Optimized Batch Policy Evaluation in the Presence of Monotone Responses

by

Wang Dong

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

In batch policy evaluation the goal is to predict the value of a policy given some historical data. A specific example, which motivated the approach pursued in this thesis, is to predict the probability of putting a natural wildfire out given some specific configuration of dispatched resources, such as helicopters, planes, trucks, and firefighters of various kinds. The general structure of problems like this is that the more resources are deployed, the higher success rate we expect to see: the response is a monotone function of the resources dispatched. In this thesis, we investigate the question of what are the best ways of estimating success probabilities of policies in problems that exhibit such a monotone structure. In particular, viewing the problem as a multiobjective optimization problem where the optimization variables are parameters describing the estimators and the objectives correspond to different problem-instances (in the wildfire application these correspond to different conditions under which the fire may happen), we propose various ways of optimizing estimation accuracy. More specifically, when resource levels are discrete and one-dimensional (or totally ordered) and focusing on minimum variance point estimation and unbiased estimators, it is found that natural optimization objectives lead to convex optimization problems that can be solved efficiently. One of the main contributions of the thesis is the careful experimental comparison of the various optimization objectives with Monte-Carlo simulations. As a second main contribution, a similar investigation is carried out for comparing alternative strategies that produce interval estimates.

Preface

Some of the research conducted for this thesis forms part of a research collaboration, led by Professor I. Lee at the University of Alberta. The technical apparatus referred to in chapter 2 was a collaboration work by my research group, Professor I. Lee, Professor C. Szepesvari, M. Rezaei and me. The other chapters are my original work. No part of this thesis has been previously published.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Lee, without whom I would not find such an interesting project. Dr. Szepesvari consistently supported me whenever I hit a wall. I would like to thank Mostafa Rezaei who worked with me most of the time. I would also like to thank the University of Alberta which gave me the opportunity to be a part of the research group. Last but not least, I would like to thank my parents and friends for their support and love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Albreta province is full of forest resources. Wildfire is one of the biggest threats in Alberta. The recent wildfires in California costs over 12 billion US dollars in 2020 [1]. Wildfires can be very expensive to fight and sometimes fires can result in the tragic loss of human life. The average damage and suppression cost of a fire could easily be more than ten thousand dollars, and sometimes it costs billions of dollars [2]. Effective suppression will quickly limit and arrest fire spread, thereby eliminating the potential for fire-caused damages and negative impacts in surrounding areas. When a new wildfire is reported, suppression resources such as airtankers, helicopters, and firefighters are dispatched. There has been some research on dispatching decisions [2–4], with limited applications. Initial attack (IA) is the set of actions taken by resources that are first to arrive at a fire [2]. The thesis is motivated by our another project, whose goal is to optimize wildfire dispatch decisions to maximize IA success while using limited resources. In wildfire application, it is nearly impossible in practice to perform experiments to evaluate the effectiveness of different combinations of suppression resources, since the consequences for real wildfires can be severe. A potential remedy to this is to use previously collected historical data to evaluate and compare success probabilities of IA policies. How to do this effectively and efficiently is the focus of research on off-policy evaluation (OPE). OPE is widely used in Computing Science and various applications. Some of the basic methods for evaluating a new policy using historical data have emerged [5–7].

In this thesis we consider a special case of this problem. Consider the problem for wildfire dispatch decisions. The dispatch decisions represent the number of resources to send. There is a higher probability for the fires to be contained as we send more resources: the success probability is monotone in the resources we send. We model this problem by introducing the notion of critical levels, which is defined as the level of resources that guarantees success. When we send the number of resources above or equal to that critical level, say 2 helicopters, we get a positive outcome. Based on limited information from fire reports, there could be many possibilities for the critical level of resources required: the critical level of resources could be hard to predict. In this thesis, we assume that the critical level is a random variable. To evaluate a new dispatching policy and to find the best one, we can utilize historical wildfire data containing dispatch decisions, and suppression outcomes.

1.2 Related Work

The literature on OPE is enormous, but they are more general. None of them mentioned here considered the monotone response case, which is the most important difference distinguishing this work from literature. Thomas and Brunskill proposed finding an estimator minimizing the mean squared error (MSE) [6]. This paper is similar to our study, which is also optimizing estimators. The difference is that, we are optimizing the variance of an unbiased estimator instead of MSE and we focus strongly on the structure of our problem. They used the weighted importance sampling (IS) estimator as a baseline and introduced several enhancements. They proposed several new OPE estimators and showed empirical results that they outperform existing estimators. They did empirical studies for the estimators. In addition to empirical results, we also show that our estimators are provably optimal for the problem we consider. Another paper by Thomas et al. in 2015 presented a technique that can evaluate a policy without executing it based on a lower confidence bound of the value of the policy to be evaluated [7]. The ability to compute a tight confidence bound comes from a novel adaptation of an existing concentration inequality to make it particularly well suited for evaluation. They performed experiments for digital marketing applications using real data. In my thesis, we also compute confidence intervals for our estimators.

Kato and Kaneko, in their 2020 preprint, considered OPE for contextual bandits when the logging policy (aka. behavior policy) evolves over time [5]. OPE in general, including this thesis, estimates the value of an evaluation policy using samples obtained under a fixed behavior policy. The cited paper proposes an estimator based on an idea of a "generalized method of moments" technique. Both of this paper and my thesis have independent and identically distributed (i.i.d.) data. In my thesis, we are optimizing variance instead of reward. Also, the bandit problem does not have monotonicity, because the reward is not monotone in actions and there is no order between actions.

To the best of our knowledge, there is no past work applying off-policy estimation to wildfire data. Although we motivate this work from wildfire application, this thesis is limited to development of off-policy estimators under monotone response, and application of the proposed estimators to wildfire data remains in future work. One of the previous research in wildfire is about IA decisions. There are two sequential decision-making stages for IA: the strategic stage at which decisions regarding the deployment of resources to operations bases are made, and the tactical-operational stage at which decisions regarding dispatching of resources to fire locations are made. Ntaimo et al formulated and implemented a two-stage standard response model (SRM) for wildfire initial attack planning [2]. The goal of the paper is to contain as many fires as possible at minimal pre-suppression costs plus expected suppression costs. To minimize the expense, they made several SRM based on simulations of historical data. They use confidence interval (CI) for various sample sizes to compare the total expense. They used a fire simulator to link dispatched resources to the outcome whereas our approach is data-driven.

Kourtz first coined a dynamic programming (DP) framework for dispatching fire forces in 1988 [3]. The paper considered the dispatching problem when resources are distributed among different places. The paper proposed a DP structure regarding minimizing the dispatching money cost for different time stages. He solved a knapsack formulation using DP to determine a set of resources that minimizes the cost and is enough to suppress the fire. Another paper by R.Wiitala which extends from Kourtz's paper by considering multiple time periods, discussed another dynamic programming based approach that offers a recursive algorithm for finding the most efficient dispatch of available suppression resources to a fire [4]. The way Wiitala solves the dispatching problem in each time period is similar to Kourtz's. They formulate the problem as that of minimizing the total cost, including suppression cost, mop-up cost and resource loss. Both of these two papers used either experts' opinion or a model for the fire line to represent the relationship between resources and suppression success whereas we use a data-driven approach.

1.3 Objectives and Contributions

In this thesis, our goal is to find a low variance unbiased estimator of the performance of a target policy using data generated by a logging policy that is potentially different from the target policy. In particular, we study the problem focused on a decision-maker who needs to choose one action out of k choices. Let $[k] = \{1, \dots, k\}$ be the set of available actions. If the agent chooses a, then the outcome is $\mathbf{1}(a \geq A^*)$, where A^* is unknown to the agent, called critical level. The goal of the agent is to choose an action that is greater or equal to the unknown critical level. We consider the critical level as a random variable from an unknown distribution P. The agent has access to i.i.d. data of the form $\{A_i, \mathbf{1}(A_i \geq A_i^*)\}$, where the distribution of A_i^* is P and the distribution of A_i is π_l, which is known and A_i and A_i^* are independent of each other. We propose various unbiased estimators of $\mathbb{E}[\mathbf{1}(A \geq A^*)]$ where A, A^* are independent, A follows π_l. A key challenge is the fact that estimators have different variances depending on the unknown critical level distribution.

The main contributions of this thesis come in three parts:

- We propose several unbiased estimators that optimize different objectives related to variance and show that they can be efficiently computed by solving convex optimization problems.
- We thoroughly compare the estimators empirically.
- We empirically evaluate several methods to obtain confidence intervals that capture the target value $\mathbb{E}[\mathbf{1}(A \geq A^*)]$ with high probability.

1.4 Thesis Outline

The rest of this document is organized as follows. Chapter 2 discusses the proposed off-policy estimators. Chapter 3 discusses the point estimation experimental results and the comparison of estimators. Chapter 4 compares several techniques to construct confidence intervals. Chapter 5 summarizes our work in this thesis and provides insights into future work.

Chapter 2

Estimators

In this section, we define our point-estimators. We first introduce the notation we will use. The agent makes decisions based on a data set as described earlier. An observation consists of an action taken and its corresponding outcome. Since the data is a sequence of n i.i.d. random variables, our estimator will take the form of the average of estimates produced on the individual observation pairs. The variance of such an estimator is the variance of the estimate based on a single observation decided by n. Hence, it suffices to consider the problem when there is only one data-point. We emphasize that the critical level is not observed in the logged data. We assume that the action in the data were chosen following $\pi_l : [k] \to [0,1]$, which is known, where $\sum_{a \in [k]} \pi_l(a) = 1$, we call π_l the logging policy.

We aim to evaluate another policy $\pi_t : [k] \to [0,1]$, called the target policy which is also part of the input. When the critical value is a^*, the value $V^\pi_{a^*}$ of a policy π is given as:

$$V^\pi_{a^*} = \sum_{a \geq a^*} \pi(a), \tag{2.1}$$

which is the probability of success of π. If p is the probability mass function of the random critical level A^*, then the value of a policy π is:

$$V^\pi_p = \sum_{a^* \in [k]} p(a^*) V^\pi_{a^*}.$$

An estimator of the value of a target policy based on a single observation assigns a real number to the observation. As such, it takes the form $f : [k] \times \{0,1\} \to \mathbb{R}$.

When an action a was taken in the logged data, our estimate of the value of π_t is $f(a,1)$ if the outcome was a success, and $f(a,0)$ if it was a failure. In order for f to be unbiased f must satisfy:

$$\sum_{a\in[k]} \pi_l(a) f(a, 1(a \geq a^*)) = V_{a^*}^{\pi_t} \text{ for } a^* \in [k]. \tag{2.2}$$

Any estimator that satisfies this equation is an unbiased estimator: the estimated value equals the true value no matter the critical level distribution. There are infinitely many unbiased estimators, since we can modify $f(a, 1(a \geq a^*))$ term without changing the value of the left hand side of the equation (2.2). For an estimator f when the critical level distribution is p, its variance is given as:

$$Var_p(f) = \sum_{a\in[k]} \sum_{a^*\in[k]} \pi_l(a) p(a^*)[f(a, 1(a \geq a^*)) - V_p^{\pi_t}]^2. \tag{2.3}$$

The variance depends on the unknown critical level distribution. We will introduce several unbiased estimators which optimize variance in different ways. Let us first review the IS estimator which we consider a baseline.

IS is a general technique in statistics for estimating properties of a distribution, while only having samples generated from a different distribution than the distribution of interest [8]. With our notation, the IS estimator takes the form

$$f(a,1) = \frac{\pi_t(a)}{\pi_l(a)}, f(a,0) = 0. \tag{2.4}$$

Clearly, this satisfies the unbiasedness condition (2.2). In what follows, we let $\mathcal{U}$ denote the set of functions f that satisfy (2.2).

2.1 Minimax Variance Estimator

Consider an unbiased estimator minimizing the worst-case variance, obtained by solving the following problem: Define Δ^{k-1} as the k-dimensional probability simplex, and solve

$$\min_{f\in\mathcal{U}} \max_{p\in\Delta^{k-1}} Var_p(f). \tag{2.5}$$

Let A denote the random variable representing the action taken by the logging policy. If f is feasible for (2.5), then we have $\mathbb{E}_A \mathbb{E}_{A^*}[f(A, 1(A \geq A^*))] = V_p^{\pi_t}$, where A^* p, A π_l. Thus, for any f that is feasible for (2.5),

$$\begin{aligned} Var_p(f) &= E_A E_{A^*}[(f(A, 1(A \geq A^*)) - V_p^{\pi_t})^2] \\ &= E_A E_{A^*}[[f(A, 1(A \geq A^*))]^2] - 2V_p^{\pi_t} E_A E_{A^*}[f(A, 1(A \geq A^*))] + (V_p^{\pi_t})^2 \\ &= E_A E_{A^*}[[f(A, 1(A \geq A^*))]^2] - (V_p^{\pi_t})^2. \end{aligned} \tag{2.6}$$

Let us introduce new notations to write (2.5) in a compact form. For any policy π, we will re-use π to denote its k-dimensional vector representation. We re-use p in the same way. Let $F^{(2)}$ be the $k \times k$ matrix whose entries are given as $F^{(2)}_{a,a^*} = [f(a, 1(a \geq a^*))]^2$. Finally, let T be the $k \times k$ matrix with $T_{a,a^*} = 1(a \geq a^*)$. Then, (2.6) can be written as:

$$\begin{aligned} Var_p(f) &= E_A E_{A^*}[[f(A, 1(A \geq A^*))]^2] - (V_p^{\pi_t})^2 \\ &= \pi_l^\intercal F^{(2)} p - (\pi_t^\intercal T p)^2. \end{aligned}$$

Thus, the inner problem of (2.5) can be written as

$$\min_{p \in \Delta^{k-1}} (\pi_t^\intercal T p)^2 - \pi_l^\intercal F^{(2)} p \tag{2.7}$$

which is minimizing a convex quadratic function on the k-dimensional unit simplex. It is a convex problem and can be solved efficiently.

Proposition 1 *The minimax problem in* (2.5) *can be reformulated as a minimization problem as follows: let α and μ be the variables in $\mathbb{R}$, F be the $k \times k$ matrix whose entries are given as $F = f(a, 1(a \geq a^*))$. Then* (2.5) *is equivalent to*

$$\textit{(P)} \min_{f,\alpha,\mu} \alpha^2 + \mu \tag{2.8}$$

$$s.t. \ \mu e^\intercal \geq \pi_l^\intercal F^{(2)} - 2\alpha \pi_t^\intercal T, \tag{2.9}$$

$$\pi_l^\intercal F = \pi_t^\intercal T. \tag{2.10}$$

(See Appendix A.1 for the proof of this proposition.)

By solving this problem, we can get the Minimax (MM) estimator which minimizes worst-case variance.

