Stochastic Methods in Optimization and Machine Learning

Fengpei Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Stochastic Methods in Optimization and Machine Learning

Fengpei Li

Stochastic methods are indispensable to the modeling, analysis and design of complex systems involving randomness. In this thesis, we show how simulation techniques and simulation-based computational methods can be applied to a wide spectrum of applied domains including engineering, optimization and machine learning. Moreover, we show how analytical tools in statistics and computer science including empirical processes, probably approximately correct learning, and hypothesis testing can be used in these contexts to provide new theoretical results. In particular, we apply these techniques and present how our results can create new methodologies or improve upon existing state-of-the-art in three areas: decision making under uncertainty (chance-constrained programming, stochastic programming), machine learning (covariate shift, reinforcement learning) and estimation problems arising from optimization (gradient estimate of composite functions) or stochastic systems (solution of stochastic PDE).

The work in the above three areas will be organized into six chapters, where each area contains two chapters. In Chapter 2, we study how to obtain feasible solutions for chance-constrained programming using data-driven, sampling-based scenario optimization (SO) approach. When the data size is insufficient to statistically support a desired level of feasibility guarantee, we explore how to leverage parametric information, distributionally robust optimization and Monte Carlo simulation to obtain a feasible solution of chance-constrained programming in small-sample situations.

In Chapter 3, We investigate the feasibility of sample average approximation (SAA) for general stochastic optimization problems, including two-stage stochastic programming without the relatively complete recourse. We utilize results from the *Vapnik-Chervonenkis* (VC) dimension and *Probably Approximately Correct* learning to provide a general framework. In Chapter 4, we design a robust importance re-weighting method for estimation/learning problem in the *covariate shift* setting that improves the best-know rate. In Chapter 5, we develop a model-free reinforcement learning approach to solve constrained Markov decision processes (MDP). We propose a two-stage procedure that generates policies with simultaneous guarantees on near-optimality and feasibility. In Chapter 6, we use multilevel Monte Carlo to construct unbiased estimators for expectations of random parabolic PDE. We obtain estimators with finite variance and finite expected computational cost, but bypassing the curse of dimensionality. In Chapter 7, we introduce unbiased gradient simulation algorithms for solving stochastic composition optimization (SCO) problems. We show that the unbiased gradients generated by our algorithms have finite variance and finite expected computational cost.

Table of Contents

Acknowledgements

I want to thank my advisor Henry Lam and co-advisor Jose Blanchet for their guidance, patience and inspiration. It has been a wonderful journey and I feel extremely fortunate to have worked with them.

I also thank my collaborators Garud Iyengar, Donald Goldfarb, Chaoxu Zhou, Xiaoou Li, Haoxian Chen and Siddharth Prusty for the opportunities to accomplish exciting work together.

I want to thank Jing Dong, Adam Elmachtoub and Garud Iyengar for serving on my committee.

I want to thank many IEOR students before me and after me, for their friendship. I especially want to thank the PhD students in our cohort, for their encouragement and for the great times we had. Some of you have become my closest friends and I wish all of you the best.

I also want to give a special thanks to a few people I've known since high school, who help me grow as a person and a friend.

Finally, I thank my parents, Wei Li, Shu Li and my grandparents Jingxiu Li, Shijie Jiang. My deepest gratitude to all of you.

To my parents.

Chapter 1: Introduction

Stochastic methods are indispensable to the modeling, analysis and design of complex systems involving randomness. In this thesis, we show how simulation techniques and simulation-based computational methods can be applied to a wide spectrum of applied domains including engineering, optimization and machine learning. Moreover, we show how analytical tools in statistics and computer science including empirical processes, probably approximately correct learning, and hypothesis testing can be used in these contexts to provide new theoretical results. In particular, we apply these techniques and present how our results can create new methodologies or improve upon existing state-of-the-art in three areas: decision making under uncertainty (chance-constrained programming in Chapter 2,stochastic programming in Chapter 3), machine learning (covariate shift in Chapter 4, reinforcement learning in Chapter 5) and unbiased estimation arising from optimization (gradient estimate of composite functions in Chapter 7) or stochastic systems (solution of stochastic PDE in Chapter 6). Most of the materials in this thesis are published or submitted works contained in [1, 2, 3, 4, 5, 6, 7].

The work in the above three areas will be organized into six chapters, where each area contains two chapters. Chapter 2 and 3 are topics on decision making under uncertainty. Chapter 2 is on how to solve the chance-constrained problem using scenario generation approach but with only limited data. We investigated a systematic approach to use simulated Monte Carlo samples in lieu of real data, under a parametric distribution, and maintain a rigorous certificate of feasibility just as solutions obtained from real data. Our approach makes use of a distributionally robust optimization (DRO) formulation that translates the data size requirement into a Monte Carlo sample size requirement drawn from what we call a generating distribution. We show that, while the optimal choice of this generating distribution is the one eliciting the data or the baseline distribution in a nonparametric divergence-based DRO, it is not necessarily so in the parametric case. Cor-

respondingly, we develop procedures to obtain generating distributions that improve upon these basic choices. Chapter 3 investigates the feasibility of sample average approximation (SAA) for general stochastic optimization problems, including two-stage stochastic programming without the relatively complete recourse assumption. In this chapter, we introduce a new framework based on the Vapnik-Chervonenkis (VC) dimension and Probably Approximately Correct learning to study the feasibility of SAA solutions which includes, but is not limited to two-stage stochastic programming. Following [8, 9], we focus on showing the exponential decrease of the portion of infeasible solutions as sample size grows. As a key contribution, we show how our framework produces feasibility bounds that are both general and explicit. In particular, for solutions of SAA, we provide feasibility bounds *with explicit and computable* constants, *with no requirement on the geometric or distributional properties of* (3.1) and *with no specific regularity conditions on the objective function* (i.e., Lipschitz continuity or the existence of certain moment generating function as in [9, 8]). Moreover, the analysis itself also does not hinge on the specific type of the problem (i.e., not limited to two-stage stochastic programming) and is widely applicable in both scenarios where some of the best-known results on SAA feasibility have been presented, and other scenarios where no similar results have been established.

Chapter 4 and 5 show how applied probability techniques can be used on topics in machine learning. Chapter 4 address how to design robust version of importance sampling weight under the context of Kernel Mean Matching (KMM) and covariate shift. In many learning problems, the training and testing data follow different distributions and a particularly common situation is the *covariate shift*. To correct for sampling biases, most approaches, including the popular kernel mean matching (KMM), focus on estimating the importance weights between the two distributions. Reweighting-based methods, however, are exposed to high variance when the distributional discrepancy is large and the weights are poorly estimated. On the other hand, the alternate approach of using nonparametric regression (NR) incurs high bias when the training size is limited. In this chapter, we propose and analyze a new estimator that systematically integrates the residuals of NR with KMM reweighting, based on a control-variate perspective. The proposed estimator can

be shown to either strictly outperform or match the best-known existing rates for both KMM and NR, and thus is a robust combination of both estimators. Chapter 5 explores how efficient simulation can speed up the finding of optimal policy for reinforcement learning problem. We develop a model-free reinforcement learning approach to solve constrained Markov decision processes, where the objective and budget constraints are in the form of infinite-horizon discounted expectations, and the rewards and costs are learned sequentially from data. We propose a two-stage procedure where we first search over deterministic policies, followed by an aggregation with a mixture parameter search, that generates policies with simultaneous guarantees on near-optimality and feasibility. We also numerically illustrate our approach by applying it to an online advertising problem. Applications of Reinforcement Learning (RL) in online advertising with recommendation systems have been a topic of major research interests ([10, 11, 12]). However, despite their tremendous success, most RL-methods are not designed to learn optimal policies under constraints, yet they appear ubiquitously when facing budget or safety considerations. A standard framework for studying RL under constraints is the Constrained Markov Decision Process (CMDP), where the objective is to maximize the long-run return, with constraints on one or several types of long-run costs. In this chapter, we consider the case where both the objective and the constraint are in the form of an infinite-horizon cumulative discounted expectation, whereas the returns, costs and transitions are revealed from sequential data. The goal is to design an efficient methodology for the constrained problem by assimilating classical optimality properties of CMDP into RL, in order to efficiently use established RL approaches and obtain policies that enjoy both near-optimality and feasibility.

Chapter 6 and 7 are estimation problems. They show how to construct unbiased estimator from biased estimators, with finite variance and computational cost, under the context of gradient estimate of composite optimization problem, as well as solutions of random partial differential equations. Partial differential equations (PDEs) are important tools for modeling physical or financial systems. However, intrinsic variability of the system or measurement errors bring uncertainty into the model and are commonly represented by random input data. In Chapter 6, we use multilevel

Monte Carlo to construct unbiased estimators for expectations of random parabolic PDE. Building on previous works of Giles (2008) and Li et al.(2018), we obtain estimators with finite variance and finite expected computational cost, but bypassing the curse of dimensionality. Regarding error analysis in the random PDE, we combine rough path theory with numerical stochastic analysis in a novel way. In Chapter 7, We introduce unbiased gradient simulation algorithms for solving stochastic composition optimization (SCO) problems. We show that the unbiased gradients generated by our algorithms have finite variance and finite expected computational cost. Therefore, the unbiased gradients can be directly used to solve SCO problems by applying the Stochastic Gradient Descent method (SGD). We also show how to combine unbiased gradient simulation with variance reduction techniques such as stochastic variance reduced gradient (SVRG) or stochastically controlled stochastic gradient (SCSG) to achieve state-of-the-art theoretical convergence rates as well as practical performances. Finally, we illustrate the effectiveness of our algorithms through experiments on datasets arising from statistics and machine learning, specifically, Cox's partial likelihood model and conditional random field models.

