\theta_t} V_{\theta_t}(x,t). \tag{4.4}$$

Therefore, the key difference between MC and TD lies in the form of $\hat{V}^\pi(x,t)$. In the next section, we propose new formulations of $\hat{V}^\pi(x,t)$ derived by stochastic gradient estimation (SGE) techniques.

4.3 New Estimators of the Value Function

In the following, we present a ratio representation of a conditional expectation, and then propose two new estimators of the value function.

4.3.1 Ratio representation of a conditional expectation

The Bellman equation shows that the problem of estimating $V^\pi(x,t)$ can be reduced to the problem of estimating a particular conditional expectation. Furthermore, Zhou et al. [6] have shown that a conditional expectation can be represented by a ratio of the gradients of two expectations. For our purpose, this technique can be applied in policy evaluation. To simplify the presentation, we consider the case where the state variable x is a one-dimensional variable.

Define $W(x, x', t; V) = R_t(x, x') + \gamma V(x', t)$. Then,

$$\begin{aligned}\mathcal{L}V(x, t+1) =& \mathbb{E}\left[W(X_t, X_{t+1}, t+1; V) | X_t = x, t\right] \\ =& \lim_{\varepsilon \to 0} \mathbb{E}\left[W(X_t, X_{t+1}, t+1; V) | X_t \in (x-\varepsilon, x+\varepsilon)\right]\end{aligned}$$

$$= \frac{\lim_{\varepsilon\to 0}(1/2\varepsilon)\mathbb{E}\left[W(X_t, X_{t+1}, t+1; V)\mathbf{1}\left\{X_t \in (x-\varepsilon, x+\varepsilon)\right\}\right]}{\lim_{\varepsilon\to 0}(1/2\varepsilon)\mathbb{E}\left[\mathbf{1}\left\{X_t \in (x-\varepsilon, x+\varepsilon)\right\}\right]} \tag{4.5}$$

$$= \left.\frac{\frac{d}{d\xi}\mathbb{E}\left[W(X_t, X_{t+1}, t+1; V)\mathbf{1}\left\{X_t \le \xi\right\}\right]}{\frac{d}{d\xi}\mathbb{E}\left[\mathbf{1}\left\{X_t \le \xi\right\}\right]}\right|_{\xi=x}. \tag{4.6}$$

If we can obtain estimators of the two stochastic gradients in (4.6), then $\mathcal{L}V(x, t+1)$ can be estimated by taking the ratio. For higher-dimension problems, the representation would involve higher-order partial derivatives. By the derived ratio representation, when estimating the value function at a specific point (x, t), we do not have to simulate the next states starting at exactly this fixed point as in the MC or the TD method. Instead, we could use sample paths from $(\tilde{x}, t)$, where $\tilde{x}$ can be different from x. In other words, given t, any sample paths starting from any state can be used to estimate the value function for any x. Therefore, the ratio representation allows much more flexibility in simulating sample paths and offers new ways of constructing different estimators of V.

4.3.2 Stochastic gradient estimation (SGE)

By the ratio representation (4.6), stochastic gradient estimation (SGE) techniques are needed to estimate the conditional expectation. There are many widely used SGE techniques, which include finite difference (FD) methods, infinitesimal perturbation analysis (IPA) and the likelihood ratio (LR) method. Among those, FD is the one of the most straightforward methods. Suppose $\mathcal{D} = \{(X_t^j, X_{t+1}^j, R_{t+1}^j), j = 1, \ldots, M\}$ are i.i.d random pairs, where $X_t^j \sim F_t$, the corresponding X_{t+1}^j is obtained by the transition function (4.1), and $R_{t+1}^j = R_{t+1}(X_t^j, X_{t+1}^j)$. Then, from (4.5),

a symmetric FD estimator is given by

$$\hat{V}^{FD}(x,t;V,\mathcal{D},\varepsilon)=\frac{\sum_{j=1}^{M}W(X_t^j,X_{t+1}^j,t+1;V)\mathbf{1}\left\{X_t^j\in(x-\varepsilon,x+\varepsilon)\right\}}{\sum_{j=1}^{M}\mathbf{1}\left\{X_t^j\in(x-\varepsilon,x+\varepsilon)\right\}}, \tag{4.7}$$

where $\varepsilon>0$ is a small fixed perturbation parameter. Although FD is easy to implement, it has several drawbacks. First, the FD estimator is biased, which usually leads to a slower convergence rate compared to unbiased estimators. Moreover, for a small ε, the event $\{X_t\in(x-\varepsilon,x+\varepsilon)\}$ could be a rare event, thus a large number of sample paths may be required to make the denominator in (4.7) nonzero. To overcome these drawbacks, we apply the generalized likelihood ratio (GLR) [9] method, which can obtain unbiased estimators for (4.6). The application of GLR can be justified under Assumption 4.1.

Assumption 4.1. *The following statements hold for all t:*

(i) Z_{t+1} *is independent of* X_t.

(ii) $f_t(x)$*, the density of* X_t*, has support* $\mathbb{R}$.

(iii) $\ln f_t(x)$ *is differentiable for all* $x\in\mathbb{R}$.

(iv) $\mathbb{E}\left[(\partial_x \ln f_t(X_t))^2\right]<\infty$.