2.2 Regret Against the IS Estimator

We find another unbiased estimator which minimizes the worst-case regret against the IS estimator. We will call this the "Regret against IS" (RIS) estimator. Let f_{IS} be the weights for IS from equation (2.4). The RIS estimator is defined by

$$\min_{f \in \mathcal{U}} \max_{p \in \Delta^{k-1}} (Var_p(f) - Var_p(f_{\text{IS}})). \tag{2.11}$$

We know rewriting this minimax problem is a more convenient form. For this, let matrix S be a $k \times k$ matrix that represents the weights of the IS estimator, $S_{a,a^*} = \frac{\pi_t(a)}{\pi_l(a)}$ if $a \geq a^*$ and zero otherwise. With this, (2.11) becomes:

$$\min_{f \in \mathcal{U}} \max_{p \in \Delta^{k-1}} \pi_l^\intercal F^{(2)} p - \pi_l^\intercal S^{(2)} p. \tag{2.12}$$

The inner problem of (2.12) is:

$$\min_{p \in \Delta^{k-1}} \pi_l^\intercal S^{(2)} p - \pi_l^\intercal F^{(2)} p. \tag{2.13}$$

Since the inner problem is a linear program we have strong duality. Therefore, the primal linear program and dual linear program have the same optimal values [9]. The dual of (2.13) is

$$\begin{aligned} &\max_{\lambda \in \mathbb{R}^k, \mu \in \mathbb{R}} -\mu \\ \text{s.t. } &\mu e^\intercal - \lambda^\intercal = \pi_l^\intercal F^{(2)} - \pi_l^\intercal S^{(2)}, \\ &\lambda \geq 0. \end{aligned}$$

Therefore, problem (2.12) can be written as a minimization problem:

$$\begin{aligned} &\min_{f,\lambda,\mu} \mu \\ \text{s.t. } &\mu e^\intercal - \lambda^\intercal = \pi_l^\intercal F^{(2)} - \pi_l^\intercal S^{(2)}, \\ &\lambda \geq 0, \\ &\pi_l^\intercal F = \pi_l^\intercal T. \end{aligned}$$

Eliminating λ gives:

$$\min_{f,\mu} \mu \tag{2.14}$$

$$\text{s.t. } \mu e^\intercal \geq \pi_l^\intercal F^{(2)} - \pi_l^\intercal S^{(2)}, \tag{2.15}$$

$$\pi_l^\intercal F = \pi_l^\intercal T. \tag{2.16}$$

By solving this problem, we can get the RIS estimator which minimizes the worst-case regret against the IS estimator.

2.3 Minimax Regret Estimator

The variance of any estimator depends on the unknown critical level distribution, and we want to find an unbiased estimator that has low variance. Thus, our estimation problem can be viewed as a multi-objective optimization problem where each possible critical level distribution defines a dimension in the objective space. In this section, we introduce an estimator that is Pareto optimal for this multi-objective optimization problem. Let $\mathcal{F} = \mathbb{R}^{[k]\times\{0,1\}}$ be the set of all real-valued functions with domain $[k]\times\{0,1\}$. For $f \in \mathcal{F}$, define $Var_f : \Delta^{k-1} \to \mathbb{R}$ as the function

$$Var_f(p) = \pi_l^\intercal F^{(2)} p - (\pi_t^\intercal T p)^2, \qquad p \in \Delta^{k-1}\,, \tag{2.17}$$

where Δ^{k-1} is the k-dimensional probability simplex. Let $\mathcal{V}^k$ denote the set of all such functions and $V(p)$ is equivalent to $Var(p)$. We equip this set with the usual partial ordering of functions: for $V, V' \in \mathcal{V}^k$ we write that $V \leq V'$ if $V(p) \leq V'(p)$

holds for all $p \in \Delta^{k-1}$. We also define $V \lneq V'$ to mean that $V \leq V'$ and $V \neq V'$ (which implies that for some $p \in \Delta^{k-1}$, $V(p) < V'(p)$). In this case we say that V dominates V', or, alternatively, that V' is dominated by V. We are interested in finding the Pareto-optimal points of $\mathcal{V}^k$, i.e., those $V^* \in \mathcal{V}^k$ that are not dominated by any other points in $\mathcal{V}^k$ (and the corresponding function f). Note that finding the Pareto optimal points belong to the topic of multi-objective optimization if we view each $p \in \Delta^{k-1}$ as defining an "objective". Unlike in most of the literature on multi-objective optimization, our multi-objective optimization problem has (uncountably) infinitely many objectives.

We define the *reference-point based achievement scalarizing function* (RPASF) for a reference point, $\bar{z} \in \mathcal{V}^k$. The RPASF: $\mathcal{V}^k \to \mathbb{R}$ of $\bar{z} \in \mathcal{V}^k$ is defined via

$$s_{\bar{z}}(V) = \sup_{p \in \Delta^{k-1}} \big(V(p) - \bar{z}(p)\big) + \rho \int_{\Delta^{k-1}} \big(V(p) - \bar{z}(p)\big)\, dp\,, \tag{2.18}$$

where $\rho > 0$.

Proposition 2 *The function $s_{\bar{z}}$ is strongly increasing, that is, for any $V, \tilde{V} \in \mathcal{V}^k$,*

$$V \lneq \tilde{V} \text{ implies that } s_{\bar{z}}(V) < s_{\bar{z}}(\tilde{V})\,. \tag{2.19}$$

Proof. We have

$$\begin{aligned} s_{\bar{z}}(V) &= \sup_{p \in \Delta^{k-1}} \big(V(p) - \bar{z}(p)\big) + \rho \int_{\Delta^{k-1}} \big(V(p) - \bar{z}(p)\big)\, dp \\ &\leq s_{\bar{z}}(\tilde{V}) + \rho \int_{\Delta^{k-1}} \big(V(p) - \tilde{V}(p)\big)\, dp \\ &< s_{\bar{z}}(\tilde{V})\,, \end{aligned}$$

where the last inequality follows because $\int_{\Delta^{k-1}} \big(V(p) - \tilde{V}(p)\big)\, dp < 0$. ■ Using the previous proposition, a slight generalization of Theorem 3.5.4 of [10], gives that the solutions of

$$\begin{aligned} &\min_f \; s_{\bar{z}}(V_f) \\ &\text{s.t. } V_f \in \mathcal{V}^k \end{aligned} \tag{2.20}$$

are Pareto optimal.

For completeness the generalized theorem is as follows. While the original theorem was stated for multicriteria optimization with *finitely* many objectives. The generalization is as follows:

Theorem 3 (Theorem 3.5.4 of [10]) *If $s : \mathcal{V} \to \mathbb{R}$ is a strongly increasing achievement function then any minimizer of s is Pareto-optimal.*

(See Appendix A.2 for the proof of this theorem.)

The generalization allows any number of objectives. We study the following version of (2.20): consider

$$\min_{f \in \mathcal{U}} \left[\max_{p \in \Delta^{k-1}} (V_f(p) - V^*(p)) + \rho \int_{\Delta^{k-1}} (V_f(p) - V^*(p)) dp \right], \tag{2.21}$$

where $V^*(p) = \min_{f \in \mathcal{U}} V_f(p)$ which is the minimum variance that can be achieved by an unbiased estimator for a critical level distribution p. Here, ρ is a small positive number. The above maximum is the worst-case regret of f against $V^*(p)$, so the above problem is interpreted as finding an estimator that minimizes the worst regret plus a regularization term. Note that there may not exist an f such that $V^*(p) = V_f(p)$ for all $p \in \Delta^{k-1}$. Because of Proposition 2 and Theorem 3, the solution of (2.21) is Pareto optimal. We study the above problem step-by-step, first $V^*(p)$.

Let $f = [f_0^\intercal, f_1^\intercal]^\intercal$, where $f_{0a} = f(a, 0)$ and $f_{1a} = f(a, 1)$ for $a \in [k]$. In this section, we will analyze the problem of minimizing the variance for a given critical level distribution p, that is, $\min_{f \in \mathcal{U}} \pi_l^\intercal F^{(2)} p - \pi_t^\intercal T p$. As the second term of the objective function does not depend on f, we analyze the following instead:

$$\min_{f \in \mathcal{U}} \pi_l^\intercal F^{(2)} p. \tag{2.22}$$

Let $A \in \mathbb{R}^{k \times 2k}$, where for $a, a' \in [k]$, $A_{a,a'} = 0$ if $a \leq a'$, $\pi_l(a')$ otherwise; and $A_{a,k+a'} = 0$ if $a > a'$, $\pi_l(a')$ otherwise. Then, we have $\mathcal{U} = \{f \in \mathbb{R}^{2k} \ : \ Af = T^\intercal \pi_t\}$. Among the entries of f, $f_{0k} = f(k, 0)$ is not well-defined because when we take the action k, the outcome cannot be zero due to the problem formulation. Thus, we fix $f_{0k} = 0$ and re-define $\mathcal{U}$ by adding this constraint.

Using matrix algebra, we can show that the nullspace of A can be represented as $\{Bz \ : \ z \in \mathbb{R}^{k-1}\}$, where $B \in \mathbb{R}^{2k \times (k-1)}$ is defined as follows:

$$B = \begin{bmatrix} I_{k-1} \\ 0 \\ I_{k-1} \\ \bar{\pi}_l \end{bmatrix}$$

and $(\bar{\pi}_l)_a = -\pi_l(a)/\pi_l(k)$ for $a \in [k-1]$. Therefore, the feasible region $\mathcal{U} = \{\bar{f} + Bz \ : \ z \in \mathbb{R}^{k-1}\}$ for any $\bar{f} \in \mathcal{U}$. Now let us represent the objective function (2.22) using the vector notation f. Let $q_0(a) = \sum_{a^* > a} p(a^*)$ and $q_1(a) = \sum_{a^* \leq a} p(a^*)$. Then, the objective function is

$$\pi_l^\intercal F^{(2)} p = \sum_a \pi_l(a) \left[q_0(a) f_{0a}^2 + q_1(a) f_{1a}^2 \right] = f^\intercal D f,$$

where D is a $2k \times 2k$ diagonal matrix defined as

$$D = \begin{bmatrix} \pi_l(1)q_0(1) & & & & & \\ & \ddots & & & & \\ & & \pi_l(k)q_0(k) & & & \\ & & & \pi_l(1)q_1(1) & & \\ & & & & \ddots & \\ & & & & & \pi_l(k)q_1(k) \end{bmatrix}.$$

Then, (2.22) is written as

$$\min_{z \in \mathbb{R}^{k-1}} g(z), \tag{2.23}$$

where $g(z) = (\bar{f} + Bz)^\intercal D (\bar{f} + Bz)$. The Hessian of g is computed as

$$\nabla^2 g(z) = 2B^\intercal D B = 2\text{diag}(\pi_l(1), \ldots, \pi_l(k-1)) + 2\pi_l(k)\bar{\pi}_l \bar{\pi}_l{}^\intercal. \tag{2.24}$$

Assuming $\pi_l(a) > 0$ for any $a \in [k]$ (similar to the positivity assumption [11]), we have that the Hessian is positive definite, and thus, g is strongly convex. Thus, the solution of (2.23) is obtained by $\nabla g(z^*) = 2B^\intercal D(\bar{f} + Bz^*) = 0$, and thus,

$$z^* = -(B^\intercal DB)^{-1} B^\intercal D\bar{f}. \tag{2.25}$$

We have that (2.22) is

$$\begin{aligned}\min_{f \in \mathcal{U}} \pi_l^\intercal F^{(2)} p = \min_z g(z) = g(z^*) &= (\bar{f} + Bz^*)^\intercal D(\bar{f} + Bz^*) \\ &= \bar{f}^\intercal D\bar{f} + 2\bar{f}^\intercal DBz^* + (z^*)^\intercal B^\intercal DBz^* \\ &= \bar{f}^\intercal D\bar{f} - \bar{f}^\intercal DB(B^\intercal DB)^{-1} B^\intercal D\bar{f}.\end{aligned}$$

Now let us study the following problem computing the maximum regret:

$$\max_{p \in \Delta^{k-1}} (V_f(p) - V^*(p)). \tag{2.26}$$

Now we consider $f \in \mathcal{U}$ to be fixed. We know that there exists z such that $f = \bar{f} + Bz$. We have

$$\begin{aligned}V_f(p) - V^*(p) &= \pi_l^\intercal F^{(2)} p - \pi_t^\intercal Tp - g(z^*) + \pi_t^\intercal Tp \\ &= (\bar{f} + Bz)^\intercal D(\bar{f} + Bz) - \bar{f}^\intercal D\bar{f} + \bar{f}^\intercal DB(B^\intercal DB)^{-1} B^\intercal D\bar{f} \\ &= 2\bar{f}^\intercal DBz + z^\intercal B^\intercal DBz + \bar{f}^\intercal DB(B^\intercal DB)^{-1} B^\intercal D\bar{f} \\ &= ||(B^\intercal DB)^{1/2} z + (B^\intercal DB)^{-1/2} B^\intercal D\bar{f}||_2^2 \\ &= ||(B^\intercal DB)^{-1/2} (B^\intercal DBz + B^\intercal D\bar{f})||_2^2 \\ &= ||(B^\intercal DB)^{-1/2} B^\intercal D(Bz + \bar{f})||_2^2 = ||(B^\intercal DB)^{-1/2} B^\intercal Df||_2^2.\end{aligned}$$

From (2.24), we know that $B^\intercal DB$ depends only on the logging policy π_l. Also, $B^\intercal Df$ is a vector in $\mathbb{R}^{k-1}$ whose ath entry is $\pi_l(a)(f_{0a} - f_{1a})q_0(a) + \pi_l(a)f_{1a} - f_{1k}\pi_l(a)$. Thus, the expression in the 2-norm of the above equation is linear in p, so for a matrix $C \in \mathbb{R}^{(k-1)\times(k-1)}$ and $d \in \mathbb{R}^{k-1}$, it can be written as $V_f(p) - V^*(p) = ||Cp + d||_2^2$. Therefore, the problem (2.26) is written as

$$\max_{p \in \Delta^{k-1}} ||Cp + d||_2^2. \tag{2.27}$$

The above problem is a maximization of a convex function: such problems cannot be solved efficiently in general. However, its feasible region allows an efficient solution method as shown by the next proposition.

Proposition 4 *The problem* (2.26) *has an optimal solution at a vertex of the* $(k-1)$-*dimensional unit simplex.*

(See Appendix A.3 for the proof of this proposition.)

By the proposition, we can solve (2.26) by comparing the k vertices of the unit simplex. In general, the number of vertices of a polytope can be exponential in dimension, but for (2.26), the above proposition renders a tractable solution method because the $(k-1)$-dimensional unit simplex has only k vertices.

We tackle the problem (2.21) based on the results. We know that we can solve the inner maximization of (2.21) for a given f by comparing $V_f(p_a) - V^*(p_a)$ where p_a for $a = 1, \ldots, k$ are the k vertices of the unit simplex, i.e., the k unit vectors. Also, we know that

$$V_f(p) - V^*(p) = ||(B^\intercal DB)^{-1/2} B^\intercal Df||_2^2.$$

From (2.24), we have

$$B^\intercal DB = \mathrm{diag}(\pi_l(1), \ldots, \pi_l(k-1)) + \pi_l(k)\bar{\pi}_l\bar{\pi}_l{}^\intercal. \tag{2.28}$$

Let $\Pi_l = \mathrm{diag}(\pi_l(1), \ldots, \pi_l(k-1))$. By the Sherman-Morrison formula, we also have

$$\begin{aligned}(B^\intercal DB)^{-1} &= \Pi_l^{-1} - \frac{\Pi_l^{-1}\pi_l(k)\bar{\pi}_l\bar{\pi}_l{}^\intercal\Pi_l^{-1}}{1 + \pi_l(k)\bar{\pi}_l{}^\intercal\Pi_l^{-1}\bar{\pi}_l} \\ &= \Pi_l^{-1} - \frac{\frac{ee^\intercal}{\pi_l(k)}}{1 - \bar{\pi}_l{}^\intercal e} \\ &= \Pi_l^{-1} - \frac{\frac{ee^\intercal}{\pi_l(k)}}{1 + \frac{1-\pi_l(k)}{\pi_l(k)}} \\ &= \Pi_l^{-1} - ee^\intercal,\end{aligned}$$

where e is the $(k-1)$-dimensional one vector. Note that $(B^\intercal DB)^{-1/2}$ does not depend on either p or f. By definition, B is independent of p and f, too. D is independent of f, but depends on p. Let D_a denote the diagonal matrix D where $p = p_a$. Then, $V_f(p_a) - V^*(p_a) = ||(\Pi_l^{-1} - ee^\intercal)^{1/2} B^\intercal D_a f||_2^2$.