Chapter 2: Parametric Scenario Optimization under Limited Data: A Distributionally Robust Optimization View

We consider optimization problems with uncertain constraints that need to be satisfied probabilistically. When sufficient data are available, a common method to obtain feasible solutions for such problems is to impose sampled constraints, following the so-called scenario optimization approach. However, when the data size is small, the sampled constraints may not statistically support a feasibility guarantee on the obtained solution. This chapter studies how to leverage parametric information and the power of Monte Carlo simulation to obtain feasible solutions for small-data situations. Our approach makes use of a distributionally robust optimization (DRO) formulation that translates the data size requirement into a Monte Carlo sample size requirement drawn from what we call a generating distribution. We show that, while the optimal choice of this generating distribution is the one eliciting the data or the baseline distribution in a nonparametric divergence-based DRO, it is not necessarily so in the parametric case. Correspondingly, we develop procedures to obtain generating distributions that improve upon these basic choices. We support our findings with several numerical examples.

It is also worth noting that there are other possible ways to approach this problem. For example, the requirement of uncertainty set of DRO to include the true distribution is usually considered to restrictive. However, to establish a theorem that would work in any black-box situation where the feasible set $\mathcal{X}_\xi$ is not specified or in most general form, we use this restrictive assumption to avoid case-by-case analysis. Also, the problem can also be efficiently solved by empirical process/learning theory concepts such as VC-dimension or Radamacher complexity, if we assume these types of structure on $\mathcal{X}_\xi$, but we again consider the most general form here. So the set-up here is most appropriate for the setting where insufficient data is still considered adequate to

characterize the parametric distribution. The main difficulty, when considering a black-box version of $\mathcal{X}_\xi$, is to transfer the feasibility under a sampling distribution, to the (unknown) true distribution.

2.1 Introduction

We consider optimization problems in the form

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x, \\ \text{s.t.} \quad & \mathbb{P}(x \in \mathcal{X}_\xi) \geq 1 - \epsilon, \end{aligned} \tag{2.1}$$

where $\mathbb{P}$ is a probability measure governing the random variable ξ (independent of decision variable x) on some space $\mathcal{Y}$ and $\mathcal{X}_\xi \subseteq \mathcal{X} \subseteq \mathbb{R}^d$ is a set depending on ξ. Problem (2.1) enforces a solution x to satisfy $x \in \mathcal{X}_\xi$ with high probability, namely at least $1 - \epsilon$. This problem is often known as a probabilistically constrained or chance-constrained program (CCP) [13]. It provides a natural framework for decision-making under stochastic resource capacity or risk tolerance, and has been applied in various domains such as production planning [14], inventory management [15], reservoir design [16, 17], communications [18], and ranking and selection [19].

We focus on the situations where $\mathbb{P}$ is unknown, but some i.i.d. data, say $\xi_1, \ldots, \xi_n$, are available. One common approach to handle (2.1) in these situations is to use the so-called scenario optimization (SO) or constraint sampling [20, 21]. This replaces the unknown constraint in (2.1) with $x \in \mathcal{X}_{\xi_i}, i = 1, \ldots, n$, namely, by considering

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x, \\ \text{s.t.} \quad & x \in \mathcal{X}_{\xi_i},\ i = 1, \ldots, n. \end{aligned} \tag{2.2}$$

Note that CCP (2.1) is generally difficult to solve even when the set $\mathcal{X}_\xi$ is convex for any given ξ and the distribution $\mathbb{P}$ is known [13]. Thus, the sampled problem (2.2) offers a tractable approximation for the difficult CCP even in non-data-driven situations, assuming the capability to generate these samples.

Our goal is to find a good feasible solution for (2.1) by solving (2.2) under the availability of i.i.d. data described above. Intuitively, as the sample size n increases, the number of constraints in (2.2) increases and one expects them to sufficiently populate the safety set $\{\xi : x \in \mathcal{X}_\xi\}$, thus ultimately give rise to a feasible solution for (2.1). To make this more precise, we first mention that because of the statistical noise from the data, one must settle for finding a solution that is feasible with a high confidence. More specifically, define, for any given solution x,

$$V(x, \mathbb{P}) = \mathbb{P}(x \notin \mathcal{X}_\xi)$$

to be the violation probability of x under probability measure $\mathbb{P}$ that generates ξ. Obviously, x is feasible for (2.1) if and only if

$$V(x, \mathbb{P}) \leq \epsilon. \tag{2.3}$$

We would like to obtain a solution, say $\hat{x}$, from the data such that

$$\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) \leq \epsilon) \geq 1 - \alpha, \tag{2.4}$$

where $\mathbb{P}_{data}$ is the distribution that generates the i.i.d. data $\xi_i, i = 1, \ldots, n$ (each sampled from $\mathbb{P}$), and $1-\alpha$ is a given confidence level (e.g., $\alpha = 5\%$). In other words, we want $\hat{x}$ to satisfy the chance constraint in (2.1) with the prescribed confidence rigorously. On the other hand, the optimality $\hat{x}$ is mostly studied empirically and we do not discuss it in detail here.

Under the convexity of $\mathcal{X}_\xi$ and mild additional assumptions (namely, that every instance of (2.2) has a feasible region with nonempty interior and a unique optimal solution), the seminal work [22] provides a tight estimate on the required data size n to guarantee (2.4). They show that a solution $\hat{x}$ obtained by solving (2.2) satisfies

$$\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) > \epsilon) \leq \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1-\epsilon)^{n-i}, \tag{2.5}$$

with equality held for the class of “fully-supported" optimization problems [22]. Thus, suppose

we have a sample size n large enough such that

$$B(\epsilon, d, n) = \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1-\epsilon)^{n-i} \leq \alpha, \tag{2.6}$$

then from (2.5) we have $\mathbb{P}_{data}(V(\hat{x}, \mathbb{P}) > \epsilon) \leq \alpha$ or (2.4).

However, in small-sample situations in which the data size n is not large enough to support (2.6), the feasibility guarantee described above may not hold. It can be shown [22] that the minimum n that achieves (2.4) is linear in d and reciprocal in ϵ, thus may impose challenges especially in high-dimensional and low-tolerance problems. Similar dependence on the key problem parameters also appears in other related methods such as [23], which uses the Vapnik-Chervonenkis dimension to infer required sample sizes, the sampling-and-discarding approach in [20], and the closely related approach using sample average approximation in [24]. Several recent lines of techniques have been suggested to overcome these challenges and reduce sample size requirements, including the use of support rank and solution-dependent support constraints [25, 26], regularization [27], and sequential approaches [28, 29, 30, 31].

In this Chapter, we offer a different path to alleviate the data size requirement than the above methods, when $\mathbb{P}$ possesses known parametric structures. Namely, we assume $\mathbb{P} \in \{\mathbb{P}_\theta\}_{\theta \in \Theta}$ for some parametric family of distribution, where $\mathbb{P}_\theta$ satisfies two basic requirements: It is estimatable, i.e., the unknown quantity or parameter θ can be estimated from data, and simulatable, i.e., given θ, samples from $\mathbb{P}_\theta$ can be drawn using Monte Carlo methods. Under these presumptions, our approach turns the CCP (2.1), with an unknown parameter, into a CCP that has a definite parameter and a suitably re-adjusted tolerance level, which then allows us to generate enough Monte Carlo samples and consequently utilize the guarantee provided from (2.5). On a high level, this approach replaces the data size requirement in using (2.2) (or, in fact, any of its variant methods) with a Monte Carlo size requirement, the latter potentially more available given cheap modern computational power. Our methodological contributions consist of the development of procedures, related statistical results on their sample size requirement translations, and also showing some key

differences between parametric and nonparametric regimes.

Our approach starts with a distributionally robust optimization (DRO) to incorporate the data-driven parametric uncertainty. The latter is a framework for decision-making under modeling uncertainty on the underlying probability distributions in stochastic problems. It advocates the search for decisions over the worst case, among all distributions contained in a so-called uncertainty set or ambiguity set (e.g., [32, 33, 34]). In CCP, this entails a worst-case chance constraint over this set (e.g., [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]). When the uncertainty set covers the true distribution with a high confidence (i.e., the set is a confidence region), then feasibility for the distributionally robust CCP converts into a confidence guarantee on the feasibility for the original CCP. We follow this viewpoint and utilize uncertainty sets in the form of a neighborhood ball surrounding a baseline distribution, where the ball size is measured by a statistical distance (e.g., [47, 48, 49, 50, 51, 52, 53, 41, 54, 55, 56, 57, 58, 59]). In the parametric case, a suitable choice of this distance (such as the ϕ-divergence that we focus on) allows easy and meaningful calibration of the ball size from the data, so that the resulting DRO provides a provable feasibility conversion to the CCP.