(v) $\mathbb{E}\left[(W(X_t,X_{t+1},t+1;V))^2\right]<\infty$.

Under Assumption 4.1, similar to what we have done in Chapter 3, the numerator in (4.6) can be represented as an ordinary expectation according to

$$\frac{d}{d\xi}\mathbb{E}\left[W(X_t,X_{t+1},t+1;V)\mathbf{1}\left\{X_t\le\xi\right\}\right]$$

$$=\mathbb{E}\left[\mathbf{1}\left\{X_t \le \xi\right\}\left(\frac{d}{dX_t}W(X_t, X_{t+1}, t+1; V) + W(X_t, X_{t+1}, t+1; V)\frac{d}{dX_t}\ln f_t(X_t)\right)\right]. \tag{4.8}$$

Using the GLR method, we can obtain an unbiased estimator for the numerator of (4.6). However, the representation (4.8) requires explicit knowledge of the density function $f_t(\cdot)$, which is not available in practice. At the same time, the derivation of the ratio representation (4.6) does not require that X_t is generated from F_t, i.e., the expectations in (4.6) can be taken with respect to a user-specified distribution F_t^s. Following this, to apply GLR, we need Assumption 4.1 to hold where F_t and f_t are replaced by F_t^s and f_t^s, respectively. For the denominator of (4.6), we can apply the same GLR technique to obtain a corresponding ordinary expectation, but the ratio of two sample means would still be a biased estimator for $\mathcal{L}V(x, t+1)$. Therefore, another benefit of using a user-specified distribution F_t^s is that the denominator of (4.8) simply becomes $f_t^s(x)$, which leads to an unbiased estimator for $\mathcal{L}V(x, t+1)$ when we use sample means of i.i.d. observations to estimate the numerator. Thus, the final form of the GLR estimator is given by

$$\begin{aligned}\hat{V}^{GLR}(x, t; V, \tilde{\mathcal{D}}) =& \frac{1}{M \cdot f_t^s(x)} \sum_{j=1}^{M}\left[\mathbf{1}\left\{\tilde{X}_t^j \le x\right\}\left(\frac{d}{d\tilde{X}_t^j}W(\tilde{X}_t^j, \tilde{X}_{t+1}^j, t+1; V)\right.\right.\\ &\left.\left. + W \cdot \frac{d}{d\tilde{X}_t^j}\ln f_t^s(\tilde{X}_t^j)\right)\right],\end{aligned} \tag{4.9}$$

where $\tilde{\mathcal{D}} = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1, \ldots, M\}$ are i.i.d random pairs, where $\tilde{X}_t^j \sim F_t^s$, the corresponding $\tilde{X}_{t+1}^j$ is obtained by (4.1), and $\tilde{R}_{t+1}^j = R_{t+1}(\tilde{X}_t^j, \tilde{X}_{t+1}^j)$. Although the GLR estimator is unbiased, it is more complicated to implement and has some limitations compared to other estimators. First, it is not clear which choices of $F_t(s)$ are "good". Moreover, since we are aiming to estimate the value function at $X_t \sim F_t$, extra simulation is needed to simulate

these evaluation points. Moreover, if W is not differentiable w.r.t the state variable, (4.9) would introduce additional conditional expectation terms conditioning on points at which W is not differentiable. In such cases, the GLR method could be ineffective.

4.4 Algorithms

In general, we update the parameters θ using

$$\theta_t^{n+1} = \theta_t^n + \alpha^n g_t^n \quad t = T, \dots, 0, \tag{4.10}$$

where α_n is the step-size and $g_t^n = g_t(\theta_t^n, \theta_{t+1}^n)$ is the gradient estimator at iteration n. For different methods, we only need to replace $\hat{V}^\pi(x,t)$ in (4.4) by the corresponding estimators. For example, for FD, the gradient estimator is

$$g_t^{FD}(\theta_t^n, \theta_{t+1}^n; \varepsilon) = (\hat{V}^{FD}(x,t; V_{\theta_{t+1}^n}, \varepsilon) - V_{\theta_t^n}(x,t))\nabla_{\theta_t^n} V_{\theta_t^n}(x,t). \tag{4.11}$$

The resulting policy evaluation algorithms with the FD and GLR estimators are given in Algorithms 4.1 and 4.2.

The convergence analysis of algorithms using FD and GLR estimators would be similar to that of TD and MC, which has been studied extensively in the literature. One common approach is to use ODE techniques to establish asymptotic convergence by showing that the ODE has a global asymptotically stable equilibrium [72]. Moreover, under appropriate assumptions, convergence rate results can also be established [75]. Here, we provide some insights into the proofs of the proposed algorithms by using previous results shown for stochastic approximation (SA) problems

Algorithm 4.1: Policy Evaluation with FD.

Input: fixed policy π, value function approximation model V_θ, initial number of sample paths M_0, simulation budget M in each iteration, number of evaluation points for each t in each iteration, N_0, step sizes sequences $\{\alpha_t^n\}$, initial parameters θ^0, initial permutation parameter $\varepsilon^0 > 0$.