In the integral term of (2.21), $V^*(p)$ does not depend on f and thus it can be omitted in the minimization. Also, in $V_f(p) = \pi_l^\intercal F^{(2)} p - \pi_t^\intercal T p$, the second term is irrelevant of f so can be omitted, too. In addition, we have

$$\int_{\Delta^{k-1}} \pi_l^\intercal F^{(2)} p dp = c\pi_l^\intercal F^{(2)} e = c f^\intercal \bar{D} f,$$

for some constant c, where e is the k-dimensional one vector and $\bar{D}$ is a $2k \times 2k$ diagonal matrix whose ath diagonal entry is $(k-a)\pi_l(a)$ and $(k+a)$th diagonal entry is $a\pi_l(a)$ for $a \in [k]$ (*i.e.*, $\bar{D}$ is defined similarly as D but with e in place of p). Therefore, (2.21) is written as

$$\min_{f \in \mathcal{U}} \left[\max_{i=1,\dots,k} ||(\Pi_l^{-1} - ee^\intercal)^{1/2} B^\intercal D_a f||_2^2 + \rho' f^\intercal \bar{D} f \right],$$

where $\rho' = c\rho$. This is equivalent to

$$\min_{f \in \mathbb{R}^{2k}, t \in \mathbb{R}} t + \rho' f^\intercal \bar{D} f \tag{2.29}$$

$$\text{s.t. } t \geq ||(\Pi_l^{-1} - ee^\intercal)^{1/2} B^\intercal D_a f||_2^2 \text{ for } a = 1, \dots, k \tag{2.30}$$

$$Af = T^\intercal \pi_t, \tag{2.31}$$

which is a quadratically constrained quadratic program and convex. Therefore, the minimax regret (MMR) estimator can be obtained efficiently by solving the above convex optimization problem. The estimator is unbiased and Pareto optimal in variance in the sense explained at the beginning of this section.

Chapter 3

Experimental Result

3.1 Experimental Design

In this section, we explain our experiment setting, starting at the policies and weights. We assume that there are 10 actions, so $k = 10$. There are eleven logging and eleven target policies sharing the same distribution as shown in Figure 3.1. We will use the identifiers for policies shown on the figure. In particular, there is a uniform distribution, as well as ten unimodal distributions centered at each action. The growth rate of those distributions are linear. Except the uniform distribution, the other k distributions are symmetrical to each other. The kth distribution is the reversal of the $(11 - k)$Th distribution from 1 to 10 (for example, policy 1 is the reversal of policy 10). The minimum probability for an action is set to 1% for every policy.

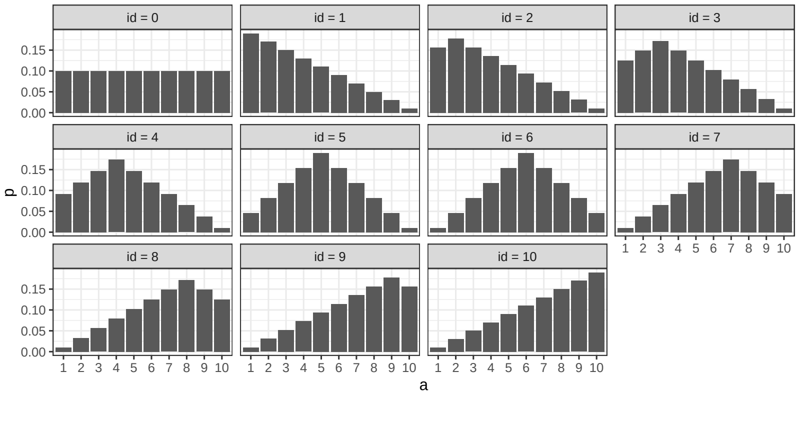
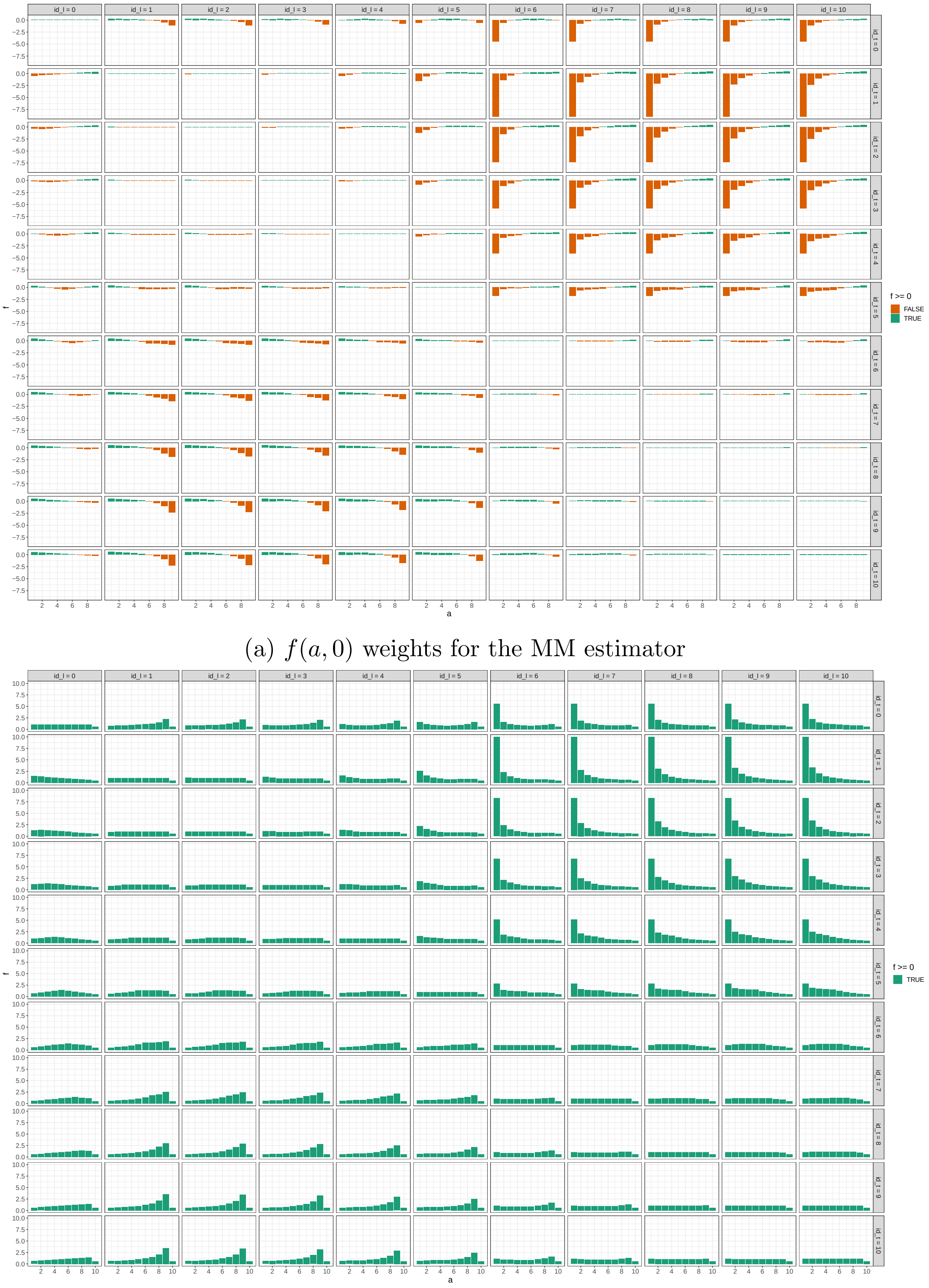


Figure 3.1: Policies



(a) $f(a, 0)$ weights for the MM estimator

(b) $f(a, 1)$ weights for the MM estimator

Figure 3.2: Weights for the MM estimator

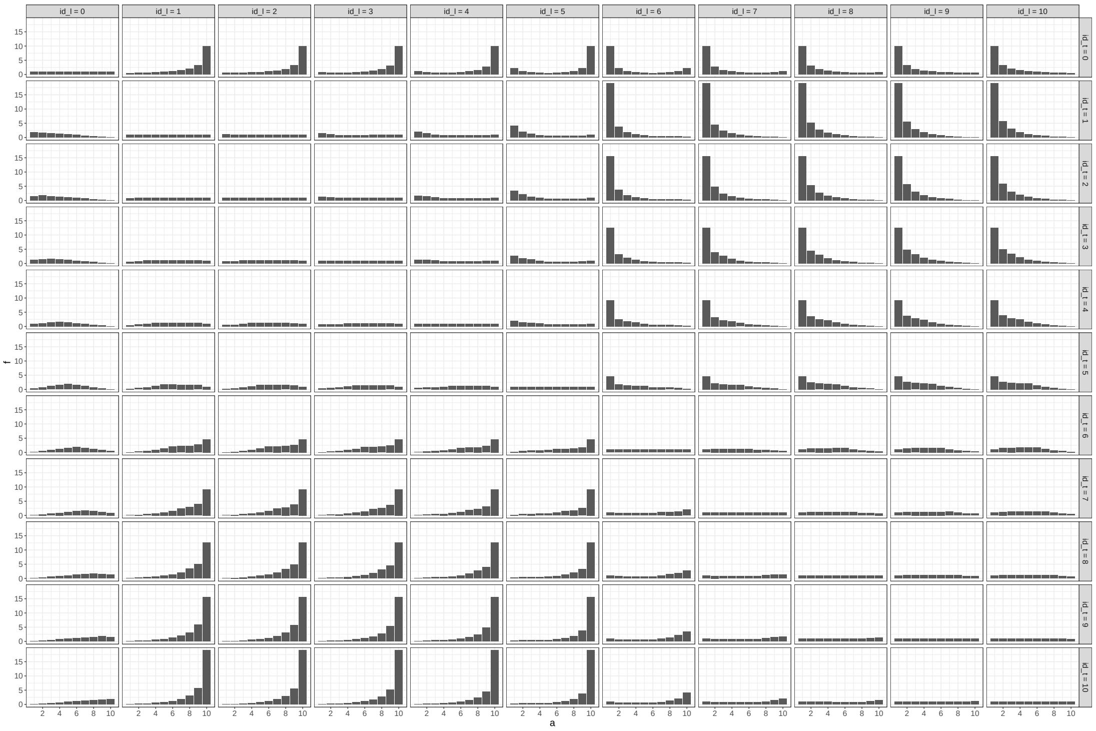
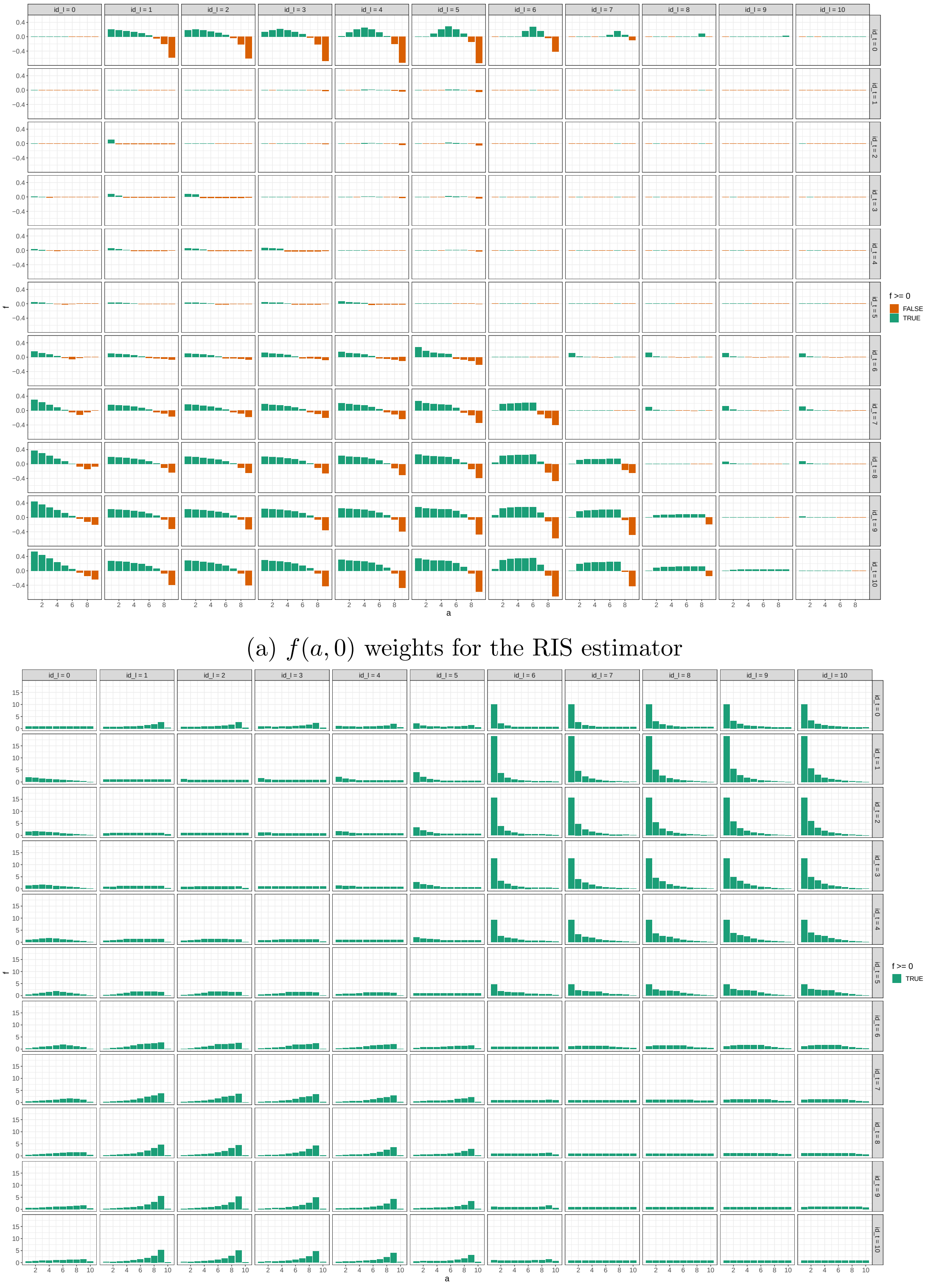


Figure 3.3: $f(a, 1)$ weights for the IS estimator

3.2 Estimator Weights

The purpose of this section is to show and compare the weights for the estimators. We want to investigate the patterns for the weights. For each pair of logging policy and target policy, we show the values (weights) for f(a,0) and f(a,1) for every possible observation based on actions and outcomes.

After solving the problem of equations (2.8) - (2.10), we obtain the weights of the MM estimator for every pair of logging and target policies. Figure 3.2 shows the weights of the MM estimator. Figure 3.2a shows the weights $f(a, 0)$; Figure 3.2b shows the weights $f(a, 1)$. Both sub-figures are organized into a matrix where a particular cell shows the weights corresponding to a particular combination of logging (columns) and target (rows) policies. The green color in Figure 3.2 means positive values and orange color means negative values. When the action a is 10, the outcome of $a \geq a^*$ would always be 1. The case of $f(10, 0)$ would never happen, since there is no value of a^* satisfying $10 < a^*$. So there are only 9 values for the $f(a, 0)$ weights to obtain. For



(a) $f(a, 0)$ weights for the RIS estimator

(b) $f(a, 1)$ weights for the RIS estimator

Figure 3.4: Weights for the RIS estimator

$f(a,1)$ weights (Figure 3.2b), the values are all positive. For $f(a,0)$ weights (Figure 3.2a), the majority of weights are negative. Whenever the $f(a,1)$ weights get a larger positive value, the corresponding $f(a,0)$ will also be a more negative value.