Our next step is to combine this DRO with Monte Carlo sampling and scenario approximation. The definition of DRO means that there are many possible candidate distributions that can govern the truth, whereas the statistical guarantee for SO assumes a specific distribution that generates the data or Monte Carlo samples. To resolve this discrepancy, we select a *generating distribution* that draws the Monte Carlo samples, and develop a translation of the guarantee from a fixed distribution into one on the DRO. We highlight the benefits in using SO to handle this DRO, as opposed to other potential methods. While there exist many good results on tractable reformulations of DRO for chance constraints (e.g., [35, 37, 46, 38, 43]), the reformulation tightness typically relies on using moment-based uncertainty sets and particular forms of the safety condition. Compared to moments, divergence-based uncertainty sets can be calibrated with data to consistently shrink to the true distribution. Importantly, in the parametric case, the calibration of divergence-based sets is especially convenient, and achieves a tight convergence rate by using maximum likelihood theory

that efficiently captures parametric information. Our condition for applying SO to this DRO is at the same level of generality as applying SO to an unambiguous CCP, which, as mentioned before, only requires the convexity of $\mathcal{X}_\xi$ and mild conditions.

To exploit the full capability of our approach, we investigate the optimal choice of the generating distribution in relation to the target DRO, in the sense of requiring the least Monte Carlo size. We show that, if there is no ambiguity on the distribution (i.e., a standard CCP), or when the uncertainty set of a DRO is constructed via a divergence ball in the nonparametric space, the best generating distribution is, in a certain sense, the true or the baseline distribution at the center of the ball. However, if there is parametric information, the optimal choice of the generating distribution can deviate from the baseline distribution in a divergence-based DRO. We derive these results by casting the problem of selecting a generating distribution into a hypothesis testing problem, which connects the sampling efficiency of the generating distribution with the power of the test and the Neyman-Pearson lemma [60]. The results on DRO in particular combine this Neyman-Pearson machinery with the established DRO reformulation of chance constraints in [39, 41], with the discrepancy between the best generating distribution and the baseline distribution in the parametric case stemming from the removal of the extremal distributions in the corresponding nonparametric uncertainty set. These connections among hypothesis testing, SO and DRO are, to our best knowledge, the first of its kind in the literature.

Finally, given the non-optimality of the baseline distribution of a divergence-based DRO in generating Monte Carlo samples, we further develop procedures to search over generating distributions that improve upon this baseline. On a high level, this can be achieved by increasing the sampling variability to incorporate the uncertainty of the distributional parameters (one may intuit this from the perspective of a posterior distribution in a Bayesian framework), which is implemented by utilizing suitable mixture distributions. We provide several classes of mixture distributions to attain such a variability enlargement, and study descent-type algorithms to search for good distributions in these classes. In the experiments, we show our methods can be combined with SO or other SO-based methods including FAST [28] to solve a variety of optimization problems and

data distributions, some of which are not amenable to RO, especially when the objective function is non-linear or the feasible sets are jointly chance-constrained. Furthermore, we also demonstrate how to search for more judicious choices of generating distributions that can significantly reduce the required number of Monte Carlo samples.

We conclude this introduction by briefly discussing a few other lines of related literature. The first is the so-called robust Monte Carlo or robust simulation that, like us, also considers using Monte Carlo sampling together with DRO [61, 62, 63, 64, 65, 66]. However, this literature focuses on approximating DRO with stochasticity in the objective function, and does not study the chance constraint feasibility and SO that constitute our main focus. We also contrast our work with [53] that also considers likelihood theory and utilizes simulation in tackling uncertain constraints. The study [53] focuses on the nonparametric regime and uses the empirical likelihood to construct uncertainty sets. Unlike our work, there is no parametric information there that can be leveraged to overcome sample size requirements in SO. Moreover, the simulation used in [53] is for calibrating the uncertainty set, instead of drawing sample constraints. Next, [67] considers a scenario approach to distributionally robust CCP with an uncertainty set based on the Prohorov distance. Like [23], [67] utilizes the Vapnik-Chervonenkis dimension in studying feasibility, in contrast to the convexity-based argument in [22] that we utilize. More importantly, we aim to optimize the efficiency of Monte Carlo sampling in handling limited-data CCP, thus motivating us to study the choice of distance, calibration schemes, and selection of generating distributions that are different from [67]. Finally, a preliminary conference version of this work has appeared in [68], which contains a basic introduction of our framework, without detailed investigation of the optimality of generating distributions, improvement strategies, and extensive numerical demonstrations.

To summarize, our main contributions of this Chapter are:

1. We propose a framework to obtain good feasible solutions in data-driven CCPs in small-sample situations, where the data size is insufficient to support the use of SO with valid statistical guarantees. Focusing on the parametric regime, our framework operates by setting up a DRO, with an uncertainty set constructed from parameter estimates using the data, that

can in turn be tackled by using SO with Monte Carlo samples. In doing so, our framework effectively leverages the parametric information to convert the SO requirement on the data size into a requirement on the Monte Carlo size, the latter can be much more abundant given cheap modern computational power. The overview of this framework and the DRO construction are in Sections 2.2.1 and 2.2.3.

2. We investigate and present the Monte Carlo size requirements needed to give statistically feasible solutions to the divergence-based DRO used in our framework. This relies on developing an implementable mechanism to connect the sample size requirement for SO, which attempts to solve a CCP with a fixed underlying distribution, to the sample size requirement needed to solve a DRO, by selecting a suitable generating distribution to draw the Monte Carlo samples. This contribution is presented in Section 2.2.2.

3. We study the optimality of generating distributions, in a sense of minimizing the Monte Carlo effort that we will describe precisely. In particular, we show that the optimal generating distributions for an unambiguous CCP, and for a distributionally robust CCP with nonparametric divergence-based uncertainty sets, are simply their respective natural choices, namely the original underlying distribution and the baseline distribution (i.e., center of the divergence ball). In contrast, the optimal generating distribution for a distributionally robust CCP in the parametric case is more delicate, and the baseline distribution there can be readily dominated by other generating distributions. These results are derived by bridging the Neyman-Pearson lemma in statistical hypothesis testing with SO and DRO, which appears to be the first of its kind in the literature as far as we know. This contribution is presented in Section 2.3.

4. Motivated by the non-optimality of the baseline distribution, we propose several approaches to construct generating distributions that dominate the baseline distributions for parametric DRO, by using mixture schemes that, on a high level, enlarge the variability of the generating distributions. We show how to use descent-type search procedures to construct these

distributions. This contribution is presented in Section 2.4.

Lastly, we also present in full detail our implementation algorithms in Section 2.5, numerically demonstrate our approach and compare with other methods in Section 5.6, and conclude in Section 2.7.

2.2 From Data-Driven DRO to Scenario Optimization

This section introduces our overall framework. Recall our goal as to find a good (good in the sense that we still try to solve for a version of SO instead of only focusing on feasibility) feasible solution $\hat{x}$ for (2.1), and suppose that we have an i.i.d. data size n possibly less than the requirement shown in (2.6). As discussed in the introduction, we first formulate a DRO that incorporates the parametric estimation noise and subsequently allows us to resort to Monte Carlo sampling to obtain a feasible solution for (2.1). In the following, Section 2.2.1 first describes the basic guarantees from DRO. Section 2.2.2 investigates Monte Carlo sampling that provides guarantees on DRO. Section 2.2.3 discusses the choice of the uncertainty set.

2.2.1 Overview of Data-Driven DRO

For concreteness, suppose the unknown true distribution $\mathbb{P} \in \mathcal{P}$, the class of possible probability distributions for ξ (to be specified later). Given the observed data $\xi_1, ..., \xi_n$, the basic steps in our data-driven DRO are:

- Step 1: Find a data-driven uncertainty set $\mathcal{U}_{data} = \mathcal{U}_{data}(\xi_1, \ldots, \xi_n) \subseteq \mathcal{P}$ such that

$$\mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1 - \alpha, \tag{2.7}$$

  where $\mathbb{P}_{data}$ denotes the measure generating the data $\xi_i, i = 1, \ldots, n$.

- Step 2: Given $\mathcal{U}_{data}$, set up the distributionally robust CCP:

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x, \\ \text{s.t.} \quad & \min_{\mathbb{Q} \in \mathcal{U}_{data}} \mathbb{Q}(x \in \mathcal{X}_\xi) \geq 1 - \epsilon, \end{aligned} \tag{2.8}$$

  where the probability measure $\mathbb{Q}$ is the decision variable in the minimization in the constraint.

- Step 3: Find a solution $\hat{x}$ feasible for (2.8).