Simulate M_0 sample paths $\mathcal{D}^0 \leftarrow \{(X_t^j, X_{t+1}^j, R_{t+1}^j), j = 1, \ldots, M_0, t = 0, ..., T-1\}$ under policy π.

for $n = 0, 1, \ldots$ **do**

  Simulate M sample paths $\mathcal{S}^n = \{(X_t^j, X_{t+1}^j, R_{t+1}^j), j = 1, \ldots, M, t = 0, ..., T-1\}$ under policy π, and $\mathcal{D}^{n+1} = \mathcal{D}^n \bigcup \mathcal{S}^n$.

  $\varepsilon^n \leftarrow \varepsilon^0 n^{-1/6}$.

  for $t = T, T-1, \ldots, 0$ **do**

    $\mathcal{C}^{n+1} \leftarrow N_0$ numbers randomly selected from $\{1, \ldots, |\mathcal{D}^{n+1}|\}$ with equal probability.

    Evaluation: For each $X_t^i, i \in \mathcal{C}^{n+1}$, calculate the FD estimator $\hat{V}^{FD}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \mathcal{D}^{n+1}, \varepsilon^n)$.

    Batch Gradient Estimator: For each $X_t^i, i \in \mathcal{C}^{n+1}$, calculate the gradient estimator by (4.11), denoted by $g_t^{n,i}$, the batch estimator $g_t^n \leftarrow (1/|\mathcal{C}^{n+1}|) \sum_{i \in \mathcal{C}^{n+1}} g_t^{n,i}$.

    Update: $\theta_t^{n+1} \leftarrow \theta_t^n + \alpha_t^n g_t^n$.

  end

end

[74].

First, we define

$$\bar{\psi}_t(\theta_t) = \mathbb{E}_{X_t \sim F_t}\left[\left(\mathcal{L}V_{\theta_{t+1}^*}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t)\right)\nabla V_{\theta_t}(X_t, t)\right],$$

$$\bar{g}_t(\theta_t, \theta_{t+1}) = \mathbb{E}_{X_t \sim F_t}\left[\left(\mathcal{L}V_{\theta_{t+1}}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t)\right)\nabla V_{\theta_t}(X_t, t)\right].$$

Let (X_t, X_{t+1}) be a random pair, where $X_t \sim F_t$ and X_{t+1} is its corresponding next state, then

$$\psi_t(\theta_t) = (\hat{V}(X_t, t; V_{\theta_{t+1}^*}) - V_{\theta_t}(X_t, t))\nabla_{\theta_t} V_{\theta_t}(X_t, t),$$

$$g_t(\theta_t, \theta_{t+1}) = (\hat{V}(X_t, t; V_{\theta_{t+1}}) - V_{\theta_t}(X_t, t))\nabla_{\theta_t} V_{\theta_t}(X_t, t)$$

Algorithm 4.2: Policy Evaluation with GLR.

Input: fixed policy π, value function approximation model V_θ, initial number of sample paths M_0 and $\tilde{M}_0$, simulation budget M and $\tilde{M}$ in each iteration, step sizes sequences $\{\alpha_t^n\}$, initial parameters θ^0, user-specified simulation distributions $F_t^s(\cdot)$, $t = 0, \ldots, T$.

construct the initial set of evaluation points: simulate M_0 sample paths $\mathcal{D}^0 \leftarrow \{X_t^j, j = 1, \ldots, M_0, t = 0, ..., T\}$ under policy π.

For each $t = 0, \ldots, T$, simulate $\tilde{M}_0$ sample paths $\tilde{\mathcal{D}}_t^0 = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1, \ldots, \tilde{M}_0\}$ under policy π, where $\tilde{X}_t^j \sim F_t^s$.

for $n = 0, 1, \ldots$ **do**

  Simulate M_0 sample paths $\mathcal{S}^n \leftarrow \{X_t^j, j = 1, \ldots, M_0, t = 0, ..., T\}$ under policy π following the dynamics. $\mathcal{D}^{n+1} \leftarrow \mathcal{D}^n \bigcup \mathcal{S}^n$.

  For each $t = 0, \ldots, T$, simulate $\tilde{M}$ sample paths $\tilde{\mathcal{S}}_t^n = \{(\tilde{X}_t^j, \tilde{X}_{t+1}^j, \tilde{R}_{t+1}^j), j = 1, \ldots, \tilde{M}\}$ under policy π, where $\tilde{X}_t^j \sim F_t^s$, and $\tilde{\mathcal{D}}_t^{n+1} \leftarrow \tilde{\mathcal{D}}_t^n \bigcup \tilde{\mathcal{S}}_t^n$.

  for $t = T, T-1, \ldots, 0$ **do**

    $\mathcal{C}^{n+1} \leftarrow N_0$ numbers randomly selected from $\{1, \ldots, |\mathcal{D}^{n+1}|\}$ with equal probability.

    Evaluation: For each $X_t^i, i \in \mathcal{C}^{n+1}$, calculate the GLR estimator $\hat{V}^{GLR}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \tilde{\mathcal{D}}_t^{n+1})$.

    Batch Gradient Estimator: For each $X_t^i, i \in \mathcal{C}^{n+1}$, calculate the gradient estimator, denoted by $g_t^{n,i}$, the batch estimator $g_t^n \leftarrow (1/|\mathcal{C}^{n+1}|) \sum_{i \in \mathcal{C}^{n+1}} g_t^{n,i}$.