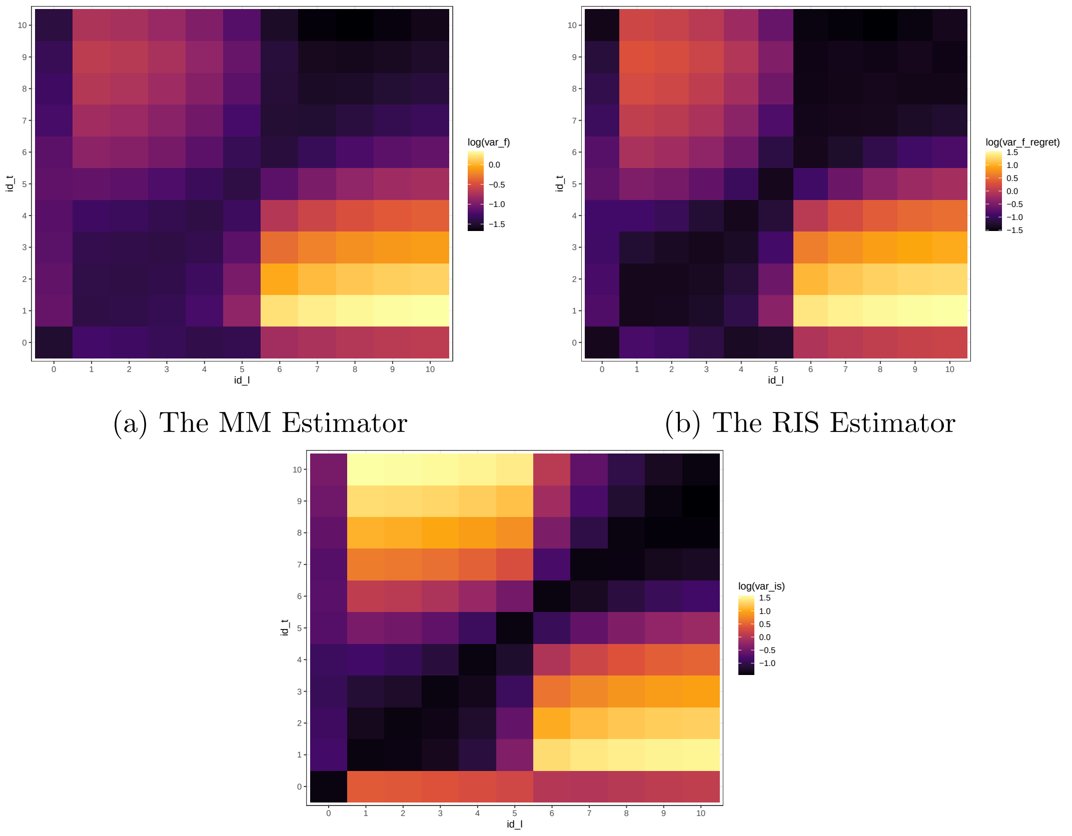
For the IS method, the results are shown in Figure 3.3. All of the $f(a,0)$ weights are zero by the definition (see Equation (2.4)). The $f(a,1)$ weights (Figure 3.3) showed a reversal pattern on top right and bottom left, since the logging and target policies are reversal to each other.

The RIS estimator weights are shown in Figure 3.4. The $f(a,1)$ weights (Figure 3.4b) are all positive which is the same as the other two estimators. And the patterns are similar to the other two estimators. For the MM estimators previously presented, $f(a,1)$ has high positive values when the corresponding $f(a,0)$ takes very small negative values. However, for the RIS estimator, we do not observe the pattern.

After obtaining the weights of the estimators, we can use those weights to compute the variance for different critical level distributions (see Equation 2.3). In the next section 3.3, we will discuss the variance comparisons of the MM, IS and RIS estimators. In section 3.4, we will compare the MMR estimator with others (MM, IS and RIS).

3.3 Variance Comparisons Results

Since the estimators we introduced in Chapter 2 are all unbiased, a natural way of comparing them is comparing their variances. However, their variance depends on the unknown critical level distribution. Thus, we compare the estimators using varying critical level distributions. We also compute the worst-case critical level distributions for each estimator (MM, IS and RIS) based on various logging-target policies pairs. Based on the worst cast critical level distributions, we can plot the worst-case variance graph based on the variance formula (2.3). Since the MM estimator guarantees the best worst-case variance, the MM estimator should outperform others when they both use the worst-case critical level distribution.



(a) The MM Estimator

(b) The RIS Estimator

(c) The IS Estimator

Figure 3.5: Heatmaps of worst-case variance for the various estimators

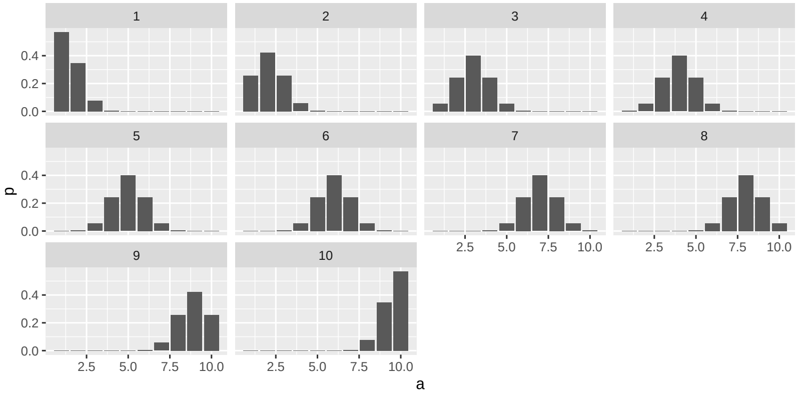


Figure 3.6: Critical Level Distributions

The Heatmaps of worst-case variance for the estimators are shown in Figure 3.5. Yellow is high variance (bad) and black is low variance (good). As before, the rows correspond to the various target policies and the columns correspond to the various logging policies. Figure 3.5a shows a heatmap of logarithm of the worst-case variance of the MM estimator. The variance is low when around the diagonal, otherwise the variance is high. Figure 3.5b and 3.5c show the heatmaps of logarithm of the worst-case variance of the RIS and IS estimators. We can see there are differences in terms of worst-case variance between the estimators. Comparing the heatmaps does not clearly show the difference between the estimators. In each subsection below, we will use tables to show the comparison of the variance for each pair of estimators by using various critical level distributions.

3.3.1 IS vs RIS

Figure 3.6, are the 10 fixed kernel critical level distributions with density 1. We call it H1 critical level distributions. We use 4 critical level distributions in total to compare the pair of estimators, the worst-case of the first estimator, the worst-case of the second estimator, 11-fixed distributions which is the same as our 11 policies (Figure 3.1), and the H1 critical level distributions 3.6.

Table 3.1 shows the variance comparison for IS and RIS estimators. The first column listed the critical level distributions we use to compare two estimators. There are 4 different critical level distribution combinations. First, we compare their variance when they are both using their worst-case critical level distributions. There are only $1 \times 11 \times 11 = 121$ scenarios since there is only one worst critical level distribution for each logging-target policy pair. Second, they are both using the worst critical level distributions for the first estimator (again 121 scenarios). Third, they are both using the worst critical level distributions for the second estimator (121 scenarios). Fourth, they are both using 11 fixed policy distributions as their critical level distributions ($11 \times 11 \times 11 = 1331$ scenarios). Mean difference is calculated by the average over different scenarios of logging policies, target policies and critical level distributions, and it is the variance difference (Variance(IS) - Variance(RIS)) for each specific cut-off. For each specific case, ρ value is calculated by the variance difference divided by the maximum of the two variances. We determined the cutoff values of ρ to be

IS vs RIS			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
Worst Case IS vs Worst Case RIS	$\rho \geq 10\%$	38.8%	1.17
	$\rho \geq 25\%$	32.2%	1.39
	$\rho \geq 50\%$	23.1%	1.80
	$\rho \geq 75\%$	9.1%	2.44
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case IS	$\rho \geq 10\%$	38.8%	1.19
	$\rho \geq 25\%$	33.1%	1.37
	$\rho \geq 50\%$	25.6%	1.66
	$\rho \geq 75\%$	9.9%	2.31
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case RIS	$\rho \geq 10\%$	38.8%	1.14
	$\rho \geq 25\%$	31.4%	1.39
	$\rho \geq 50\%$	22.3%	1.82
	$\rho \geq 75\%$	8.3%	2.59
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using 11 fixed (policy)	$\rho \geq 10\%$	37.7%	1.16
	$\rho \geq 25\%$	34.0%	1.27
	$\rho \geq 50\%$	25.5%	1.61
	$\rho \geq 75\%$	8.3%	2.21
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA

Table 3.1: Variance Comparison for IS and RIS (Cutoff)

10%, 25%, 50%, 75%, -10%, -25%, -50% and -75%. If rho is positive, then it means RIS is better. The percentage is calculated by the number of the scenarios that reach the cutoff, then divided by the total number of scenarios. “NA” appears in the mean

IS vs RIS				
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5
Worst vs	IS $\geq$ RIS	98.3%	100%	92%
Worst	IS $\leq$ RIS	1.65%	0%	8%
	Mean difference	0.46	1.81	0
	Mean ρ	21.8%	59.7%	0%
Worst IS	IS $\geq$ RIS	98.3%	100%	92%
	IS $\leq$ RIS	1.65%	0%	8%
	Mean difference	0.46	1.82	0
	Mean ρ	22.9%	59.8%	0%
Worst RIS	IS $\geq$ RIS	98.3%	100%	92%
	IS $\leq$ RIS	1.65%	0%	8%
	Mean difference	0.45	1.81	0
	Mean ρ	20.9%	59.7%	0%
11 fixed	IS $\geq$ RIS	98.6%	100%	95.64%
	IS $\leq$ RIS	1%	0%	4.36%
	Mean difference	0.44	1.76	0
	Mean ρ	23.0%	62.2%	0%

Table 3.2: Variance Comparison for IS and RIS (Scenarios)

difference column when there is no such case. Since RIS tries to minimize the variance difference against IS, there is not a single case in which IS performs 10% better than RIS estimator. Overall, the RIS performs better than the IS estimator.

In Table 3.1, all results are aggregated over all logging and target policies. We also analyzed for what kind of combinations of logging and target policies one estimator is better than the other. Table 3.2 shows the comparison for general scenarios as well as two corner scenarios. The critical level distributions on the first column are the same as Table 3.1, but the name is shorter (for example Worst vs Worst means Worst Case

IS vs RIS			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
H1 10 fixed Distributions	$\rho \geq 10\%$	42.9%	1.03
	$\rho \geq 25\%$	36.1%	1.21
	$\rho \geq 50\%$	28.8%	1.80
	$\rho \geq 75\%$	14.4%	1.97
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
H1 5 of 10 fixed Distributions (1-5)	$\rho \geq 10\%$	40.4%	1.05
	$\rho \geq 25\%$	34.6%	1.22
	$\rho \geq 50\%$	28.6%	1.41
	$\rho \geq 75\%$	14.4%	1.83
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
H1 5 of 10 fixed Distributions (6-10)	$\rho \geq 10\%$	45.4%	1.01
	$\rho \geq 25\%$	31.4%	1.20
	$\rho \geq 50\%$	22.3%	1.50
	$\rho \geq 75\%$	14.4%	2.12
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA

Table 3.3: Variance Comparison for IS and RIS - H1 (Cutoff)

IS vs Worse Case RIS). Unlike the previous table, this one shows the percentage of scenarios for which one estimator is better than the other. The percentage calculated

IS vs RIS							
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5	L1-5, T1-5	L6-10, T6-10	Diagonal
10 fixed H1	IS $\geq$ RIS	95.8%	100%	82%	100%	100%	100%
	IS $\leq$ RIS	4%	0%	8%	0%	0%	0%
	Mean difference	0.44	1.76	0	0.01	0.1	0
	Mean ρ	21.8%	59.7%	0%	6%	25%	0%
5 fixed H1 (1-5)	IS $\geq$ RIS	97.8%	100%	91.2%	100%	100%	100%
	IS $\leq$ RIS	2%	0%	8.8%	0%	0%	0%
	Mean difference	0.43	1.73	0	0.01	0.08	0
	Mean ρ	25%	63%	0%	3%	28%	0%
5 fixed H1 (6-10)	IS $\geq$ RIS	93.8%	100%	72.8%	100%	100%	100%
	IS $\leq$ RIS	6%	0%	27.2%	0%	0%	0%
	Mean difference	0.45	1.79	0	0.01	0.12	0
	Mean ρ	27%	68%	0%	9%	23%	0%

Table 3.4: Variance Comparison for IS and RIS - H1 (Scenarios)

by the number of scenarios of variance for one estimator is either greater or smaller than the other, divided by the total scenarios in the corresponding scenario. For example, for the first row, the 98.3% means for all the scenarios that the IS estimator has greater variance compared to RIS (RIS is better). “L” and “T” in the table means the logging policy ID and target policy IS. So, “L1-5, T6-10” are the top left scenarios in Figure 3.5 and “L6-10, T1-5” are the bottom right scenarios. The result indicates that the RIS estimator outperforms IS over 98% of all the scenarios ($10 \times 11 \times 11 = 1210$ scenarios) in general in terms of variance. For top left scenarios, RIS completely beats IS for all the scenarios.

Additionally, we analyzed for what kind of critical level distributions one estimator is better than the other. Since the H1 critical level distribution does not include uniform distribution and the critical level distribution is more concentrated, we use H1 critical level distributions to do the following analysis. Table 3.3 shows the result of comparing IS and RIS using H1 critical level distributions. There are three parts in total: using all 10 H1 critical level distributions (Figure 3.6) as well as only half of the H1 critical level distributions, H1(1-5) or h1(6-10), with all the combinations for logging policies and target policies. H1(1-5) means the 5 critical level distributions from 1 to 5 in Figure 3.6. The structure of table 3.3 is similar to table 3.1.

IS vs MM			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
Worst Case IS vs Worst Case MM	$\rho \geq 10\%$	83.5%	1.11
	$\rho \geq 25\%$	68.6%	1.33
	$\rho \geq 50\%$	50.4%	1.73
	$\rho \geq 75\%$	21.5%	2.35
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case IS	$\rho \geq 10\%$	85.1%	1.16
	$\rho \geq 25\%$	74.4%	1.32
	$\rho \geq 50\%$	56.2%	1.69
	$\rho \geq 75\%$	22.3%	2.3
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case MM	$\rho \geq 10\%$	53.7%	1.14
	$\rho \geq 25\%$	39.7%	1.39
	$\rho \geq 50\%$	28.1%	1.82
	$\rho \geq 75\%$	19%	2.59
	$\rho \leq -10\%$	9.1%	-0.08
	$\rho \leq -25\%$	3.3%	-0.15
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using 11 fixed (policy)	$\rho \geq 10\%$	48.7%	1.08
	$\rho \geq 25\%$	39.8%	1.31
	$\rho \geq 50\%$	30.1%	1.69
	$\rho \geq 75\%$	20.7%	2.25
	$\rho \leq -10\%$	32.2%	-0.29
	$\rho \leq -25\%$	24.5%	-0.37
	$\rho \leq -50\%$	17.9%	-0.46
	$\rho \leq -75\%$	13.2%	-0.54

Table 3.5: Variance Comparison for IS and MM (Cutoff)

Table 3.4 is similar to 3.2, but it has more columns. The diagonal column includes the scenarios when the target policies are the same as logging policies. We also considered near diagonal scenarios in table 3.4 which are bottom left ("L1-5, T1-5") and top right ("L6-10, T6-10") part in Figure 3.5. In general, the RIS still outperforms IS estimator. However, for bottom right scenarios which indicated by the column "L6-10, T1-5", IS and RIS estimators can perform similarly. For that specific case, the average for the mean difference of variance and ρ are almost 0.