It is straightforward to see that $\hat{x}$ obtained from the above procedure is feasible for (2.1) with confidence at least $1 - \alpha$: If $\mathbb{P} \in \mathcal{U}_{data}$, then any $\hat{x}$ feasible for (2.8) satisfies

$$\mathbb{P}(\hat{x} \in \mathcal{X}_\xi) \geq \min_{\mathbb{Q} \in \mathcal{U}_{data}} \mathbb{Q}(\hat{x} \in \mathcal{X}_\xi) \geq 1 - \epsilon$$

Thus

$$\mathbb{P}_{data}(\mathbb{P}(\hat{x} \in \mathcal{X}_\xi) \geq 1 - \epsilon) \geq \mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1 - \alpha, \tag{2.9}$$

which gives our conclusion.

2.2.2 Monte Carlo Sampling for DRO

To use the above procedure, we need to provide a way to construct the depicted $\mathcal{U}_{data}$ and to find a (confidently) feasible solution for (2.8). We postpone the set construction to the next subsection and focus on finding a feasible solution here. We resort to SO, via Monte Carlo sampling, to handle (2.8). Note that, unlike in the standard SO discussed in the introduction, the distribution $\mathbb{Q}$ here can be any candidate within the set $\mathcal{U}_{data}$. Thus, let us select a generating distribution, called $\mathbb{P}_0$ (which can depend on the data), to generate Monte Carlo samples $\xi_i^{MC}, i = 1, \ldots, N$, and solve

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x, \\ \text{s.t.} \quad & x \in \mathcal{X}_{\xi_i^{MC}}, \ i = 1, \ldots, N. \end{aligned} \tag{2.10}$$

For convenience, denote, for any $\epsilon, \beta > 0$,

$$N_{exact}(\epsilon, \beta, d) = \min \left\{ n : \sum_{i=0}^{d-1} \binom{n}{i} \epsilon^i (1-\epsilon)^{n-i} \leq \beta \right\}. \tag{2.11}$$

From the result of [22] discussed in the introduction, using $N_{exact}(\epsilon, \beta, d)$ or more Monte Carlo samples from $\mathbb{P}_0$ in (2.10) would give a solution $\hat{x}^{MC}$ that satisfies $V(\hat{x}^{MC}, \mathbb{P}_0) \leq \epsilon$ with confidence level $1 - \beta$. This is not exactly the distributionally robust feasibility statement for problem (2.8). To address this discrepancy, we consider, conditional on the data $\xi_1, \ldots, \xi_n$,

$$\begin{aligned} \max_{\mathbb{Q} \in \mathcal{U}_{data}} \quad & V(\hat{x}^{MC}, \mathbb{Q}) \\ \text{s.t.} \quad & V(\hat{x}^{MC}, \mathbb{P}_0) \leq \delta. \end{aligned} \tag{2.12}$$

This optimization problem serves to translate a guarantee on the violation probability under $\mathbb{P}_0$ to any $\mathbb{Q}$ in $\mathcal{U}_{data}$. If we can bound the optimal value in (2.12), then we can trace back the level of δ that is required to ensure a chance constraint validity of tolerance level ϵ. However, the event involved in defining $V(\hat{x}^{MC}, \mathbb{P}_0)$ and $V(\hat{x}^{MC}, \mathbb{Q})$, namely $\{\xi : \hat{x}^{MC} \notin \mathcal{X}_\xi\}$, can be challenging to handle in general. Thus, we relax (2.12) to

$$\begin{aligned} \max_{\mathbb{Q} \in \mathcal{U}_{data}, A \subset \mathcal{Y}} \quad & \mathbb{Q}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.13}$$

where the decision variables now include the set A in addition to the probability measure $\mathbb{Q}$. Conditional on the data $\xi_1, \ldots, \xi_n$, the optimal value of optimization problem (2.13), which we denote $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$, is clearly an upper bound for that of (2.12). In fact, it is also clear from (2.13) that $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ is non-decreasing in $\delta > 0$ and

$$\max_{\mathbb{Q} \in \mathcal{U}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \leq M(\mathbb{P}_0, \mathcal{U}_{data}, V(\hat{x}^{MC}, \mathbb{P}_0)), \tag{2.14}$$

by simply taking $A = \{\xi : \hat{x}^{MC} \notin \mathcal{X}_\xi\}$ and $\delta = V(\hat{x}^{MC}, \mathbb{P}_0)$ in (2.13). We have the following guarantee:

Theorem 2.2.1. *Given $\mathbb{P}_0$, $\mathcal{U}_{data}$ and $\epsilon > 0$, suppose there exists $\delta_\epsilon > 0$ small enough such that*

$$M(\mathbb{P}_0, \mathcal{U}_{data}, \delta_\epsilon) \leq \epsilon, \tag{2.15}$$

then if we solve (2.10) *with $N_{exact}(\delta_\epsilon, \beta, d)$ number of samples drawn from $\mathbb{P}_0$, the obtained solution $\hat{x}^{MC}$ would be feasible for* (2.8) *with confidence at least $1 - \beta$. Furthermore, if*

$$\mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1 - \alpha, \tag{2.16}$$

where $\mathbb{P}_{data}$ is the measure governing the real-data generation under the true distribution $\mathbb{P}$, then the obtained solution $\hat{x}^{MC}$ would be feasible for (2.1) *with confidence at least $1 - \alpha - \beta$.*

Proof. By results in [22], we know that by solving (2.10) with $N_{exact}(\delta_\epsilon, \beta, d)$ number of samples from $\mathbb{P}_0$, the obtained solution $\hat{x}^{MC}$ would satisfy

$$\mathbb{P}_{MC,0}(V(\hat{x}^{MC}, \mathbb{P}_0) > \delta_\epsilon) \leq \beta \tag{2.17}$$

where $\mathbb{P}_{MC,0}$ is the measure with respect to the Monte Carlo samples drawn from $\mathbb{P}_0$. Moreover, based on the monotonicity property of $M(\cdot)$ and (2.14), we have

$$V(\hat{x}^{MC}, \mathbb{P}_0) \leq \delta_\epsilon \implies \max_{\mathbb{Q} \in \mathcal{U}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) \leq M(\mathbb{P}_0, \mathcal{U}_{data}, \delta_\epsilon). \tag{2.18}$$

Thus (2.15) implies that

$$\mathbb{P}_{data}\left(\max_{\mathbb{Q} \in \mathcal{U}_{data}} V(\hat{x}^{MC}, \mathbb{Q}) > \epsilon\right) \leq \mathbb{P}_{data}(V(\hat{x}^{MC}, \mathbb{P}_0) > \delta_\epsilon) \leq \beta$$

and hence $\hat{x}^{MC}$ is feasible for (2.8) with confidence at least $1 - \beta$. Furthermore, if $\mathbb{P} \in \mathcal{U}_{data}$, then

a $\hat{x}^{MC}$ feasible for (2.8) is also feasible for (2.1) since $\max_{\mathbb{Q}\in\mathcal{U}_{data}} V(\hat{x}^{MC},\mathbb{Q}) \geq V(\hat{x}^{MC},\mathbb{P})$ and hence

$$\max_{\mathbb{Q}\in\mathcal{U}_{data}} V(\hat{x}^{MC},\mathbb{Q}) \leq \epsilon \implies V(\hat{x}^{MC},\mathbb{P}) \leq \epsilon. \tag{2.19}$$

Thus, if we denote $\Xi = \{\xi_1, ..., \xi_n, \xi_1^{MC}, ..., \xi_N^{MC}\}$ to be entire sequence consisting of real data and the generated Monte Carlo samples, it then follows that

$$\{\Xi : V(\hat{x}^{MC},\mathbb{P}) > \epsilon\} \subseteq \{\Xi : \mathbb{P} \notin \mathcal{U}_{data}\} \cup \{\Xi : V(\hat{x}^{MC},\mathbb{P}_0) > \delta_\epsilon\}. \tag{2.20}$$

It now follows by (2.16) and (2.17) that $\hat{x}^{MC}$ is feasible for (2.1) with probability at least $1-\alpha-\beta$. □

Theorem 2.2.1 can be cast in terms of asymptotic instead of finite-sample guarantees by following the same line of arguments. We summarize it as the following corollary.

Corollary 2.2.1.1. *In Theorem 2.2.1, if the condition $\mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1-\alpha$ is substituted by the asymptotic condition*

$$\liminf_{n\to\infty} \mathbb{P}_{data}(\mathbb{P} \in \mathcal{U}_{data}) \geq 1-\alpha, \tag{2.21}$$

then the feasibility of $\hat{x}^{MC}$ in the last conclusion of Theorem 2.2.1 holds with confidence asymptotically tending to at least $1-\alpha-\beta$.

To summarize, in the presence of data insufficiency, if we choose $\mathcal{U}_{data}$ to satisfy the confidence property (2.7), and are able to evaluate the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ that translates the violation probability under $\mathbb{P}_0$ to a worst-case violation probability over $\mathcal{U}_{data}$, then we can run SO with $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples from $\mathbb{P}_0$ to obtain a solution for (2.1) with confidence $1-\alpha-\beta$.

We also note that the above scheme still holds if the $N_{exact}(\epsilon, \beta, d)$ in (2.11) is replaced by the sample size requirements of other variants of SO (e.g., FAST [28]) that are potentially smaller. This works as long as we stay with the same SO-based procedure in using the Monte Carlo samples. For

clarity, throughout most of our exposition we will focus on the sample size requirement depicted in (2.11), but we will discuss other variants in our implementation and numerical sections.