    Update: $\theta_t^{n+1} \leftarrow \theta_t^n + \alpha_t^n g_t^n$.

  end

end

are stochastic estimators of $\bar{\psi}_t(\theta_t)$ and $\bar{g}_t(\theta_t, \theta_{t+1})$, respectively. It is easily seen that $\bar{\psi}_t(\theta_t)$ is the negative of the gradient of the objective in

$$\min_{\theta_t} \frac{1}{2} \left\| \left(\mathcal{L} V_{\theta_{t+1}^*}(X_{t+1}, t+1) - V_{\theta_t}(X_t, t) \right)^2 \right\|_{F_t}, \tag{4.12}$$

which is very similar to (4.3), except that in (4.12) the optimal θ_{t+1}^* is given. Using the update

$$\theta_t^{n+1} = \theta_t^{n+1} + \alpha_t^n \psi_t(\theta_t^n)$$

with $\alpha_t^n = O(n^{-1})$, an $O\left(n^{-\min\{2\beta, 1+\nu\}}\right)$ convergence rate of $\mathbb{E}\left[\|\theta_t^n - \theta_t^*\|^2\right]$ can be established

under assumptions on the decay rates of (4.13) and (4.14), along with other assumptions on $\bar{\psi}$ (see Theorem 3.1 in L'Ecuyer and Yin [74]):

$$\left\|\bar{\psi}(\theta_t^n) - \mathbb{E}[\psi_t^n|\theta_t^n]\right\| \leq K_\beta n^{-\beta} \; w.p.1 \tag{4.13}$$

$$\mathbb{E}\left[\left\|\bar{\psi}(\theta_t^n) - \mathbb{E}[\psi_t^n|\theta_t^n]\right\|^2\right] \leq K_\nu n^{-\nu}. \tag{4.14}$$

Similar results could be expected to hold for our update (4.10) under similar assumptions, i.e., we would require the decaying behavior of the conditional bias and variance of the estimator g_t^n to follow (4.13) and (4.14). In practice, for FD, the conditional bias can be reduced by using a decreasing perturbation parameter along with an increasing number of sample paths, while for GLR, the bias is 0, since the estimator is unbiased.

4.5 Numerical Experiments

In this section, we compare four algorithms with different gradient estimators: the MC estimator, the temporal difference estimator, the FD estimator and the generalized likelihood ratio estimator. Algorithms are tested on two benchmark problems. In both experiments, we use a 2nd-degree polynomial model to approximate the value function.

4.5.1 Optimal Asset Allocation and Consumption

We consider policy evaluation in the problem of finding optimal asset allocation and consumption to maximize aggregated utility of consumption [76]. Suppose we have one risky asset and one riskless asset. The value of the risky asset S_t follows a geometric Brownian process

$dS_t = \mu S_t dt + \sigma S_t dZ_t$, where μ is a drift constant and σ is a volatility constant, Z_t is a standard Brownian motion. In addition, the riskless asset $\tilde{S}_t$ follows $d\tilde{S}_t = r\tilde{S}_t dt$, where r is the riskfree rate. Let X_t be the wealth at time t, and denote by $p(X_t, t)$ the percentage of wealth allocated to the risky asset and $1 - p(X_t, t)$ the percentage of wealth allocated to the riskless asset. Let $c(X_t, t)$ be the wealth consumption per unit time. Then the wealth X_t evolves according to

$$dX_t = ((p(X_t, t)(\mu - r) + r)X_t - c(X_t, t))dt + p(X_t, t)\sigma X_t dZ_t.$$

The optimal value function is given by

$$V^*(x, t) = \max_{p,c} \mathbb{E}\left[\int_t^T e^{-\rho(s-t)} U(c(X_s, s))ds + e^{-\rho(s-t)} B(T)U(X_T)|X_t = x, t\right]$$

where $B(T) = \epsilon^\eta$ is a fixed function, $\epsilon, \eta \in (0, 1)$, $U(x)$ is the utility of consumption function, $\rho \geq 0$ is the utility discount rate. Suppose $U(x) = \frac{x^{1-\eta}}{1-\eta}$, and the optimal action policy and the associated optimal value function have closed forms. Thus, we can apply policy evaluation algorithms for the optimal action policy (p^*, c^*) and compare the estimated value function with the optimal value function. First, we approximate the original continuous-time problem by discretization. The one-step transition thus can be written as

$$X_{t+1} = X_t + ((p_t^*(X_t, t)(\mu - r) + r)X_t - c^*(X_t, t))\Delta t + p_t^*(X_t, t)\sigma X_t \cdot \tilde{Z}_{t+1}, \tag{4.15}$$

where Δt is a small time interval, $\tilde{Z}_{t+1} \sim \mathcal{N}(0, \sqrt{\Delta t})$. The Bellman operator can be approximated by

$$\mathcal{L}V(x, t+1) = \mathbb{E}\left[e^{-\rho \cdot \Delta t} U(c^*(X_t, t))\Delta t + e^{-\rho \cdot \Delta t} V(X_{t+1}, t+1) | X_t = x, t\right]. \tag{4.16}$$