3.3.2 IS vs MM

IS vs MM				
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5
Worst vs	IS $\geq$ MM	100%	100%	100%
Worst	IS $\leq$ MM	0%	0%	0%
	Mean difference	0.93	2.29	1.52
	Mean ρ	47%	80.4%	58.5%
Worst IS	IS $\geq$ MM	98.3%	100%	92%
	IS $\leq$ MM	0%	0%	0%
	Mean difference	0.99	2.30	1.74
	Mean ρ	22.9%	59.8%	0%
Worst MM	IS $\geq$ MM	78.5%	100%	80%
	IS $\leq$ MM	21.5%	0%	20%
	Mean difference	0.52	2.05	0.1
	Mean ρ	26.7%	77.2%	6.6%
11 fixed	IS $\geq$ MM	60%	100%	12.4%
	IS $\leq$ MM	40%	0%	87.6%
	Mean difference	0.43	2.14	-0.37
	Mean ρ	14.7%	80%	-41%

Table 3.6: Variance Comparison for IS and MM (Scenarios)

IS vs MM			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
H1 10 fixed Distributions	$\rho \geq 10\%$	54.4%	1.04
	$\rho \geq 25\%$	47.9%	1.2
	$\rho \geq 50\%$	37.2%	1.5
	$\rho \geq 75\%$	22.7%	2.09
	$\rho \leq -10\%$	33.7%	-0.3
	$\rho \leq -25\%$	28%	-0.31
	$\rho \leq -50\%$	20.7%	-0.42
	$\rho \leq -75\%$	13.3%	-0.49
H1 5 of 10 fixed Distributions (1-5)	$\rho \geq 10\%$	62.4%	0.99
	$\rho \geq 25\%$	56.6%	1.10
	$\rho \geq 50\%$	43.6%	1.40
	$\rho \geq 75\%$	24.4%	2.03
	$\rho \leq -10\%$	25.2%	-0.21
	$\rho \leq -25\%$	18.4%	-0.21
	$\rho \leq -50\%$	10.4%	-0.33
	$\rho \leq -75\%$	3.6%	-0.37
H1 5 of 10 fixed Distributions (6-10)	$\rho \geq 10\%$	46.4%	1.09
	$\rho \geq 25\%$	39.2%	1.30
	$\rho \geq 50\%$	30.8%	1.61
	$\rho \geq 75\%$	21%	2.15
	$\rho \leq -10\%$	42.2%	-0.38
	$\rho \leq -25\%$	37.6%	-0.42
	$\rho \leq -50\%$	31%	-0.5
	$\rho \leq -75\%$	23%	-0.62

Table 3.7: Variance Comparison for IS and MM - H1 (Cutoff)

We use the same comparing method for critical level distributions to compare the MM estimator and the IS estimator. Table 3.5 and 3.6 shows the variance comparison

IS vs MM							
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5	L1-5, T1-5	L6-10, T6-10	Diagonal
10 fixed H1	IS $\geq$ MM	61%	100%	24%	56%	63%	64%
	IS $\leq$ MM	39%	0%	76%	44%	37%	36%
	Mean difference	0.44	2.12	-0.31	0.01	0.1	0.01
	Mean ρ	13.3%	82%	-47%	-2%	18%	-1%
5 fixed H1	IS $\geq$ MM	67.6%	100%	48%	58%	59%	28%
(1-5)	IS $\leq$ MM	32%	0%	52%	42%	41%	72%
	Mean difference	0.54	2.2	0.03	0.03	0.08	-0.01
	Mean ρ	27.4%	86%	-5%	6%	17%	-16%
5 fixed H1	IS $\geq$ MM	54%	100%	0%	54%	67%	100%
(6-10)	IS $\leq$ MM	46%	0%	100%	46%	33%	0%
	Mean difference	0.34	2.05	-0.64	-0.01	0.11	0.02
	Mean ρ	-0.8%	79%	-90%	-11%	18%	15%

Table 3.8: Variance Comparison for IS and MM - H1 (Scenarios)

for IS and MM estimators. The results in Tables 3.5 and 3.6 are obtained in the same way as in Tables 3.1 and 3.2, respectively. The results showed that in general, MM has lower variance than IS. Even for some scenarios when IS has lower variance against MM estimator, the average difference is not big.

Based on the comparison for H1 critical level distributions in table 3.7 and 3.8, we can see the variance differences for various scenarios. For most scenarios, MM estimator has either similar or lower variance compared to IS estimator. The differences in Table 3.7 are not as big as those in Table 3.3 which compared IS and RIS. However, for the case indicated by the column “L6-10, T1-5” in Table 3.8, IS estimator outperforms MM estimator on average.

3.3.3 RIS vs MM

Now we know that RIS and MM estimators both outperform IS in general. Then we use the same technique to compare the RIS and the MM estimators. Table 3.9 shows the variance comparison for RIS and MM estimators using the same method as Table 3.1. If $\rho > 0$, the MM estimator is better. The MM estimator completely outperforms RIS whenever they are both using their worst-case critical level distri-

RIS vs MM			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
Worst Case RIS vs Worst Case MM	$\rho \geq 10\%$	75.2%	0.62
	$\rho \geq 25\%$	68.6%	0.84
	$\rho \geq 50\%$	50.4%	1.2
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case RIS	$\rho \geq 10\%$	77.7%	0.68
	$\rho \geq 25\%$	58.7%	0.88
	$\rho \geq 50\%$	39.7%	1.22
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	0	NA
	$\rho \leq -25\%$	0	NA
	$\rho \leq -50\%$	0	NA
	$\rho \leq -75\%$	0	NA
Both Using Worst Case MM	$\rho \geq 10\%$	39.7%	0.21
	$\rho \geq 25\%$	21.5%	0.32
	$\rho \geq 50\%$	0	NA
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	11.6%	-0.08
	$\rho \leq -25\%$	4.1%	-0.16
	$\rho \leq -50\%$	1.7%	-0.16
	$\rho \leq -75\%$	0	NA
Both Using 11 fixed (policy)	$\rho \geq 10\%$	35.5%	0.26
	$\rho \geq 25\%$	24.4%	0.35
	$\rho \geq 50\%$	10%	0.53
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	38.5%	-0.26
	$\rho \leq -25\%$	27.1%	-0.34
	$\rho \leq -50\%$	18.8%	-0.45
	$\rho \leq -75\%$	13.9%	-0.53

Table 3.9: Variance Comparison for RIS and MM (Cutoff)

bution, since MM estimator guarantees minimum worst-case variance. However, for the 11 fixed distributions, those two estimators showed a similar performance. For

RIS vs MM				
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5
Worst vs Worst	RIS $\geq$ MM	100%	100%	100%
	RIS $\leq$ MM	0%	0%	0%
	Mean difference	0.47	0.48	1.52
	Mean ρ	32.9%	57.41%	58.52%
Worst RIS	RIS $\geq$ MM	100%	100%	100%
	RIS $\leq$ MM	0%	0%	0%
	Mean difference	0.53	0.5	1.74
	Mean ρ	26.7%	77.2%	6.6%
Worst MM	RIS $\geq$ MM	70.2%	100%	80%
	RIS $\leq$ MM	29.8%	0%	20%
	Mean difference	0.08	0.29	0.1
	Mean ρ	9.63%	38.9%	6.6%
11 fixed	RIS $\geq$ MM	51%	100%	12.4%
	RIS $\leq$ MM	49%	0%	87.6%
	Mean difference	-0.01	0.38	-0.37
	Mean ρ	-1.9%	47.2%	-40.9%

Table 3.10: Variance Comparison for RIS and MM (Scenarios)

13.9% scenarios, the variance of MM is worse than RIS. Table 3.10 shows the variance comparison by scenarios. Generally, for scenarios "L1-5, T6-10", MM estimator perform better.

Then we use H1 critical level distributions in Table 3.11 and Table 3.12 to get more detailed information. In general, the RIS and the MM estimators perform similar on average. The performance of the RIS and the MM are various for different critical level distributions and logging policy - target policy pairs. When critical level distribution is heavier on left (H1(1-5)), the MM estimator perform better on average. On the

RIS vs MM			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
H1 10 fixed Distributions	$\rho \geq 10\%$	40.2%	0.27
	$\rho \geq 25\%$	32%	0.33
	$\rho \geq 50\%$	17.5%	0.42
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	41.5%	-0.26
	$\rho \leq -25\%$	33%	-0.31
	$\rho \leq -50\%$	23%	-0.42
	$\rho \leq -75\%$	15%	-0.46
H1 5 of 10 fixed Distributions (1-5)	$\rho \geq 10\%$	51.4%	0.34
	$\rho \geq 25\%$	45%	0.38
	$\rho \geq 50\%$	29%	0.5
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	33%	-0.18
	$\rho \leq -25\%$	24.4%	-0.23
	$\rho \leq -50\%$	14%	-0.33
	$\rho \leq -75\%$	5.4%	-0.31
H1 5 of 10 fixed Distributions (6-10)	$\rho \geq 10\%$	29%	0.19
	$\rho \geq 25\%$	19%	0.27
	$\rho \geq 50\%$	5.8%	0.34
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	50%	-0.33
	$\rho \leq -25\%$	41.6%	-0.40
	$\rho \leq -50\%$	32.6%	-0.51
	$\rho \leq -75\%$	24.6%	-0.6

Table 3.11: Variance Comparison for RIS and MM - H1 (Cutoff)

other hand, when critical level distribution is heavier on right (H1(6-10)), the RIS estimator perform better on average. The MM estimator perform better on "L1-5,

RIS vs MM							
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5	L1-5, T1-5	L6-10, T6-10	Diagonal
10 fixed H1	RIS $\geq$ MM	49.7%	91.6%	24%	49%	33%	64%
	RIS $\leq$ MM	50%	8.3%	76%	51%	67%	36%
	Mean difference	0	0.36	-0.31	0.00	0.0	0.01
	Mean ρ	-6%	45.4%	-47.4%	-6%	-11%	-1%
5 fixed H1 (1-5)	RIS $\geq$ MM	67.6%	100%	48%	58%	59%	28%
	RIS $\leq$ MM	32%	0%	52%	42%	41%	72%
	Mean difference	0.54	2.2	0.03	0.03	0.08	-0.01
	Mean ρ	27.4%	86%	-5%	6%	17%	-16%
5 fixed H1 (6-10)	RIS $\geq$ MM	41.2%	83.2%	0%	44%	43%	100%
	RIS $\leq$ MM	59%	16.8%	100%	56%	57%	0%
	Mean difference	-0.1	0.26	-0.64	-0.02	-0.01	0.02
	Mean ρ	-22%	30%	-89.8%	-16%	-5%	15%

Table 3.12: Variance Comparison for RIS and MM - H1 (Scenarios)

T6-10" and the RIS perform better on "L6-10, T1-5" logging policy - target policy pairs on average. The two estimators shows a similar performance near diagonal ("L1-5, T1-5", "L6-10, T6-10", "Diagonal"). Overall, there is no clear evidence for one estimator is better than another between MM and RIS.

3.3.4 MM vs MMR and RIS vs MMR

In this subsection, we compare the MMR estimator with the MM estimator and the RIS estimator. First, we compare the MM estimator to MMR estimator using H1 critical level distribution. Table 3.13 and 3.14 shows the detailed comparison result between the MM estimator and the MMR estimator. Table 3.13 and 3.14 are similar to 3.3 and 3.4. Positive difference and ρ means the MMR estimator is better. The result shows MMR outperforms MM on average. The result also shows there are not many differences whether the critical level distribution is concentrated on left (H1(1-5)) or right (H1(6-10)).

Table 3.15 and 3.16 shows the comparison result for the RIS estimator and MMR estimator. Again, positive value means the MMR estimator is better. The comparison between the RIS estimator and MMR estimator is quite similar to the comparison

MM vs MMR			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
H1 10 fixed Distributions	$\rho \geq 10\%$	63.8%	0.08
	$\rho \geq 25\%$	36.6%	0.09
	$\rho \geq 50\%$	6%	0.11
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	12.7%	-0.04
	$\rho \leq -25\%$	7.5%	-0.05
	$\rho \leq -50\%$	2.1%	-0.06
	$\rho \leq -75\%$	0	NA
H1 5 of 10 fixed Distributions (1-5)	$\rho \geq 10\%$	63.6%	0.07
	$\rho \geq 25\%$	34.5%	0.09
	$\rho \geq 50\%$	5.7%	0.13
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	15.4%	-0.04
	$\rho \leq -25\%$	8.4%	-0.05
	$\rho \leq -50\%$	2.1%	-0.05
	$\rho \leq -75\%$	0	NA
H1 5 of 10 fixed Distributions (6-10)	$\rho \geq 10\%$	64%	0.08
	$\rho \geq 25\%$	38.7%	0.09
	$\rho \geq 50\%$	6.8%	0.10
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	10.1%	-0.04
	$\rho \leq -25\%$	6.6%	-0.05
	$\rho \leq -50\%$	2.1%	-0.06
	$\rho \leq -75\%$	0	NA

Table 3.13: Variance Comparison for MM and MMR - H1 (Cutoff)

between RIS estimator and MM estimator (Table 3.11 and 3.12). The RIs estimator is better on average when the critical level is concentrated on right (H1(6-10)) and

MM vs MMR							
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5	L1-5, T1-5	L6-10, T6-10	Diagonal
10 fixed H1	MM $\geq$ MMR	80.5%	84.3%	89.3%	75.6%	73.3%	66.3%
	MM $\leq$ MMR	19%	16%	11%	24%	27%	33%
	Mean difference	0.05	0.04	0.06	0.05	0.04	0.04
	Mean ρ	16%	13.4%	13.4%	21.2%	16.5%	14.3%
5 fixed H1	MM $\geq$ MMR	77.7%	93.3%	78.7%	84.8%	58.9%	60%
(1-5)	MM $\leq$ MMR	22%	7%	21%	15%	41%	40%
	Mean difference	0.05	0.04	0.06	0.08	0.02	0.04
	Mean ρ	14%	14%	10%	33%	6%	11%
5 fixed H1	MM $\geq$ MMR	83.3%	75.3%	100%	66.4%	87.8%	72.7%
(6-10)	MM $\leq$ MMR	17%	25%	0%	34%	12%	27%
	Mean difference	0.05	0.05	0.07	0.03	0.06	0.05
	Mean ρ	17%	12%	16.8%	10%	27%	17%

Table 3.14: Variance Comparison for MM and MMR - H1 (Scenarios)

L6-10,T1-5. On average, the MMR estimator is better than the RIS estimator.

We have seen the result shows the MMR estimator outperforms the MM estimator and the RIS estimator. Since both the MM estimator and the RIS estimator outperform the IS estimator, we can say that the MMR estimator is the best estimator in terms of average variance over the four estimators (MMR, IS, RIS, MM).

3.3.5 Combined Results

After the comparison of the four estimators: IS, RIS, MM and MMR. We know that in general, the RIS and MM outperform IS. The MMR estimator outperforms both the MM estimator and RIS estimator. The MM estimator guarantees the best worst-case variance. Table 3.17, shows the combined comparison for each pair of estimators using H1 fixed critical level distributions. The estimator name in the table shows which estimator is better for the corresponding case. The percentage in parentheses shows the percentage of scenarios in which the estimator beats the others in terms of variance.