Finally, let us take a step back and justify why we use SO to tackle (2.8), as opposed to other potential means. Indeed, as pointed out in the introduction, there exist many good results on tractable reformulations of DRO. As will be discussed in detail in the next subsection, in the present context we will choose an uncertainty set that can leverage parametric information efficiently. Sets based on the neighborhoods of distributions measured by ϕ-divergences are particularly attractive choices, as they can be calibrated easily (both the ball center and the size) in a way that efficiently uses parametric information. The dependence on the parameter dimension in particular is reflected in the degree of freedom in the χ^2-distribution used in the calibrating the ball size, which shrinks to zero at a canonical rate as the data size increases. Other sets, such as moment-based ones, though possibly amenable to tight tractable reformulations, do not enjoy these statistical properties in the parametric context. Thus, in view of tackling ϕ-divergence-based DRO, SO appears to be a natural choice, and we have set up a framework to utilize it under conditions at the same level of generality as required for the unambiguous counterpart. Sections 2.3 and 2.4 will study this framework in further depth and enhance its efficiency. We caution, however, that the conservativeness in our proposed uncertainty set (which affects the optimality of the obtained solution) relies on the dimensionality of the distributional parameters. Our approach is expected to work well when this dimension is moderate, but not in high-dimensional problems where other approaches could be better choices.

2.2.3 Constructing Uncertainty Sets

In this section we discuss the construction of the uncertainty set $\mathcal{U}_{data}$, using the ϕ-divergence approach [47]. We assume the true distribution $\mathbb{P}$ of ξ lies in a parametric family. We denote the true parameter as θ_{true}. To highlight the parametric dependence, we call the true distribution $\mathbb{P}_{\theta_{true}} \in \mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta\in\Theta\subset\mathbb{R}^D}$ indexed by θ, where D is the dimension of parameter space. Given

data $\xi_1, \xi_2, ..., \xi_n$, we want to construct an uncertainty set $\mathcal{U}_{data}$ satisfying

$$\lim_{n\to\infty} \mathbb{P}_{data}(\mathbb{P}_{\theta_{true}} \in \mathcal{U}_{data}) = 1 - \alpha \tag{2.22}$$

so that Corollary 2.2.1.1 applies. To do so, we first estimate θ_{true} from the data. There are various approaches to do so; here we apply the common maximum likelihood estimator (MLE) $\hat{\theta}_n$, and set $\mathcal{U}_{data}$ to be

$$\mathcal{U}_{data} = \left\{ \mathbb{Q} \in \mathcal{P}_{para} : d_\phi(\mathbb{P}_{\hat{\theta}_n}, \mathbb{Q}) \leq \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n} \right\}, \tag{2.23}$$

where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of χ^2_D, the χ^2-distribution with degree of freedom D, and $d_\phi(\cdot,\cdot)$ is the ϕ-divergence between two probability measures, i.e., given a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, with $\phi(1) = 0$, a distance between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ defined as

$$d_\phi(\mathbb{P}_1, \mathbb{P}_2) = \int_{\mathcal{Y}} \phi\left(\frac{d\mathbb{P}_2}{d\mathbb{P}_1}\right) \mathbb{P}_1(dy), \tag{2.24}$$

assuming $\mathbb{P}_2$ is absolutely continuous with respect to $\mathbb{P}_1$ with Radon-Nikodym derivative $\frac{d\mathbb{P}_2}{d\mathbb{P}_1}$ on $\mathcal{Y}$. Moreover, we assume that ϕ is twice continuously differentiable with $\phi''(1) \neq 0$, and if necessary set the continuation of ϕ to $\mathbb{R}_-$ as $\phi(x) = +\infty$ for $x < 0$. In (2.23), we call the center $\mathbb{P}_{\hat{\theta}_n}$ of the divergence ball, the baseline distribution.

To guarantee desirable asymptotic properties of our uncertainty set, we make the following assumption:

Assumption 1. *Let $\theta_{true} \in \Theta$ be the true parameter and let $\hat{\theta}_n$ be the MLE of θ_{true} estimated from n i.i.d. data points. Then, as $n \to \infty$, $\hat{\theta}_n$ satisfies consistency and asymptotic normality condition:*

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_{true} \quad \textit{and} \quad \sqrt{n}(\hat{\theta}_n - \theta_{true}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_{true})), \tag{2.25}$$

where $\mathcal{I}(\theta)$ is the Fisher information for the parametric family $\mathcal{P}_{para}$ with well-defined inverse that is continuous in the domain $\theta \in \Theta$.

Assumption 1 of MLE estimator is known to hold under various regularity conditions [69, 70]. We list a set of such conditions in supplementary section 2.8.

Under Assumption 1, it can be shown [71, 69] that $\mathcal{U}_{data}$ in (2.23) satisfies the confidence guarantee (2.22). Furthermore, since we can identify each $\mathbb{P}_\theta$ in $\mathcal{P}_{data}$ with θ, we can equivalently view $\mathcal{U}_{data}$ as a subset of $\theta \in \Theta$, and write it as

$$\mathcal{U}_{data} \triangleq \left\{ \theta \in \Theta : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_\theta) \leq \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n} \right\}. \tag{2.26}$$

For convenience, we shall use the two definitions of $\mathcal{U}_{data}$ interchangeably depending on the context. It is also known that the asymptotic confidence properties of (2.23) or (2.26) are the same among different choices within the ϕ-divergence class. These can be seen via a second order expansion of the ϕ-divergences. Moreover, they are asymptotically equivalent to

$$\left\{ \theta \in \Theta : (\theta - \hat{\theta}_n)^T \mathcal{I}(\hat{\theta}_n)(\theta - \hat{\theta}_n) \leq \frac{\chi^2_{1-\alpha,D}}{n} \right\}, \tag{2.27}$$

where $\mathcal{I}(\hat{\theta}_n)$ is the estimated Fisher information, under the regularity conditions above [71, 72, 69]. In other words, under Assumption 1, both (2.26) and (2.27) satisfy

$$\lim_{n\to\infty} \mathbb{P}_{data}(\theta_{true} \in \mathcal{U}_{data}) = 1 - \alpha. \tag{2.28}$$

Note that the convergence rate of (2.22) or (2.28) depends on the higher-order properties of the parametric model, which in turn can depend on the parameter dimension. Different from the sample size requirements in SO, this convergence rate is a consequence of MLE properties. Some details on finite-sample behaviors of MLE can be found in [73].

The $\mathcal{U}_{data}$ discussed above is a set over the parametric class of distributions (or parameter values). Considering tractability, DRO over nonparametric space could be easier to handle than parametric, which suggests a relaxation of the parametric constraint to estimate the bounding function M. This also raises the question of whether one can possibly contain $\mathcal{U}_{data}$ in a nonparametric

ball with a shrunk radius and subsequently obtain a better M. These would be the main topics of Sections 2.3 and 2.4.

2.3 Bounding Functions and Generating Distributions

Given the uncertainty set $\mathcal{U}_{data}$ in (2.26), we turn to the choice of the generating measure $\mathbb{P}_0$ and the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ which, as we recall, is the optimal value of optimization problem (2.13). In the discussed parametric setup, the latter becomes

$$\begin{aligned} \max_{\theta \in \mathcal{U}_{data}, A \subset \mathcal{Y}} \quad & \mathbb{P}_\theta(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.29}$$

From Theorem 2.2.1 and the fact that $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ is non-decreasing in δ, we want to choose $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ so that we can take the maximum δ_ϵ and subsequently achieve overall confident feasibility with the least Monte Carlo sample size. Note that $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ is a multi-input function depending on both $\mathbb{P}_0$ and δ, and so a priori it is not clear that a uniform minimizer $\mathbb{P}_0$ can exist across all values of δ so that the described task is well-defined. It turns out that this is possible in some cases, which we shall investigate in detail. In the following, we discuss results along this line at three levels: The unambiguous case, namely when $\mathcal{U}_{data}$ in (2.29) is a singleton (Section 2.3.1), the case where $\mathcal{U}_{data}$ is nonparametric (Section 2.3.2), and the case where $\mathcal{U}_{data}$ is parametric (Section 2.3.3). The first two cases pave the way to the last one, which is most important to our development and also motivates Section 2.4. With these results in hand, we also discuss the possibility of using other statistical distances in our framework in Section 2.3.4.