The MC, TD and FD estimators can be easily derived from previous discussions. For GLR, we have to specify a distribution F_t^s. Intuitively, we would hope F_t^s be close to F_t (see Figure 4.1), but this usually cannot be achieved in practice. Instead, we set F_t^s to be $exp(\lambda_t)$, where $\lambda_t = 1/(at+b)$, and a and b are parameters fitted by a linear regression of the sample means $\{\bar{X}_t,\ t = 0, \ldots, T\}$ that is calculated from some pre-simulated sample paths starting from an initial wealth x_0. We use an exponential distribution for F_t^s, since the support of the X_t is non-negative under the optimal action policy. However, from Assumption 4.1, the support of f_t^s be all of $\mathbb{R}$ for GLR to be valid. Therefore, we need to perform a change of variable, e.g., $Y_t = \ln(X_t)$ in (4.16). Above all, in nth iteration, the GLR estimator (see Algorithm 4.2) is:

$$\begin{aligned}
&\hat{V}^{GLR}(X_t^i, X_{t+1}^i; V_{\theta_{t+1}^n}, \tilde{\mathcal{D}}_t^{n+1}) \\
=&\frac{e^{-\rho \cdot \Delta t}}{|\tilde{\mathcal{D}}_t^{n+1}| \cdot f_t^s(X_t^i) \cdot X_t^i} \sum_{j=1}^{|\tilde{\mathcal{D}}_t^{n+1}|} \Bigg[\mathbf{1}\left\{\tilde{X}_t^j \le X_t^i\right\} \Bigg(\frac{dV_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1)}{d\tilde{Y}_t^j} \\
&+ V_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1) \cdot q(\tilde{Y}_t^j) \Bigg) \Bigg] + e^{-\rho \cdot \Delta t} U(c^*(X_t^i, t))\Delta t,
\end{aligned}$$

where $f_t^s(x) = \lambda e^{-\lambda x}$, $\tilde{X}_t^j \sim exp(\lambda_t)$, $\tilde{Y}_t^j = \ln(\tilde{X}_t^j)$ with density $f_Y(y) = \lambda e^{y - \lambda e^y}$, and $q(y) := \frac{d \ln(f_y(y))}{dy} = 1 - \lambda e^y$. The gradient $\frac{dV_{\theta_{t+1}^n}(\tilde{X}_{t+1}^j, t+1)}{d\tilde{Y}_t^j}$ can be calculated by taking derivatives of the parameterized approximation model and the transition function (4.15) and applying the chain

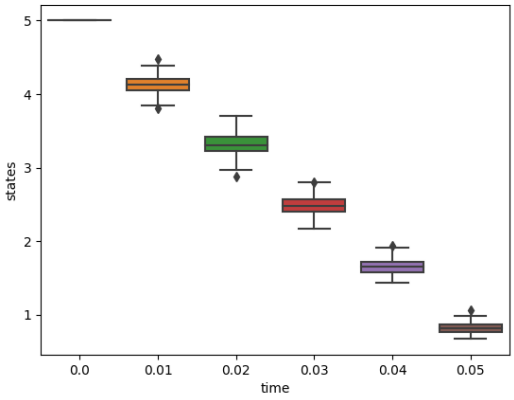

Figure 4.1: Boxplot of $F_t(\cdot)$, distribution of the state variable X_t. Parameters used: $T = 0.05$, $\Delta t = 0.01$, $x_0 = 10$, $\mu = 1$, $\sigma = 2$, $r = 0.8$, $\rho = 1.2$, $\eta = 0.4$, $\epsilon = 0.01$.

rule.

Numerical results are shown in Figure 4.2, from which we can see the algorithm with FD has the best performance as it obtains a closest approximation of the true value function, and is the most stable. The GLR estimator has the worst performance, but becomes comparable to other methods after many epochs. It is worth mentioning that for the GLR method, there are many hyperparameters that can be tuned, such as F_t^s. The choice of F_t^s could have a great impact on the performances, and how to design a "good" one still remains a challenging problem.

4.5.2 (s, S) Inventory Policy Evaluation

We consider a classic periodic review inventory control problem with zero lead time, i.e., an order arrives as soon as it is placed. Let X_t be the inventory level at the beginning of day t, $t = 0, \ldots, T$, a_t be the order amount between t and $t + 1$, c_h be the unit holding cost, c_s be the unit shortage cost, $c_s > c_h$, c be the unit order cost and $K > 0$ be the fixed reorder cost. The cost