RIS vs MMR			
Critical Level Distribution	Cutoff	Percentage	Mean Difference
H1 10 fixed Distributions	$\rho \geq 10\%$	57.9%	0.26
	$\rho \geq 25\%$	51.1%	0.29
	$\rho \geq 50\%$	29.8%	0.41
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	33.4%	-0.29
	$\rho \leq -25\%$	30.2%	-0.32
	$\rho \leq -50\%$	22.4%	-0.40
	$\rho \leq -75\%$	13.8%	-0.52
H1 5 of 10 fixed Distributions (1-5)	$\rho \geq 10\%$	68.0%	0.31
	$\rho \geq 25\%$	60.6%	0.34
	$\rho \geq 50\%$	39.8%	0.44
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	23.3%	-0.24
	$\rho \leq -25\%$	20%	-0.27
	$\rho \leq -50\%$	12.4%	-0.35
	$\rho \leq -75\%$	5.1%	-0.43
H1 5 of 10 fixed Distributions (6-10)	$\rho \geq 10\%$	47.8%	0.20
	$\rho \geq 25\%$	41.5%	0.22
	$\rho \geq 50\%$	19.8%	0.36
	$\rho \geq 75\%$	0	NA
	$\rho \leq -10\%$	43.6%	-0.33
	$\rho \leq -25\%$	40.3%	-0.35
	$\rho \leq -50\%$	32.6%	-0.42
	$\rho \leq -75\%$	22.5%	-0.54

Table 3.15: Variance Comparison for RIS and MMR - H1 (Cutoff)

RIS vs MMR							
Critical Level Distribution	Comparing Items	All Cases	L1-5, T6-10	L6-10,T1-5	L1-5, T1-5	L6-10, T6-10	Diagonal
10 fixed H1	RIS $\geq$ MMR	62.3%	86%	24.6%	70%	68.9%	70%
	RIS $\leq$ MMR	38%	14%	75%	30%	11%	30%
	Mean difference	0.05	0.35	-0.24	0.05	0.05	0.05
	Mean ρ	6%	45.2%	-45.3%	16%	11%	16%
5 fixed H1	RIS $\geq$ MMR	73.1%	100%	44%	86.4%	65.6%	60%
(1-5)	RIS $\leq$ MMR	27%	0%	56%	14%	34%	40%
	Mean difference	0.16	0.44	0.03	0.09	0.06	0.03
	Mean ρ	22%	63%	-12%	36%	7%	2%
5 fixed H1	RIS $\geq$ MMR	51.7%	72%	5%	53.6%	72%	80%
(6-10)	RIS $\leq$ MMR	48%	18%	95%	46%	28%	20%
	Mean difference	-0.05	0.25	-0.51	0.01	0.05	0.07
	Mean ρ	-9.4%	27%	-78.8%	-4.2%	14.6%	29%

Table 3.16: Variance Comparison for RIS and MMR - H1 (Scenarios)

3.4 Comparing Regret Variance Against the Best Variance

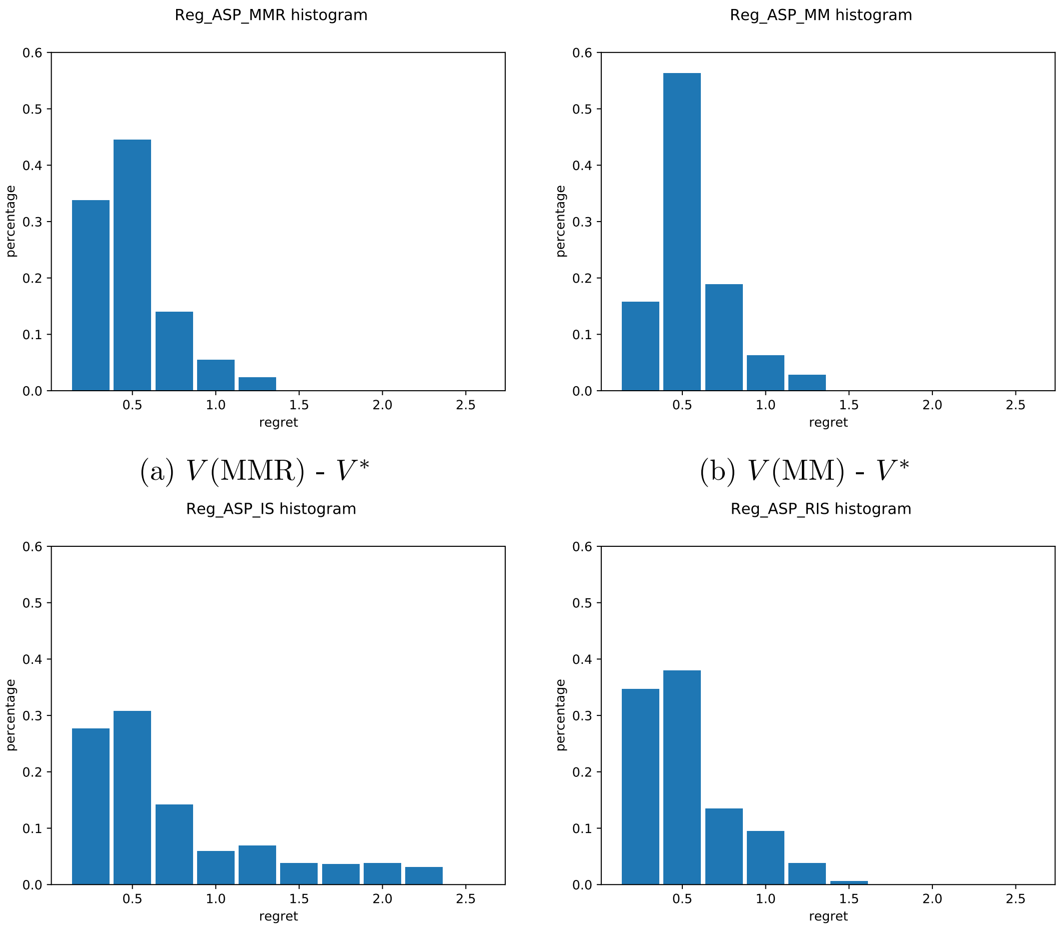
In this section, we introduce the experimental results for the MMR estimator along with comparisons. The MMR estimator aim to minimize the worst case regret against the best variance (V^*) for each critical level distribution. The MMR estimator should have the smallest difference compared to V^* in the worst-case. Then we use an experiment to compare the regret of variance against V^* for the MMR and other estimators. For logging and target policies, we use the same 11 distributions as before (Figure 3.1). We are also using the H1 critical level distributions (Figure 3.6). We generate the histogram of the regret variance for each of the four estimators with all the combinations of logging policies, target policies and critical level distributions.

Figure 3.7 shows the distributions of the regret variance for each estimator. All the regret are computed by subtracting the variance of target estimator by the fixed V^*. By subtracting the regret, we get the actual variance difference for each estimator. The x-Axis is the regret of the variance and the y-Axis is the percentage value for each range of regret value. The value on x-Axis is the upper-bound for each range, and the range is 0.25. For example, the bar at 0.5 means the regret from 0.25 to 0.5. The

Variance Comparisons		
IS vs MM	H1 (1-5)	H1(6-10)
L1-5, T1-5	MM(58%)	MM(54%)
L1-5, T6-10	MM(100%)	MM(100%)
L6-10, T1-5	50-50	IS(100%)
L6-10, T6-10	MM(59%)	MM(67%)
Diagonal	IS(72%)	MM(100%)
IS vs RIS	H1 (1-5)	H1(6-10)
L1-5, T1-5	RIS(100%)	RIS(100%)
L1-5, T6-10	RIS(100%)	RIS(100%)
L6-10, T1-5	RIS(91%)	RIS(72%)
L6-10, T6-10	RIS(100%)	RIS(100%)
Diagonal	RIS(100%)	RIS(100%)
RIS vs MM	H1 (1-5)	H1(6-10)
L1-5, T1-5	MM(54%)	RIS(56%)
L1-5, T6-10	MM(100%)	MM(83%)
L6-10, T1-5	50-50	RIS(100%)
L6-10, T6-10	RIS(77%)	RIS(57%)
Diagonal	RIS(72%)	MM(100%)
MM vs MMR	H1 (1-5)	H1(6-10)
L1-5, T1-5	MMR(84%)	MMR(66%)
L1-5, T6-10	MMR(93%)	MMR(75%)
L6-10, T1-5	MMR(79%)	MMR(100%)
L6-10, T6-10	MMR(59%)	MMR(88%)
Diagonal	MMR(60%)	MMR(73%)
RIS vs MMR	H1 (1-5)	H1(6-10)
L1-5, T1-5	MMR(86%)	MMR(54%)
L1-5, T6-10	MMR(100%)	MMR(72%)
L6-10, T1-5	RIS(56%)	RIS(95%)
L6-10, T6-10	MMR(66%)	MMR(72%)
Diagonal	MMR(60%)	MMR(80%)

Table 3.17: Variance Comparison Conclusion - H1

mode of the regret for all four estimators are at 0.25-0.5 range. The MMR method has the thinnest tail. For the tail part (regret > 0.5), MMR has smallest percentage



(a) V(MMR) - V^*

(b) V(MM) - V^*

(c) V(IS) - V^*

(d) V(RIS) - V^*

Figure 3.7: Regret Variance for Estimators Against V^*

22%; RIS has percentage of 27%; MM has percentage of 28%; IS has worst tail of 41%.

Variance Difference for MMR vs Others (IS, RIS and MM)				
MMR vs Others	Minimum	Average	Maximum	Range(Max - Min)
V(MMR)-V(MM)	-0.26	-0.05	0.18	0.44
V(MMR)-V(RIS)	-0.93	-0.05	1.02	1.95
V(MMR)-V(IS)	-1.75	-0.24	1.02	2.77

Table 3.18: Variance Comparison Conclusion for MMR Against Others - H1

In table 3.18, the minimum, average, maximum and range are calculated based on 1210 scenarios using H1 critical level distribution, 11 target policies and 11 logging policies. As can be seen from Table 3.18, the average variance differences between the MMR estimator and others are all negative. So, the MMR estimator gets the smallest variance on average. The MM estimator gets the smallest range for the variance difference.

3.5 Conclusion

From section 3.3 and section 3.4, we showed that the MMR has the lowest average variance compared to IS, RIS and MM. So, in general, the MMR is the best estimator in terms of average variance. MM estimator is slightly worse than MMR in terms of average variance but MM is the second best estimator compared to MMR and it guarantees the minimum worst-case variance. So, both the MM and MMR estimator could be useful depending on whether we want to minimize the worst variance or average variance.

Chapter 4

Confidence Intervals

So far, we have studied only point estimates of the values of target policies. Those point estimators involve uncertainty because they are based on a limited amount of observed data. Evaluating uncertainty is critical in high-stakes applications such as wildfire. If we do not account for the uncertainty, there may be a high risk that the derived decisions will not perform well in practice. In this chapter, we estimate CIs for the performance estimates to further study the uncertainties of the MM and MMR estimator.

We first describe our simulation experiment in Section 4.1. We use four method for estimating CIs. We did CI estimations based on two estimators: MM and MMR. We analyze the coverages of the CIs for different methods and estimators. The results in Section 4.2 shows the CI coverages for the estimation are dependent on the variance for the estimator.

4.1 Experimental Design

Our simulation experiment involves multiple iterations. Each iteration uses a different seed value and in each iteration sampled observations are simulated for each combination of logging policy and critical level distribution. Then, we obtain a set of observations (action - outcome pair) for each combination of logging policy and critical level distribution based on the actions and critical values. If the action is greater or equal to the critical value, the outcome is 1, otherwise the outcome is 0. We use the simulated observations from each combination of logging policy and critical level distribution to estimate the performance of a target policy. We use the eleven logging policies and eleven target policies, shown in Figure 3.1, as well as ten H1 critical level distributions shown in Figure 3.6. In total, there are 1210 scenarios. We use the $f(a,1)$ and $f(a,0)$ vectors of the estimator for each combination of logging and target policy, and we use it to calculate an estimate for the target policy value.

We compute CIs for the value of the target policy based on these sample observations. We use four different methods for estimating CIs. We have sampled observations of size n. Then we compute the estimator for each sample and then take the mean of the estimated target values from the sampled observations. The formula for getting confidence bounds for central limit theorem (CLT) method is:

$$\bar{x} \pm \mathbf{z}^* s,$$

where the sample mean is $\bar{x}$ with sample standard deviation s. Here $\mathbf{z}^*$ is the upper 0.05/2 critical value for the standard normal distribution. The interval bounds for student-t distribution (T method) is calculated by:

$$\bar{x} \pm t\sqrt{\frac{s^2}{n}},$$

where t is the upper 0.05/2 critical value of t-distribution with $n-1$ degrees of freedom.

Algorithm 1 Pseudo algorithm for getting r_0

```
initialize: $n$ (number of observations), $\alpha = 0.05$
if $method == F$ then
    $Q \leftarrow$Percent point function of F$(1-\alpha, df_1 = 1, df_2 = n-1)$
end
if $method == Chi$ then
    $Q \leftarrow$Percent point function of chi-square$(1-\alpha, df = 1)$
end
$r_0$ = exp(-$Q$/2)
```

For Chi and F methods, we use Owen's Empirical Likelihood(EL) confidence region for the expectation of target policy value of the Chi and F distribution with 95% confidence level [12]. EL is a non-parametric maximum likelihood approach that does not assume any parametric distribution for the estimate. $r_0 \in \mathbb{R}$ is an important parameter to get CI for Chi and F method. For Chi and F methods, we get the lower bound by solving the following problem:

$$
\begin{aligned}
& \min_{w \in \mathbb{R}^n} V^\intercal w && (4.1)\\
\text{s.t. } & \sum_{i=1}^{n} \log(n w_i) \geq \log(r_0), && (4.2)\\
& w \geq 0, && (4.3)\\
& \sum_{i=1}^{n} w_i = 1. && (4.4)
\end{aligned}
$$

The upper bound is computed by maximizing (4.1), where V is the target policy value vector with length n, w is an n-dimensional vector variable, and r_0 was calculated by Algorithm 1. After solving this problem, we take the objective value and make it the lower or upper bound.

For each combination of logging policy, critical level distribution, and target policy, we repeat the process of generating data and computing CIs M times, where M is the number of iterations. In each iteration we generate a simulated data set of size n using the logging policy and the critical level distribution. We compute the four types of CIs. We also compute the true value of the target policy and then see if the CIs contain the true value. Then we calculate the average coverage over M iteration for the 1210 (11 logging policy, 11 target policy and 10 critical level distribution) scenarios.

4.2 Experimental Results

4.2.1 CI with 30 Observations (MM)

We obtain $n = 30$ observations for each logging policy, target policy and critical level distribution. Then we calculate CIs by the four methods. For each of the 1210

CI Analysis for 30 Observations(MM)		
Estimation Method	Minimum Coverage	Average Coverage
CLT	0.64	0.91
T	0.64	0.92
Chi	0.61	0.92
F	0.62	0.93

Table 4.1: CI Analysis for 30 Observations, with 250 Iterations

scenarios, we repeat it $M = 250$ times. We evaluate the four confidence intervals on simulated data by observing how many CIs contain the true value. The coverage for a single scenario in one iteration is either 0 or 1. Table 4.1 shows the results. The results show the average coverage over the 1210 scenarios is around 92%, with a minimum coverage around 60%. Chi-square and F methods have a slightly higher average coverage compared to CLT and T. The coverage for a single scenario in one iteration is either 0 or 1 and The coverage of each scenario is the mean coverage over 250 iterations (250 binary numbers). Most of the small coverage scenarios are distributed at L1-5, T1-5 and the H1(10) critical level distribution.