2.3.1 Neyman-Pearson Connections and A Least Powerful Null Hypothesis

We first consider, for a given $\theta_1 \in \mathcal{U}_{data}$, the optimization problem

$$\begin{aligned} \max_{A \subset \mathcal{Y}} \quad & \mathbb{P}_{\theta_1}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.30}$$

This problem can be viewed as choosing a most powerful decision rule in a statistical hypothesis test. More precisely, one can think of A as a rejection region for a simple test with null hypothesis $\mathbb{P}_0$ and alternate hypothesis $\mathbb{P}_{\theta_1}$. Subject to a tolerance of δ Type-I error, optimization problem (2.30) looks for a decision rule that maximizes the power of the test. By the Neyman-Pearson lemma [60], under mild regularity conditions on the parametric family, the optimal set $A^\star_{0,\theta_1,\delta}$ of (2.30) takes the form

$$A^\star_{0,\theta_1,\delta} = \{\xi \in \mathcal{Y} : \frac{d\mathbb{P}_{\theta_1}}{d\mathbb{P}_0}(\xi) > K^\star_{0,\theta_1,\delta}\}, \tag{2.31}$$

with $K^\star_{0,\theta_1,\delta}$ chosen so that $\mathbb{P}_0(A^\star_{0,\theta_1,\delta}) = \delta$. Also, then, the optimal value of (2.30) is $\mathbb{P}_{\theta_1}(A^\star_{0,\theta_1,\delta})$. Generalizing the above analysis to all $\theta \in \mathcal{U}_{data}$, we conclude that

$$M(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = \sup_{\theta \in \mathcal{U}_{data}} \mathbb{P}_\theta(A^\star_{0,\theta,\delta}), \tag{2.32}$$

is the optimal value of (2.29). These observations will be useful for deriving our subsequent results.

Our goal is to choose $\mathbb{P}_0$ to minimize (2.32). To start our analysis, let us first consider the extreme case where the uncertainty set $\mathcal{U}_{data}$ consists of only one point $\mathbb{Q}$. In this case, we look for $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$, the optimal value of

$$\begin{aligned} \max_{A \subset \mathcal{Y}} \quad & \mathbb{Q}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.33}$$

That is, for a given measure $\mathbb{Q}$, we seek for the maximum discrepancy between $\mathbb{Q}$ and $\mathbb{P}_0$ over all $\mathbb{P}_0$-measure sets that have δ or less content. This is similar to minimizing the total variation

distance between $\mathbb{Q}$ and $\mathbb{P}_0$, and hints that the optimal choice of $\mathbb{P}_0$ is $\mathbb{Q}$. The following theorem, utilizing the Neyman-Pearson lemma depicted above, confirms this intuition. We remark that the assumptions of the theorem can be relaxed by using more general versions of the lemma, but the presented version suffices for most purposes and also the subsequent examples we will give.

Theorem 2.3.1. *Given a measure $\mathbb{Q}$ with continuous density on $\mathcal{X}$, among all $\mathbb{P}_0$ such that $\frac{d\mathbb{Q}}{d\mathbb{P}_0}$ exists and is continuous and positive almost surely, the minimum $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$ is obtained by choosing $\mathbb{P}_0 = \mathbb{Q}$, giving $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) = \delta$.*

Proof. Under the assumptions, by the Neyman-Pearson lemma, for a fixed measure $\mathbb{P}_0$, the set achieving the optimal value of (2.33) takes the form $A^\star = \{\xi \in \mathcal{Y} : \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K^\star\}$ for some $K^\star \geq 0$ with $\mathbb{P}_0(A^\star) = \delta$. It then follows that

$$M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) - \delta = \mathbb{Q}(A^\star) - \mathbb{P}_0(A^\star) = \int_{\frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K^\star} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi).$$

Under the absolute continuity assumption, we define

$$g(K) = \int_{\frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi),$$

which can be seen to be a non-increasing function for $K \geq 1$ and a non-decreasing function for $K \leq 1$. To see this, take $K_1 \geq K_2$, and we have

$$g(K_2) = g(K_1) + \int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi).$$

Thus, when $K_1 \geq K_2 \geq 1$, we have $g(K_2) \geq g(K_1)$ because

$$\int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) \geq (K_2 - 1)\mathbb{P}_0(K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2) \geq 0,$$

while when $1 \geq K_1 \geq K_2$, we have $g(K_2) \leq g(K_1)$ because

$$\int_{K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2} (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) \leq (K_1 - 1)\mathbb{P}_0(K_1 \geq \frac{d\mathbb{Q}}{d\mathbb{P}_0}(\xi) > K_2) \leq 0.$$

Then, to identify the minimum of $g(K)$, we either decrease K from 1 to 0 which gives

$$\liminf_{K \to 0} g(K) = \int (\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1) d\mathbb{P}_0(\xi) = 0, \tag{2.34}$$

by using the dominated convergence theorem (e.g., by considering the set $\{1 > d\mathbb{Q}/d\mathbb{P}_0(\xi) > K\}$) or we increase K from 1 to ∞ which gives

$$\liminf_{K \to \infty} g(K) \geq 0. \tag{2.35}$$

by Fatou's lemma. Observations (2.34) and (2.35) suggest that $g(K) \geq 0$ for all $K \geq 0$ and imply that $g(K^\star) \geq 0$. Thus, we must have $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta) \geq \delta$. Note that this holds for any $\mathbb{P}_0$. Now, since choosing $\mathbb{P}_0 = \mathbb{Q}$ gives $M(\mathbb{Q}, \{\mathbb{Q}\}, \delta) = \delta$, an optimal choice of $\mathbb{P}_0$ is $\mathbb{Q}$. □

Theorem 2.3.1 shows that under mild regularity conditions, in terms of choosing the generating distribution $\mathbb{P}_0$ and minimizing $M(\mathbb{P}_0, \{\mathbb{Q}\}, \delta)$, we cannot do better than simply choosing $\mathbb{Q}$ itself. This means that if we had known the true distribution was $\mathbb{Q}$, and without additional knowledge of the event of interest, the safest choice (in the minimax sense) for sampling would be $\mathbb{Q}$, a quite intuitive result. In the language of hypothesis testing, given the simple alternate hypothesis $\mathbb{Q}$, the null hypothesis $\mathbb{P}_0$ that provides the least power for the test, i.e., makes it most difficult to distinguish between the two hypotheses, is $\mathbb{Q}$.

2.3.2 Nonparametric DRO

Building on the discussion in Section 2.3.1, we now consider the choice of generating distribution $\mathbb{P}_0$ to minimize the bounding function obtained from (2.29). Before so, we first discuss the

nonparametric case, where the analog of (2.29) is in the form:

$$\begin{aligned} \max_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda, A\subset\mathcal{Y}} \quad & \mathbb{Q}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A)\le\delta. \end{aligned} \tag{2.36}$$

for some ball radius $\lambda > 0$, where the decision variables are $\mathbb{Q}$ in the space of all distributions absolutely continuous with respect to $\mathbb{P}_{\hat{\theta}}$, and A.

We show that the above setting can be effectively reduced to the unambiguous case, i.e., when $\mathbb{Q}$ lies in a singleton discussed in Section 2.3.1. This comes from an established equivalence between a distributionally robust chance constraint and an unambiguous chance constraint evaluated by the center of the divergence ball, when the event A is fixed [41, 39]. In particular, suppose the stochasticity space is $\mathcal{Y} = \mathbb{R}^k$, and $\mathbb{P}_{\hat{\theta}}$ admits a density $p_{\hat{\theta}}$. Theorem 1 in [39] shows that for any A,

$$\max_{d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le\lambda} \mathbb{Q}(A)\le\epsilon \iff \mathbb{P}_{\hat{\theta}}(A)\le\epsilon', \tag{2.37}$$

where $\epsilon' = \epsilon'(\epsilon,\lambda,\phi) > 0$ can be explicitly determined by ϵ, λ and ϕ as

$$\epsilon'(\epsilon,\lambda,\phi) = \max\left\{1 - \inf_{\substack{z>0, z+\pi z\le\ell_\phi \\ \underline{m}(\phi^*)\le z_0+z\le\overline{m}(\phi^*)}} \left\{\frac{\phi^*(z_0+z)-z_0-\epsilon z+\lambda}{\phi^*(z_0+z)-\phi^*(z_0)}\right\}, 0\right\} \tag{2.38}$$

with $\phi^*(t) = \sup_x\{tx - g(x)\}$ being the conjugate function of ϕ and $\underline{m}(\phi^*) = \sup\{m\in\mathbb{R} : \phi^* \text{ is a finite constant on } (-\infty, m]\}$, $\overline{m}(\phi^*) = \inf\{m\in\mathbb{R} : \phi^*(m) = +\infty\}$, $\ell_\phi = \lim_{x\to+\infty}\phi(x)/x$, and $\pi = -\infty$ if $Leb\{[p_{\hat{\theta}} = 0]\} = 0$, 0 if $Leb\{[p_{\hat{\theta}} = 0]\} > 0$ and $Leb\{[p_{\hat{\theta}} = 0]\setminus A\} = 0$, and 1 otherwise, where $Leb\{\cdot\}$ is the Lebesgue measure on $\mathbb{R}^k$.

The above equivalence can be used to obtain the following result.