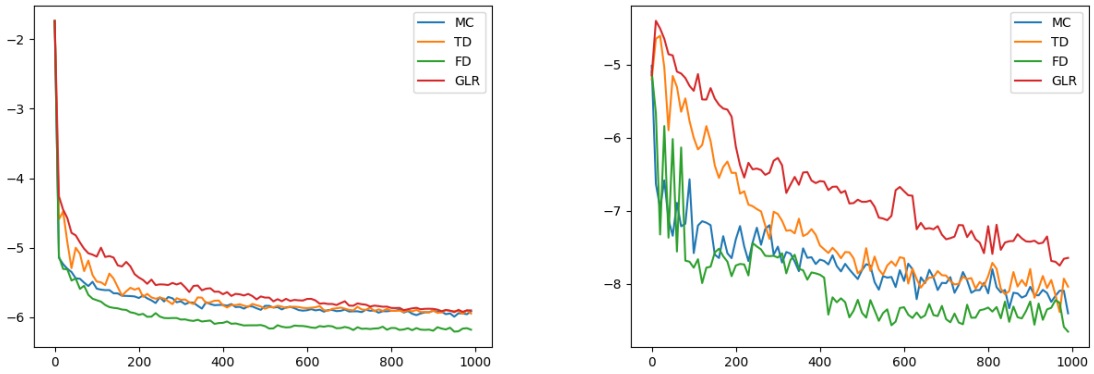

Figure 4.2: The left and right plots represent the logarithm of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Results are reported every 10 epochs. For FD, $M_0 = 200$, $M = 1$, $N_0 = 200$. For GLR, $M_0 = \tilde{M}_0 = 200$, $M = \tilde{M} = 1$, $N_0 = 200$.

R_{t+1} incurred between t and $t+1$ is given by

$$c_h \max\{X_t, 0\} + c_s \max\{-X_t, 0\} + ca_t + K\mathbf{1}\{a_t > 0\}.$$

Assume that the demand Z_{t+1} between t and $t+1$, $t = 0, \ldots, T-1$ are i.i.d. $\exp(\lambda)$ random variables. Then the inventory X_{t+1} is obtained by the following transition

$$X_{t+1} = X_t + a_t - Z_{t+1}.$$

The policy we would like to evaluate is the well-known (s, S) policy, $S > s$, that is

$$a_t(X_t) = \begin{cases} S - X_t & X_t \leq s \\ 0 & X_t > s \end{cases}.$$

For this problem, the action policy $a_t(x)$ is not differentiable at $x = s$, therefore, the GLR

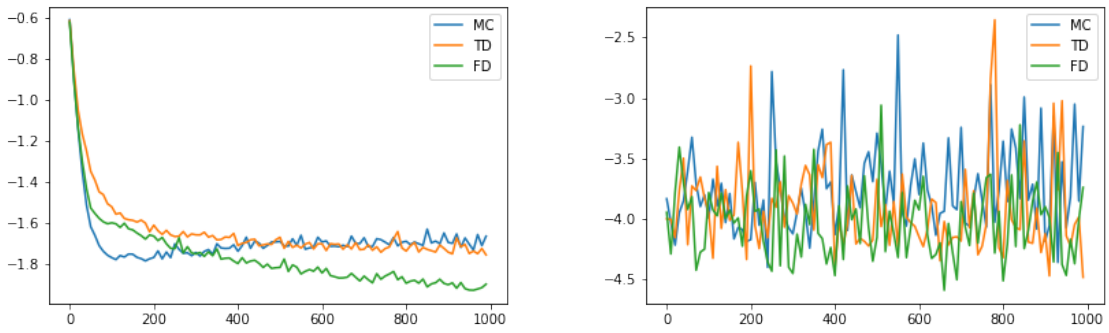

Figure 4.3: (s, S) policy evaluation. The left and right plots represent the logarithm of sample averages and the standard errors of $\sum_{t=0}^{T} \frac{\|V-V^*\|_{F_t}}{\|V^*\|_{F_t}}$ vs epochs over 50 macro-replications, respectively. Parameters used: $T = 5$, $\gamma = 0.9$, $\lambda = 0.2$, $c_h = 3$, $c_s = 5$, $c = 3$, $K = 5$. Results are reported every 10 epochs. For FD, $M_0 = 200$, $M = 1$, $N_0 = 200$.

method would yield an additional conditional expectation conditioning on $X_t = s$. Considering this issue, we only compare MC, TD and FD for this experiment. Numerical results are shown in Figure 4.3, from which we can see that FD achieves a higher accuracy compared to MC and TD after around 350 epochs. In addition, the stability of the three methods are comparable. In practice, in one iteration, FD requires a higher computational cost than MC and TD, since FD uses a set of sample paths instead of a single one. However, the benefit is that this set of sample paths can be used to construct estimators of the value function for many points in one iteration. This feature makes the policy evaluation algorithm with FD suitable for parallel computing, which could greatly reduce the computational time.

4.6 Conclusion

We have proposed two new estimators for policy evaluation, the FD estimator and the GLR estimator, based on the key observation that the conditional expectation in the Bellman error can be represented as a ratio of two ordinary expectations. These two estimators allow

us the flexibility to estimate the value of any given state using sample paths starting from other states. However, there are still many practical issues that need further investigation. For example, the number of sample paths used in each iteration could greatly affect the performance of the algorithms. Moreover, for GLR, although theoretically it can yield an unbiased estimator, how to choose an appropriate F_t^s remains another practical challenge. One possible way is to generate sample paths following F_t as the iteration proceeds and dynamically adjusting F_t^s based on collected replications. GLR is theoretically appealing, but has many challenges for practical implementation and is more suitable when the action policy and the value function are differentiable.

Bibliography

[1] Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen E Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

[2] Peter I Frazier, Warren B Powell, and Savas Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[3] Peter I Frazier, Warren B Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[4] Yi Zhou, Michael C Fu, and Ilya O Ryzhov. Sequential learning with a similarity selection index. *submitted to Operations Research*, 2023.