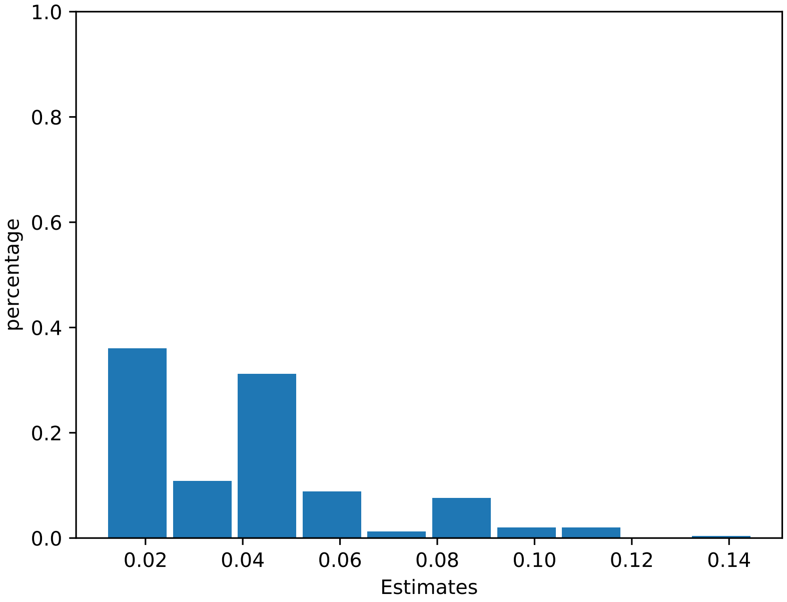


Figure 4.1: Distribution of the estimate for scenario (low coverage with 30 Observation 250 iteration)

We want to further analyze the low coverage scenarios to find out what the cause is. We plotted a histogram (Figure 4.1) for one of the lowest coverage scenarios (L1, T1, H1(10)). The plot shows the distribution of the estimated values for 400 iterations of the MM estimator. Then, we did D'Agostino and Pearson's normality test. The null hypothesis for the normality test is: the distribution of those 400 data points come from a normal distribution. The test shows that the null hypothesis of normality is rejected which indicates the distribution is not normal. In the next section, we run the same simulation using more observations and more iterations.

4.2.2 CI with 100 Observations (MM)

CI Analysis for 100 Observations(MM)		
Estimation Method	Minimum Coverage	Average Coverage
CLT	0.80	0.93
T	0.80	0.94
Chi	0.81	0.94
F	0.82	0.94

Table 4.2: CI Analysis for 100 Observations, with 400 Iterations

In this section we use the same simulation procedure as in Section 4.2.1. We increase the number of observations in each iteration to 100 and increase the number of iterations to 400. The logging policies, target policies and critical level distributions stay the same. The overall average coverage for CLT and T distributions are around 93.5% and for Chi and F distributions are around 94%. The minimum coverage of all four distributions are now around 80%. Compared to the previous experiment, the minimum coverage is increased when we increase the number of observations. The histogram of the scenario "L7, T1, H1(9)" which has the lowest coverage is shown in Figure 4.2. We did the normality test on the distributions for 9 out of the 1210 scenarios which had low coverage. Then we obtained the result that the null hypothesis cannot be rejected. On the other hand, we also tested normality for the

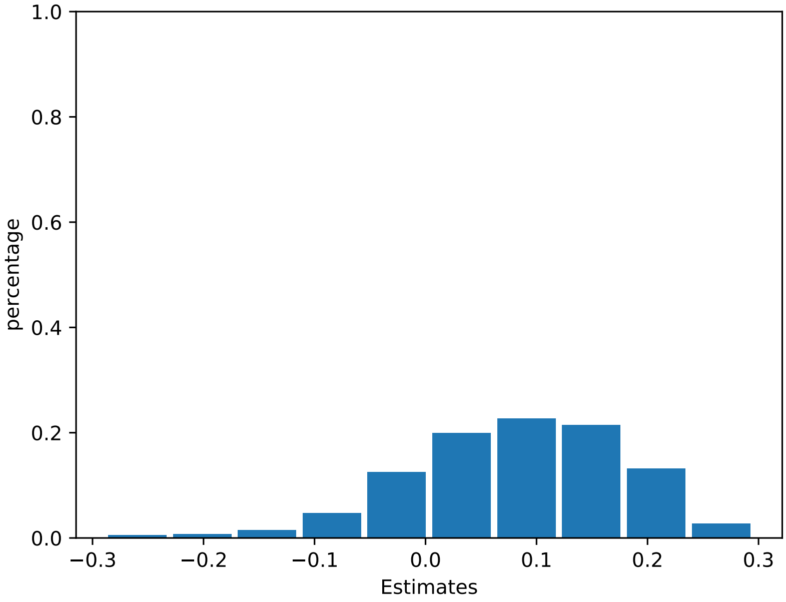


Figure 4.2: Distribution of the estimate for scenario (low coverage with 100 Observation 400 iteration)

scenarios with high coverage. The histogram for one of the highest coverage scenarios

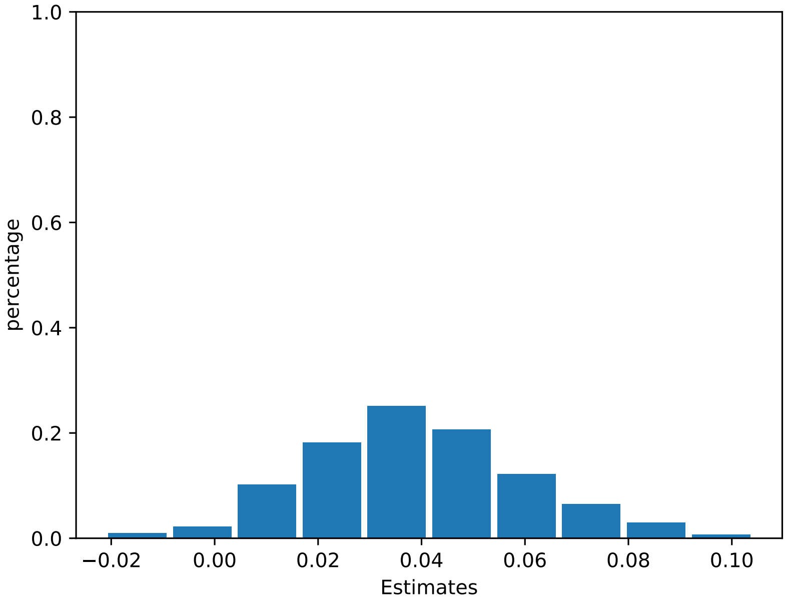


Figure 4.3: Distribution of the estimate for scenario (high coverage with 100 Observation 400 iteration)

"L6, T5, H1(10)" is shown in Figure 4.3. The normality test results for the high coverage scenarios also show a failure to reject the null hypothesis.

For each scenario, we run multiple iterations. So we have an empirical distribution of the estimated values. Then we computed the empirical variance and compared it to the true variance computed by the formula (2.3). The scatter plot (Figure 4.4) of variance by formula and variance by the simulation shows a strong positive correlation, which is expected. Each point in the scatter plot is one of the 1210 scenarios.

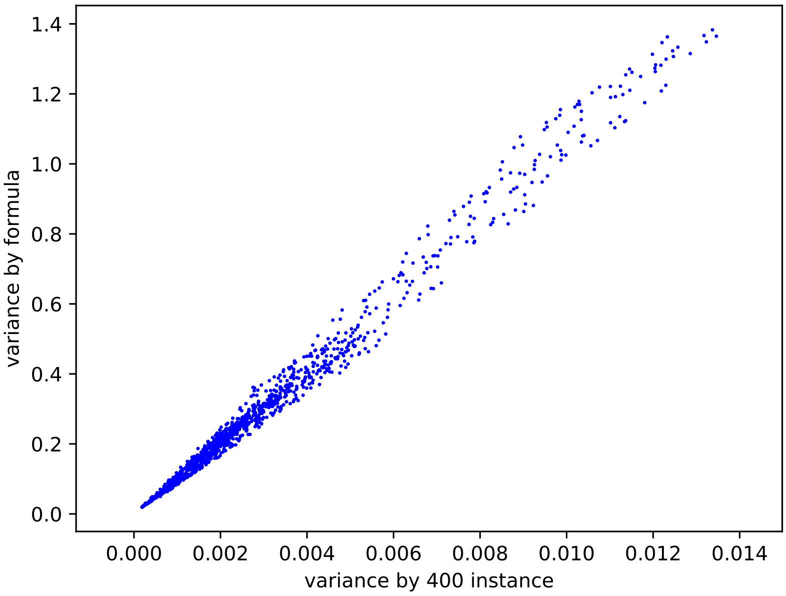
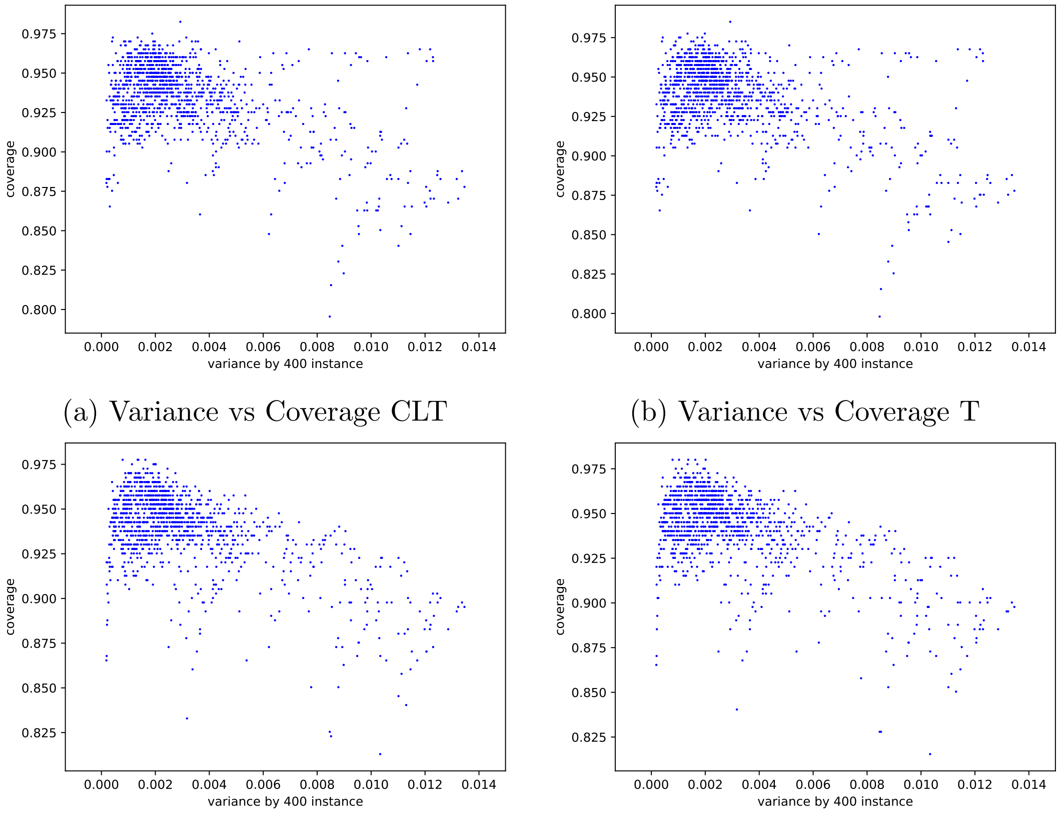


Figure 4.4: Simulation Variance vs Variance by Formula

After that, we tried to find out the relationship between variance and the coverage. We generated scatter plots showing the empirical variance against coverage for all four methods in Figure 4.5. As can be seen from the scatter plots, coverage and variance have a negative Pearson's correlation. For CLT (Figure 4.5a) and T (Figure 4.5b) methods, the correlation coefficient is about -0.48. When variance is below 0.5, the correlation is weak which is around -0.01; when variance is above 0.5, the correlation is stronger, around -0.49. For Chi (Figure 4.5c) and F (Figure 4.5d) methods, the correlation coefficient is about -0.60. When variance is below 0.5, the correlation is weak which is around -0.10; when variance is above 0.5, the correlation is stronger, around -0.62.



(a) Variance vs Coverage CLT

(b) Variance vs Coverage T

(c) Variance vs Coverage Chi

(d) Variance vs Coverage F

Figure 4.5: Variance against Coverage for different CIs.

4.2.3 CI with 100 Observations (MMR)

CI Analysis for 100 Observations(MMR)		
Estimation Method	Minimum Coverage	Average Coverage
CLT	0.76	0.94
T	0.77	0.94
Chi	0.78	0.94
F	0.78	0.94

Table 4.3: CI Analysis for 100 Observations, with 400 Iterations

In this section we analyze the uncertainty of the MMR estimator. We ran the same simulation using 100 observations and 400 iterations as in the previous section. Table 4.3 shows the minimum coverage and average coverage over the 1210 scenarios. Compared to the CIs of MM estimator, that of MMR estimator has higher average

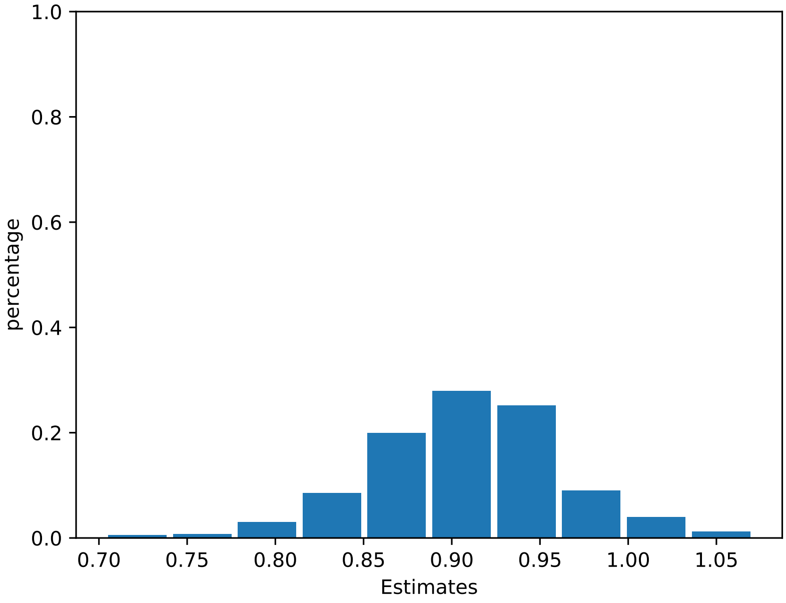


Figure 4.6: Distribution of the estimate for scenario (low coverage with 100 Observation 400 iteration)

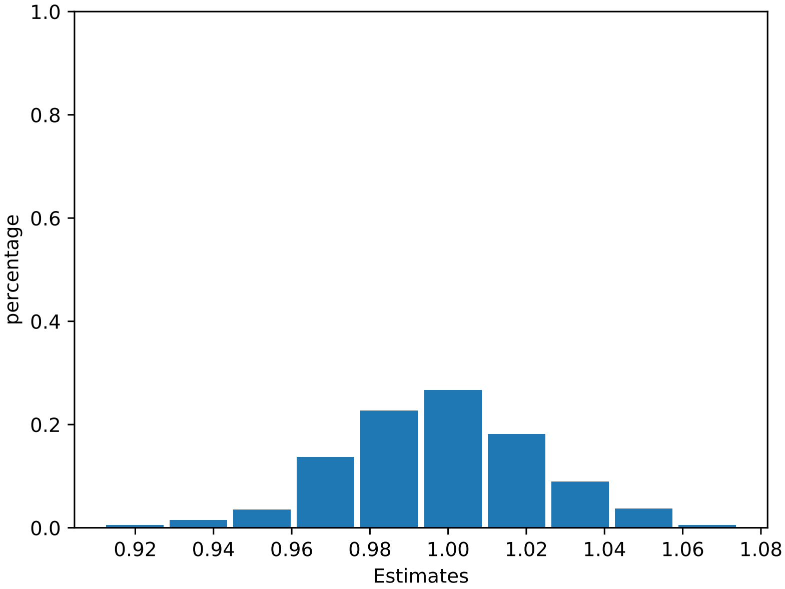


Figure 4.7: Distribution of the estimate for scenario (high coverage with 100 Observation 400 iteration)

coverages and lower minimum coverages. The histograms of a low coverage scenario “L6, T1, H1(2)” and a high coverage scenario “L4, T6, H1(1)” are shown in Figures 4.6 and 4.7. The normality test shows that they both fail to reject the null hypothesis.