Theorem 2.3.2. *Suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density. Among all $\mathbb{P}_0$ such that $\frac{d\mathbb{P}_{\hat{\theta}}}{d\mathbb{P}_0}$ exists and is continuous, positive almost surely, an optimal choice of $\mathbb{P}_0$ that minimizes $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}},\mathbb{Q})\le$*

$\lambda\}, \delta)$*, namely the optimal value of* (2.36)*, is the center of the* ϕ*-divergence ball* $\mathbb{P}_{\hat{\theta}}$*. Moreover, this gives* $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta) = \epsilon'^{(-1)}(\delta, \lambda, \phi)$*, where* $\epsilon'^{(-1)}(\cdot, \lambda, \phi)$ *is the inverse of the function* $\epsilon' = \epsilon'(\epsilon, \lambda, \phi)$ *defined in* (2.38) *with respect to* ϵ*, given by*

$$\epsilon'^{(-1)}(x, \lambda, \phi) \triangleq \min\{\epsilon \geq 0 : \epsilon'(\epsilon, \lambda, \phi) \geq x\} \tag{2.39}$$

Proof. From Theorem 1 in [39], we know that, for any $A \subset \mathcal{Y}$ and $0 \leq \epsilon \leq 1$, (2.37) holds. We can rewrite the optimal value of problem (2.36) in the form:

$$\begin{aligned} \min_{\epsilon \geq 0} \quad & \epsilon \\ \text{s.t.} \quad & \max_{d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda} \mathbb{Q}(A) \leq \epsilon \text{ for all } A \subset \mathcal{Y} \text{ such that } \mathbb{P}_0(A) \leq \delta, \end{aligned} \tag{2.40}$$

which, according to (2.37), has the same optimal value as

$$\begin{aligned} \min_{\epsilon \geq 0} \quad & \epsilon \\ \text{s.t.} \quad & \mathbb{P}_{\hat{\theta}}(A) \leq \epsilon' \text{ for all } A \subset \mathcal{Y} \text{ such that } \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.41}$$

Since, fixing ϕ and λ, ϵ' is a non-decreasing function of ϵ, we see that minimizing ϵ is equivalent to minimizing ϵ'. Denoting $\nu^\star$ as the optimal value of the optimization problem

$$\begin{aligned} \max_{A \subset \mathcal{Y}} \quad & \mathbb{P}_{\hat{\theta}}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta, \end{aligned} \tag{2.42}$$

then the optimal value of (2.41) is $\epsilon'^{(-1)}(\nu^*, \lambda, \phi)$. Moreover, this is achievable by setting $\mathbb{P}_0 = \mathbb{P}_{\hat{\theta}}$ that gives the optimal value $\nu^* = \delta$ to (2.42) by Theorem 2.3.1.

□

An implication of Theorem 2.3.2 is that, by noting that a parametric divergence ball lies inside a corresponding nonparametric ball, we can compute a bound for M to obtain a required Monte Carlo

size, drawn from the baseline $\mathbb{P}_{\hat{\theta}}$, to get a feasible solution for the distributionally robust CCP (2.8) and subsequently the CCP (2.1). More precisely, recall the bounding function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ with $\lambda = \phi''(1)\chi^2_{1-\alpha,D}/(2n)$, given by (2.29), as the optimal value of

$$\begin{aligned} \max_{d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}, A \subset \mathcal{Y}} \quad & \mathbb{Q}(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.43}$$

We have:

Corollary 2.3.2.1. *Given a data size n, suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density, where $\hat{\theta}$ is the MLE under Assumption 1. If we choose $\delta_\epsilon = \epsilon'(\epsilon, \phi''(1)\chi^2_{1-\alpha,D}/(2n), \phi)$ and draw $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples from the generating distribution $\mathbb{P}_{\hat{\theta}}$ to construct the sampled problem* (2.10)*, then the obtained solution will be feasible for* (2.1) *with asymptotic confidence level at least* $1 - \alpha - \beta$.

Proof. Note that a parametric divergence ball lies inside a corresponding nonparametric ball in the sense that

$$\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}$$

Thus, by the definition of M, we have

$$M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \leq M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta)$$

In particular,

$$M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \leq M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta) = \epsilon'^{-1}(\delta, \lambda, \phi)$$

where the equality follows from Theorem 2.3.2. Thus, if we choose δ_ϵ such that $\epsilon'^{-1}(\delta_\epsilon, \lambda, \phi) \leq \epsilon$, or $\delta_\epsilon = \epsilon'(\epsilon, \lambda, \phi)$, where $\lambda = \phi''(1)\chi^2_{1-\alpha,D}/(2n)$ as presented in (2.23), and the generating distribution as $\mathbb{P}_{\hat{\theta}}$, then Corollary 2.2.1.1 guarantees that running SO on $N_{exact}(\delta_\epsilon, \beta, d)$ Monte Carlo samples gives a feasible solution for (2.1) with confidence asymptotically at least $1 - \alpha -$

β. □

Corollary 2.3.2.1 thus provides an implementable procedure to handle (2.1) through (2.8).

2.3.3 Parametric DRO

Next we discuss further the choice of generating distributions in parametric DRO beyond $\mathbb{P}_{\hat{\theta}}$. While the ball center $\mathbb{P}_{\hat{\theta}}$ is a valid choice, the equivalence relation (2.37) does not apply when the divergence ball is in a parametric class, and the optimal choice of the generating distribution may no longer be $\mathbb{P}_{\hat{\theta}}$, as shown in the next result.

Theorem 2.3.3. *In terms of selecting a generating distribution $\mathbb{P}_0$ to minimize $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$, the optimal value of* (2.43)*, the choice $\mathbb{P}_{\hat{\theta}}$ can be strictly dominated by other distributions.*

Intuitively, Theorem 2.3.3 arises because the extreme distribution that achieves the equivalence relation (2.37) may not be in the considered parametric family. It implies more flexibility in choosing the generating measure $\mathbb{P}_0$, in the sense of requiring less Monte Carlo samples than using $\mathbb{P}_{\hat{\theta}}$.

From the standpoint of hypothesis testing in Section 2.3.1, the imposed minimax problem (2.43) in searching for the best $\mathbb{P}_0$ can be viewed as finding a simple null hypothesis that is uniformly least powerful across the uncertainty set. This question is related and appears more general than finding the least favorable or powerful prior in testing against composite null hypothesis [60]. In the latter context, given a set Θ_1, one aims to find a distribution $\mu^\star(d\theta_0)$ such that $\Gamma(\mu^\star) \leq \Gamma(\mu)$ for all distributions $\mu(d\theta_0)$ on Θ_0, where $\Gamma(\mu)$ is the optimal value of

$$\begin{aligned} \max_{\theta_1 \in \Theta_1} \quad & \mathbb{P}_{\theta_1}(A) \\ \text{s.t.} \quad & \int_{\Theta_0} \mathbb{P}_{\theta_0}(A)\mu(d\theta_0) \leq \delta. \end{aligned} \tag{2.44}$$

The distribution $\mu(d\theta_0)$ is interpreted as a prior on a composite null hypothesis parametrized by θ_0, and $\mu^*(d\theta_0)$ is the least favorable prior. The difference between (2.44) and our formulation (2.43) lies in the restriction to measures of the form $\mathbb{P}_0 = \int_{\Theta_0} \mathbb{P}_{\theta_0}\mu(d\theta_0)$ for the former, leading

to a smaller search space than ours. This mixture-type $\mathbb{P}_0$ and the Bayesian connection will partly motivate our investigation in Section 2.4.

To prove Theorem 2.3.3, we present a counter example and also some related discussion.

Consider the uncertainty set $\mathcal{U}_{data} = \{\mathbb{P}_\theta, : -1 \leq \theta \leq 1\}$ within Gaussian location family on $\mathbb{R}$ with $\mathbb{P}_\theta(dy) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\theta)^2}{2}}$. This can be thought of, e.g., as an uncertainty set based on the χ^2-distance, the latter defined between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ as

$$\chi^2(\mathbb{P}_1, \mathbb{P}_2) = \int_{\mathcal{Y}} (\frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1)^2 \mathbb{P}_1(dy). \tag{2.45}$$

Note that the χ^2-distance is in the family of ϕ-divergences, by choosing $\phi = (x-1)^2$. We aim to find a generating distribution $\mathbb{P}_0$ to minimize $M(\mathbb{P}, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, \delta)$, the optimal value of

$$\begin{aligned} \max_{\theta \in \mathcal{U}_{data}, A \subset \mathbb{R}} \quad & \mathbb{P}_\theta(A) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.46}$$

We consider several symmetric distributions as $\mathbb{P}_0$ (symmetry is reasonably conjectured as a good property since an imbalanced shift might increase the power for the alternative hypothesis on one side and the worst case overall). We list these symmetric distributions in increasing variability:

$$\begin{aligned} \mathbb{P}_0^1(dy) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\ \mathbb{P}_0^2(dy) &= \frac{1}{\sqrt{2\pi \cdot 2}} e^{-\frac{y^2}{2\cdot 2}} \\ \mathbb{P}_0^3(dy) &= \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(y-1)^2}{2}} + e^{-\frac{(y+1)^2}{2}} \right). \end{aligned} \tag{2.47}$$

Given $0 \leq \theta \leq 1$, it can be shown by the Neyman-Pearson lemma that the rejection region $A^\star$ (i.e. the set giving the optimal value of (2.46) for a given θ) for $\mathbb{P}_0^1$ has the form $\{y : y > c_1\}$, for $\mathbb{P}_0^2$ the form $\{y : y - 2\theta \leq c_2\}$ and for $\mathbb{P}_0^3$ the form $\{y : \frac{e^{\theta y}}{e^y + e^{-y}} > c_3\}$, for some c_1, c_2 and c_3. Let $\delta = 0.05$

be the tolerance level, it can be shown through numerical verification that

$$\begin{aligned} M(\mathbb{P}_0^1, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.2595 \\ M(\mathbb{P}_0^2, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.1160 \\ M(\mathbb{P}_0^3, \{\mathbb{P}_\theta : \theta \in \mathcal{U}_{data}\}, 0.05) &= 0.0995. \end{aligned} \tag{2.48}$$

Thus, the natural choice $\mathbb{P}_{\hat{\theta}} = \mathbb{P}_0^1$ based on relaxing to nonparametric DRO yields a bounding function $M(\cdot)$ that is outperformed by $\mathbb{P}_0^2$ or $\mathbb{P}_0^3$. Later in Section 2.4 we will see numerically how $\mathbb{P}_0^2$ and $\mathbb{P}_0^3$ can lead to a smaller sample size requirements.