[5] Guowei Sun, Yunchuan Li, and Michael C Fu. A spectral index for selecting the best alternative. In N. Mustafee, K.-H. G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, editors, *Proceedings of the 2019 Winter Simulation Conference*, pages 3404–3415, 2019.

[6] Yi Zhou, Michael C Fu, and Ilya O Ryzhov. Estimating a conditional expectation with the generalized likelihood ratio method. In S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, editors, *Proceedings of the 2021 Winter Simulation Conference*, pages 1–12, Piscataway, New Jersey, 2021. Institute of Electrical and Electronics Engineers, Inc.

[7] Paul Glasserman and Yu-Chi Ho. *Gradient Estimation via Perturbation Analysis*, volume 116. Springer Science & Business Media, 1991.

[8] Yu Chi Ho and Xiren Cao. Perturbation analysis and optimization of queueing networks. *Journal of Optimization Theory and Applications*, 40(4):559–582, 1983.

[9] Yijie Peng, Michael C Fu, Jian-Qiang Hu, and Bernd Heidergott. A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Research*, 66(2):487–499, 2018.

[10] Ralph Neuneier. Optimal asset allocation using adaptive dynamic programming. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems, Vol 8*, page 952–958, Cambridge, Massachusetts, 1995. MIT Press.

[11] Afshin Oroojlooyjadid, MohammadReza Nazari, Lawrence V Snyder, and Martin Takáč. A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management*, 24(1):285–304, 2022.

[12] Francis A Longstaff and Eduardo S Schwartz. Valuing American options by simulation: a simple least-squares approach. *The Review of Financial Studies*, 14(1):113–147, 2001.

[13] John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

[14] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems, Vol 12*, page 1057–1063, Cambridge, Massachusetts, 1999. MIT Press.

[15] LA Prashanth and Michael C Fu. Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends in Machine Learning*, 15(5):537–693, 2022.

[16] Yi Zhou, Michael C Fu, and Ilya O Ryzhov. Policy evaluation with stochastic gradient estimation techniques. In B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, editors, *Proceedings of the 2022 Winter Simulation Conference*, pages 3039–3050, Piscataway, New Jersey, 2022. Institute of Electrical and Electronics Engineers, Inc.

[17] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[18] Chun-Hung Chen, Stephen E Chick, Loo Hay Lee, and Nugroho A. Pujowidianto. Ranking and selection: efficient simulation budget allocation. In M. C. Fu, editor, *Handbook of Simulation Optimization*, pages 45–80. Springer, 2015.

[19] Seong-Hee Kim and Barry L Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation*, 11(3): 251–273, 2001.

[20] Stephen E Chick, Jürgen Branke, and Christian Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

[21] Donghai He, Stephen E Chick, and Chun-Hung Chen. Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Transactions on Systems, Man, and Cybernetics*, C37(5):951–961, 2007.

[22] Weiwei Fan, L Jeff Hong, and Barry L Nelson. Indifference-zone-free selection of the best. *Operations Research*, 64(6):1499–1514, 2016.

[23] Sijia Ma and Shane G Henderson. An efficient fully sequential selection procedure guaranteeing probably approximately correct selection. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, editors, *Proceedings of the 2017 Winter Simulation Conference*, pages 2225–2236, 2017.

[24] Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In R. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 577–585, 2004.

[25] Susan R Hunter and Raghu Pasupathy. Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS Journal on Computing*, 25(3):527–542, 2013.

[26] Tianhan Gao and Wei Lu. Machine learning toward advanced energy storage devices and systems. *iScience*, 24(1):101936:1–101936:33, 2021.

[27] Fu S Xue, Yan M Zhang, X Liao, Jian H Liu, and Gang An. Influences of age and gender on dose response and time course of effect of atracurium in anesthetized adult patients. *Journal of Clinical Anesthesia*, 11(5):397–405, 1999.

[28] Bin Han, Ilya O Ryzhov, and Boris Defourny. Optimal learning in linear regression with combinatorial feature selection. *INFORMS Journal on Computing*, 28(4):721–735, 2016.

[29] Haihui Shen, L Jeff Hong, and Xiaowei Zhang. Ranking and selection with covariates. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, editors, *Proceedings of the 2017 Winter Simulation Conference*, pages 2137–2148, 2017.

[30] Mark W Brantley, Loo Hay Lee, Chun-Hung Chen, and Argon Chen. Efficient simulation budget allocation with regression. *IIE Transactions*, 45(3):291–308, 2013.

[31] Warren R Scott, Warren B Powell, and Hugo P Simão. Calibrating simulation models using the knowledge gradient with continuous parameters. In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference*, pages 1099–1109, 2010.

[32] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

[33] Huashuai Qu, Ilya O Ryzhov, Michael C Fu, and Zi Ding. Sequential selection with unknown correlation structures. *Operations Research*, 63(4):931–948, 2015.

[34] Qiong Zhang and Yongjia Song. Moment-matching-based conjugacy approximation for Bayesian ranking and selection. *ACM Transactions on Modeling and Computer Simulation*, 27(4):26:1–26:23, 2017.

[35] Chun-Hung Chen and Loo Hay Lee. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. World Scientific, 2010.

[36] Michael C Fu, Jian-Qiang Hu, Chun-Hung Chen, and Xiaoping Xiong. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing*, 19(1):101–111, 2007.