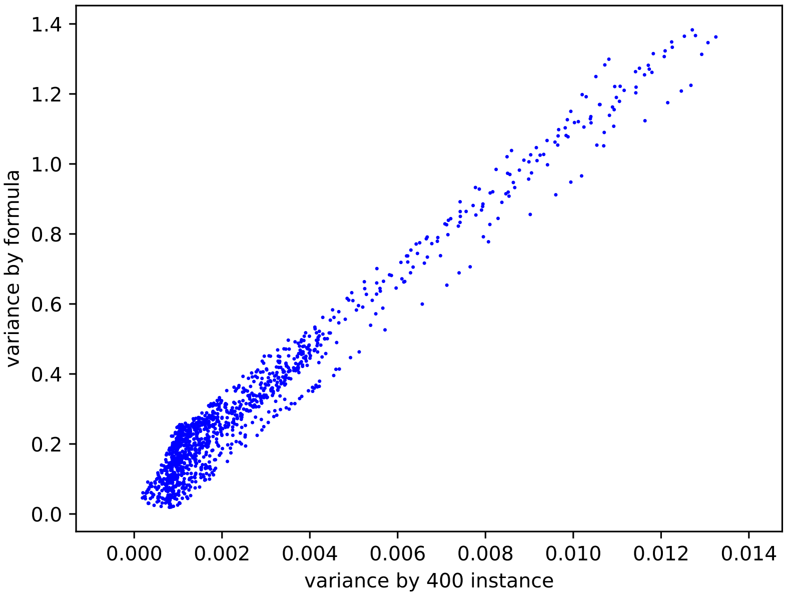
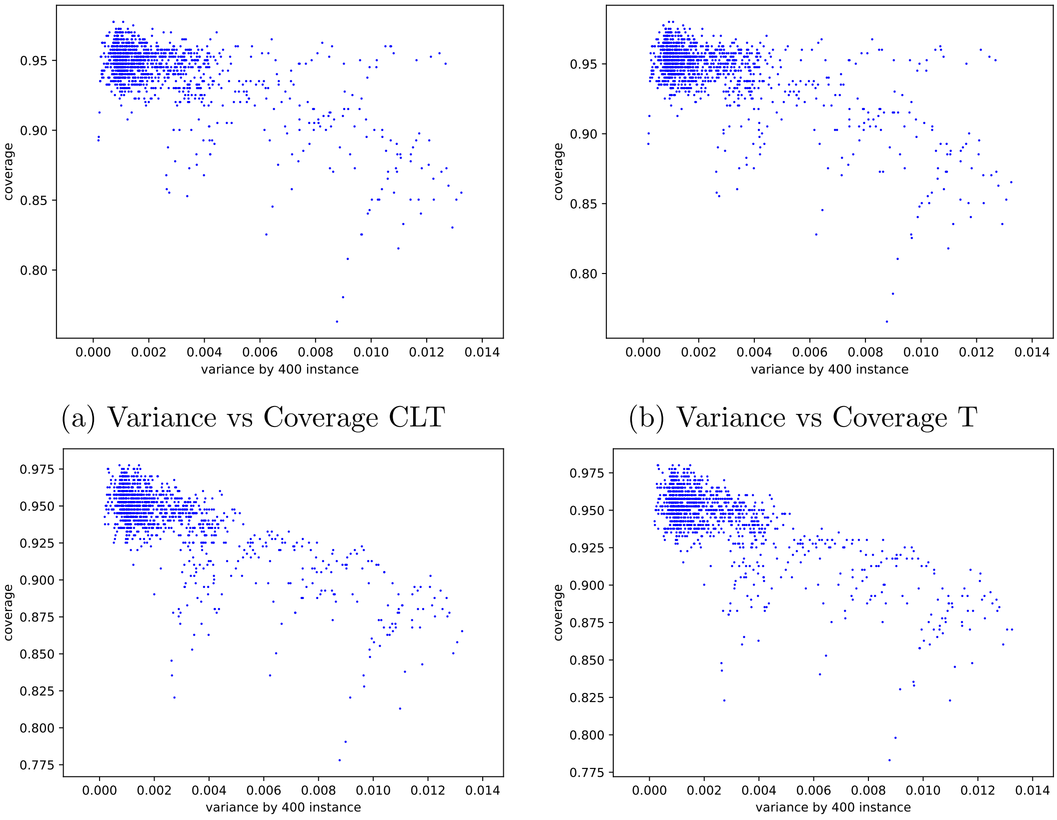


Figure 4.8: Simulation Variance vs Variance by Formula (MMR)



(a) Variance vs Coverage CLT

(b) Variance vs Coverage T

(c) Variance vs Coverage Chi

(d) Variance vs Coverage F

Figure 4.9: Variance against Coverage for different CIs.

Figure 4.8 shows the scatter plot between the empirical variance and the variance by formula (2.3). It shows a strong positive correlation with the correlation coefficient 0.98. Then, we generated scatter plots showing the relationship between variance and coverage. Figure 4.9 shows four scatter plots of empirical variance against coverage. The negative correlations are stronger compared to the results of the MM estimator. The correlation for CLT (Figure 4.9a) and T method (Figure 4.9b) are around -0.64 and the correlation for Chi (Figure 4.9c) and F (Figure 4.9d) are around -0.75.

4.3 Conclusions

In this section, we computed the CIs for the MM estimator and the MMR estimator using four methods. The average coverages are above 93% when we have enough samples for all of the four methods and the two estimators we used; and the minimum coverages are above 76%. The CIs for the MMR estimator has higher average coverage as well as lower minimum coverage compared to the MM estimator. The result also shows there is a negative correlation between variance and coverage. The MMR estimator has lower average variance compated to the MM estimator (Section 4.2) which explains why the MMR estimator has higher average coverage. There is a stronger negative correlation between variance and coverage when the variance is high.

We have used 4 methods for CIs: CLT, T, Chi and F. The F and Chi methods performed better which has the higher average coverage for both of the MM and MMR estimators when we have 100 sample observations. The F and Chi methods are EL approaches and EL is a non-parametric method that does not assume any distribution for the data. It can therefore perform better than parametric methods when the distribution of the data is different from an assumed parametric distribution. The other two methods have nearly the same performance in terms of the average coverage. The normality test failed to reject the null hypothesis when we had 100 observations. The parametric estimators CLT and T of CIs that assume a normal distribution, therefore, also perform well.

Chapter 5

Conclusions, & Future Work

5.1 Conclusions

We have studied batch policy evaluation in the presence of monotone responses, which is the first time in literature to the best of our knowledge. In summary, we have found multiple unbiased estimators for off-policy evaluation. The estimators are MM which minimizes the worst-case variance, IS which is our baseline estimator using the Important Sampling technique, RIS which minimizing the worst-case regret against IS, and MMR which minimizing the worst regret against V^*, the best variance for each critical level distribution. We have used both point estimation and CI to evaluate the estimators. The results have shown that the MMR, MM and RIS estimators have lower variance compared to the IS estimator. MM estimator guarantees the smallest worst-case variance and almost as good as MMR Estimator. MMR estimator gets the smallest overall variance. Although we have not applied to the real historical data yet, the estimators we found could be useful for similar applications. This led us to some future work.

5.2 Future Work

Our current study is limited only to the four estimators. There are some future works we can work on. On the one hand, we can further optimize the estimator. There are infinitely many unbiased estimators. We can pursue to find more of them and determine the best one using different criteria. We can also analyze CIs for other estimators. In addition, we can try to combine some of the estimators to create a more general estimator that performs well in most of the scenarios. On the other hand, we can apply the theory and estimators we developed to real fire data instead of using simulated observations. Besides, making an application for dispatch prediction would be the final goal of the wildfire project. In the practice of wildfire operations, we have multiple resources. We can also consider multi-dimensional action space.

Bibliography

[1] "Fire statistics," *CAL FIRE*, 2021.

[2] L. Ntaimo, J. Arrubla, C. Stripling, J Young, and T. Spencer, "A stochastic programming standard response model for wildfire initial attack planning.," *Canadian Journal of Forest Research. 42. 987-1001. 10.1139/x2012-032.*, 2012.

[3] P Kourtz, "Two dynamic programming algorithms for forest fire resource dispatching," *Canadian Journal of Forest Research*, vol. 19, no. 1, pp. 106–112, 10.1139/x89–014, 1989.

[4] M. R. Wiitala, *A Dynamic Programming Approach to Determining Optimal Forest Wildfire Initial Attack Responses*, ser. Symposium on Fire Economics, Planning, and Policy: Bottom Lines. U.S. Department of Agriculture Forest Service, Pacific Southwest Research Station, 1999.

[5] M. Kato and Y Kaneko, "Off-policy evaluation of bandit algorithm from dependent samples under batch update policy," *Papers 2010.13554, arXiv.org.*, 2020.

[6] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning.," *Proceedings of The 33rd International Conference on Machine Learning, in PMLR 48:2139-2148*, 2016.

[7] P. Thomas, G. Theocharous, and M. Ghavamzadeh, "High-confidence off-policy evaluation.," *Proceedings of the AAAI Conference on Artificial Intelligence, 29(1).*, 2015.

[8] R. Srinivasan, "Importance sampling—applications in communications and detection," Jan. 2002.

[9] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[10] K Miettinen, *Nonlinear multiobjective optimization.* Springer, 2012, vol. 12.

[11] B. Fredrickson, *Positivity.* Three Rivers Press, 2009, ISBN: 9780307393746.

[12] A Owen, *Empirical Likelihood Ratio Confidence Regions.* Ann. Statist., 1990, vol. 18.

Appendix A: Proposition and Theorem Proof

A.1 Proposition 1 Proof

Proof. We prove the proposition by converting the inner problem to its dual. The Lagrangian of the inner problem is:

$$\begin{aligned} L^{\text{inner}}(p,\lambda,\mu) &= (\pi_t^\intercal Tp)^2 - \pi_l^\intercal F^{(2)}p - \lambda^\intercal p + \mu(e^\intercal p - 1) \\ &= (\pi_t^\intercal Tp)^2 + (\mu e^\intercal - \lambda^\intercal - \pi_l^\intercal F^{(2)})p - \mu, \end{aligned}$$

where $\lambda \in \mathbb{R}^k$ and $\mu \in \mathbb{R}$. The KKT condition is

$$\begin{aligned} \nabla_p L^{\text{inner}}(p,\lambda,\mu) &= 2(T^\intercal \pi_t \pi_t^\intercal T)p - (F^{(2)})^\intercal \pi_l - \lambda + \mu e = 0, \\ e^\intercal p &= 1, p \geq 0, \\ \lambda &\geq 0 \\ \lambda_i p_i &= 0 \text{ for } i \in [k]. \end{aligned}$$

Since the inner problem (2.7) is convex and has only linear constraints, the KKT condition is necessary and sufficient for an optimal solution [9]. Also, (2.7) and its Lagrangian dual satisfy strong duality.

Now we derive the Lagrangian dual. Let $g^{\text{inner}}(\lambda,\mu) = \inf_p L^{\text{inner}}(p,\lambda,\mu)$. The minimizer p in the infimum must satisfy the first equation of the KKT condition. Note that in the first equation, the first term $(T^\intercal \pi_t \pi_t^\intercal T)p$ is a constant multiple of $T^\intercal \pi_t$. Thus, only when $(F^{(2)})^\intercal \pi_l + \lambda - \mu e$ is a constant multiple of $T^\intercal \pi_t$, the equation has a solution, and thus, $g(\lambda,\mu)$ is well defined. If $(F^{(2)})^\intercal \pi_l + \lambda - \mu e = 2\alpha T^\intercal \pi_t$ for some $\alpha \in \mathbb{R}$, then the first equation of the KKT condition reduces to $2\pi_t^\intercal Tp - 2\alpha = 0$, or $\pi_t^\intercal Tp = \alpha$. Thus, for a minimizer p of $L^{\text{inner}}(p,\lambda,\mu)$, we have

$$L^{\text{inner}}(p,\lambda,\mu) = (\pi_t^\intercal Tp)^2 + (\mu e^\intercal - \lambda^\intercal - \pi_l^\intercal F^{(2)})p - \mu = -\alpha^2 - \mu.$$

Recall that the Lagrangian dual of (2.7) is defined as $\max_{\lambda \geq 0, \mu} g^{\text{inner}}(\lambda,\mu)$. By a change of variable $\lambda = 2\alpha T^\intercal \pi_t - (F^{(2)})^\intercal \pi_l + \mu e$, the Lagrangian dual can be written as

$$\min \alpha^2 + \mu \text{ s.t. } 2\alpha T^\intercal \pi_t + \mu e \geq (F^{(2)})^\intercal \pi_l \,.$$

Therefore, (2.5) can be written as a minimization problem:

$$\begin{aligned} \text{(P)} \ \min_{f,\alpha,\mu} \ & \alpha^2 + \mu \\ \text{s.t. } & \mu e^\intercal \geq \pi_l^\intercal F^{(2)} - 2\alpha \pi_t^\intercal T \\ & \pi_l^\intercal F = \pi_t^\intercal T, \end{aligned}$$

■

A.2 Theorem 3 Proof

Proof. Let $z^* \in \mathcal{V}$ be a minimizer of s. Let us suppose that it is not Pareto optimal. In this case, there exists a $z \in \mathcal{V}$ such that $z \leq z^*$ and also for some $p \in D$, $z(p) < z^*(p)$. Because s is strongly increasing, we know that $s(z) < s(z^*)$, which contradicts the assumption that z^* minimizes s. Thus z^* is Pareto optimal. ■

A.3 Proposition 4 Proof

Proof. Suppose that (2.26) does not have an optimal solution that is a vertex. Let p^* be an optimal solution. Since Δ^{k-1} is convex, $p^* = \sum_{a=1}^{k} \lambda_a p^a$, where $p^1, \ldots, p^k$ are vertices of the unit simplex and $\lambda_1, \ldots, \lambda_k$ are nonnegative weights that sum up to one. Because the objective function is convex, we have $||Cp^* + d||_2^2 \leq \sum_{a=1}^{k} \lambda_a ||Cp^a + d||_2^2$. However, this contradicts the assumption that $||Cp^* + d||_2^2 > ||Cp^a + d||_2^2$ for $a \in [k]$. ■