[37] Yijie Peng, Edwin KP Chong, Chun-Hung Chen, and Michael C Fu. Ranking and selection as stochastic control. *IEEE Transactions on Automatic Control*, 63(8):2359–2373, 2018.

[38] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[39] Michael Grant and Stephen P Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[40] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications (2nd ed.)*. Springer Berlin Heidelberg, 2009.

[41] Jiaqi Zhou and Ilya O Ryzhov. A new rate-optimal sampling allocation for linear belief models. *Operations Research (to appear)*, 2022.

[42] Raghu Pasupathy, Susan R Hunter, Nugroho A Pujowidianto, Loo Hay Lee, and Chun-Hung Chen. Stochastically constrained ranking and selection via score. *ACM Transactions on Modeling and Computer Simulation*, 25(1):1–26, 2014.

[43] Siyang Gao, Weiwei Chen, and Leyuan Shi. A new budget allocation framework for the expected opportunity cost. *Operations Research*, 65(3):787–803, 2017.

[44] Chih-Jen Lin, Stefano Lucidi, Laura Palagi, Arnaldo Risi, and Marco Sciandrone. Decomposition algorithm model for singly linearly-constrained problems subject to lower and upper bounds. *Journal of Optimization Theory and Applications*, 141(1):107–126, 2009.

[45] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 2013.

[46] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[47] Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5381–5391, 2017.

[48] Ye Chen and Ilya O Ryzhov. Complete expected improvement converges to an optimal budget allocation. *Advances in Applied Probability*, 51(1):209–235, 2019.

[49] Morris H DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 2005.

[50] Ilya O Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 64(6):1515–1528, 2016.

[51] Yijie Peng and Michael C Fu. Myopic allocation policy with asymptotically optimal sampling rate. *IEEE Transactions on Automatic Control*, 62(4):2041–2047, 2017.

[52] Jiu Ding and Aihui Zhou. Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, 20(12):1223–1226, 2007.

[53] Peng Zhang. *Industrial Control Technology: A Handbook for Engineers And Researchers*. William Andrew, 2008.

[54] Dheeru Dua and Casey Graff. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.

[55] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[56] Yijie Peng, Chun-Hung Chen, Michael C Fu, and Jian-Qiang Hu. Non-monotonicity of probability of correct selection. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 3678–3689, 2015.

[57] Amir Ali Nasrollahzadeh and Amin Khademi. Optimal stopping of adaptive dose-finding trials. *Service Science*, 12(2-3):80–99, 2020.

[58] Christian Ritz, Florent Baty, Jens C Streibig, and Daniel Gerhard. Dose-response analysis using R. *PloS ONE*, 10(12):e0146021, 2015.

[59] Linda Pei, Barry L Nelson, and Susan R Hunter. Parallel adaptive survivor selection. *Operations Research*, 2022.

[60] Yijie Peng, Jie Xu, Loo Hay Lee, Jianqiang Hu, and Chun-Hung Chen. Efficient simulation sampling allocation using multifidelity models. *IEEE Transactions on Automatic Control*, 64(8):3156–3169, 2018.

[61] Travis Goodwin, Jie Xu, Nurcin Celik, and Chun-Hung Chen. Real-time digital twin-based optimization with predictive simulation learning. *Journal of Simulation (to appear)*, 2023.

[62] René Carmona and Michael Ludkovski. Valuation of energy storage: An optimal switching approach. *Quantitative Finance*, 10(4):359–374, 2010.

[63] Robert C Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, pages 141–183, 1973.

[64] Guangxin Jiang, L Jeff Hong, and Barry L Nelson. Online risk monitoring using offline simulation. *INFORMS Journal on Computing*, 32(2):356–375, 2020.

[65] Eric Fournié, Jean-Michel Lasry, Jérôme Lebuchoux, and Pierre-Louis Lions. Applications of malliavin calculus to monte-carlo methods in finance. ii. *Finance and Stochastics*, 5(2): 201–236, 2001.

[66] Catherine Daveloose, Asma Khedher, and Michèle Vanmaele. Representations for conditional expectations and applications to pricing and hedging of financial products in lévy and jump-diffusion setting. *Stochastic Analysis and Applications*, 37(2):281–319, 2019.

[67] Reuven Y Rubinstein and Alexander Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, 1993.

[68] Reuven Y Rubinstein. Sensitivity analysis of discrete event systems by the "push out" method. *Annals of Operations Research*, 39(1):229–250, 1992.

[69] Ram P Kanwal. *Generalized Functions Theory and Technique: Theory and Technique*. Springer Science & Business Media, 2012.

[70] Joseph Beyene and Rahim Moineddin. Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology*, 5 (1):1–7, 2005.

[71] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, Massachusetts, 2 edition, 2018.

[72] J.N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

[73] Harold J Kushner and G Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2003.

[74] Pierre L'Ecuyer and George Yin. Budget-dependent convergence rate of stochastic approximation. *SIAM Journal on Optimization*, 8(1):217–247, 1998.

[75] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

[76] Ashwin Rao and Tikhon Jelvis. Foundations of reinforcement learning with applications in finance, 2022. `https://stanford.edu/~ashlearn/RLForFinanceBook/book.pdf`, accessed 12nd March 2022.