Although Theorem 2.3.3 reveals room to search for the best generating distribution, the involved optimization, or even just finding an improved distribution over $\mathbb{P}_{\hat{\theta}}$, appears to be nontrivial. In particular, the maximization problem in (2.43) depends on the computation of $A^\star$ for each alternative of $\theta \in \mathcal{U}_{data}$. Section 2.4 discusses some approaches to search for improvements. We conclude the current section with some discussion on the choice of statistical distances used in the uncertainty set.

2.3.4 Choice of Statistical Distance

We have chosen to use ϕ-divergence to construct our uncertainty set $\mathcal{U}_{data}$, and we have seen how this allows us to effectively translate sample size requirements from the data to Monte Carlo. Note that another common type of distance is the Wasserstein distance (e.g., [56, 57, 59]). If one can translate the violation probability under a generating distribution into the worst-case violation probability over a Wasserstein ball, then the same line of arguments in Section 2.2 applies to using SO on this DRO. Presuming that the size of a parametric Wasserstein-based confidence region can be properly calibrated from data, it is conceivable that the above can give rise to an alternate solution route. It is known (Theorem 3 in [57]), under suitable regularity conditions, that one can equate a Wasserstein-ambiguous probability $\sup_{d_W(\mathbb{Q},\mathbb{P}_{\hat{\theta}})\leq\lambda} \mathbb{Q}(\xi \in A)$, where d_W denotes a Wasserstein distance of order 1 and cost function c, and A is an event, to $\mathbb{P}_{\hat{\theta}}(c(\xi, A) \leq 1/\nu^*)$ where $\nu^* \geq 0$ is a dual multiplier for the associated optimization problem, and $c(\xi, A)$ denotes the cost-induced

distance between a point ξ and a set A. Thus, $M(\mathbb{P}_0, \{\mathbb{Q} : d_W(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta)$ can be written as

$$\begin{aligned} \max_{A \subset \mathcal{Y}} \quad & \mathbb{P}_{\hat{\theta}}(c(\xi, A) \leq 1/\nu^*) \\ \text{s.t.} \quad & \mathbb{P}_0(A) \leq \delta. \end{aligned} \tag{2.49}$$

Compared to the evaluation of $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta)$ in Theorem 2.3.2, the tightening of the tolerance level from ϵ to ϵ' is now replaced by the set inflation from A to the $(1/\nu^*)$-neighborhood of A given by $\{\xi : c(\xi, A) \leq 1/\nu^*\}$. Note that, regardless of the distance used, one could reduce the conservativeness of our analysis by focusing on A in the form $\{x \notin \mathcal{X}_\xi\}$, but this would require looking at the specific form of the safety set $\mathcal{X}_\xi$.

2.4 Improving Generating Distributions

This section discusses some approaches to search for better generating distributions beyond the baseline distribution in a divergence ball of DRO. Section 2.4.1 first states a general result to create better generating distributions. Section 2.4.2 then specializes to using a mixture distribution on θ to exploit this result. Sections 2.4.3 and 2.4.4 then provide two specific ways to construct these mixtures. Finally, Section 2.4.5 demonstrates some numerical comparisons in using these new mixing generating distributions and also simply using the baseline.

2.4.1 A Framework to Reduce Divergence Ball Size by Incorporating Parametric Information

The reason why the best choice of generating distribution $\mathbb{P}_0$ is not the baseline of the divergence ball, $\mathbb{P}_{\hat{\theta}}$, in minimizing $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ is that the equivalence relation (2.37) does not hold when $\mathbb{Q}$ is restricted to a parametric class. In some sense the reduction to the unambiguous chance constraint in the right hand side of (2.37) is over-conservative as it does not account for parametric information. Suppose we would still like to use the analytically tractable relation (2.37), but at the same time be less conservative. Then, one approach is to find a new baseline distribution, say $\tilde{\mathbb{P}}$, such that the parametrically restricted divergence ball

$\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$ lies inside a new nonparametric divergence ball at the center $\tilde{\mathbb{P}}$, namely $\{\mathbb{Q} : d_\phi(\tilde{\mathbb{P}}, \mathbb{Q}) \leq \tilde{\lambda}\}$. If we can obtain a nonparametric ball size $\tilde{\lambda}$ such that $\tilde{\lambda} < \lambda$ and the set inclusion holds, then this new ball is also a valid uncertainty set, and, when simply setting the generating distribution as $\mathbb{P}_0 = \tilde{\mathbb{P}}$ and applying Theorem 2.3.2, we have a smaller upper bound for $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ than $\epsilon'^{-1}(\delta, \lambda, \phi)$ obtained from using Theorem 2.3.2 directly with the parametric constraint relaxed.

To above mechanism can be executed as follows. Let $\mathcal{U}_{data} = \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$. For any $\mathbb{P}_0$, let

$$\mathcal{D}_{data}(\mathbb{P}_0, \phi) \triangleq \sup_{\mathbb{Q} \in \mathcal{U}_{data}} d_\phi(\mathbb{P}_0, \mathbb{Q}). \tag{2.50}$$

Then we clearly have

$$\mathcal{U}_{data} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_0, \mathbb{Q}) \leq \mathcal{D}_{data}(\mathbb{P}_0, \phi)\}, \tag{2.51}$$

since the right-hand-side set includes distributions outside of the parametric family as well.

Our goal is to find $\mathbb{P}_0$ to minimize $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ or any upper bound of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ so that it is smaller than the ball size λ appearing in the original parametric divergence ball $\mathcal{U}_{data}$. We state the implication of this as follows:

Theorem 2.4.1. *Suppose $\mathcal{Y} = \mathbb{R}^k$ and $\mathbb{P}_{\hat{\theta}}$ admits a density. Consider the parametric divergence ball $\mathcal{U}_{data} = \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}$. Suppose we can find $\mathbb{P}_0$ such that $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ defined in* (2.50) *satisfies $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$. Then we have*

$$\begin{aligned}\min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) &\leq \min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \mathcal{D}_{data}(\mathbb{P}_0, \phi)\}, \delta) \\ &\leq \min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}, \delta) \end{aligned} \tag{2.52}$$

and

$$\min_{\mathbb{P}_1} M(\mathbb{P}_1, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta) \leq \epsilon'^{-1}(\delta, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi) \leq \epsilon'^{-1}(\delta, \lambda, \phi) \tag{2.53}$$

where $\epsilon'^{-1}(\epsilon, \lambda, \phi)$ *is defined in* (2.39).

Proof. By the definition of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$, (2.51) holds. Together with the condition $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$, we have the set inclusions

$$\{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_0, \mathbb{Q}) \leq \mathcal{D}_{data}(\mathbb{P}_0, \phi)\} \subseteq \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\} \tag{2.54}$$

The inequalities (2.52) then follow from the definition of M. The inequalities (2.53) in turn follow immediately from Theorem 2.3.2. □

Theorem 2.4.1 stipulates that choosing $\mathbb{P}_0$ depicted in the theorem as the generating distribution, and setting $\epsilon'^{-1}(\delta, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi)$ as an upper bound for $M(\mathbb{P}_0, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda, \mathbb{Q} \in \mathcal{P}_{para}\}, \delta)$ to obtain the required Monte Carlo size $N_{exact}(\delta_\epsilon, \beta, d)$ implied by Corollary 2.2.1.1, will give a lighter Monte Carlo requirement than using the bound $\epsilon'^{-1}(\delta, \lambda, \phi)$ directly obtained by relaxing the parametric constraint and using $\mathbb{P}_{\hat{\theta}}$ as the generating distribution as in Corollary 2.3.2.1.

2.4.2 Mixture as Generating Distribution

Since optimization (2.50) can be difficult to solve generally, we focus on finding improved generating distribution $\mathbb{P}_0$ so that the implication of Theorem 2.4.1 holds, instead of fully optimizing (2.50). In this and the next subsections, we design a search space $\mathcal{P}_0$ for $\mathbb{P}_0$ that allows the construction of tractable procedures to achieve such improvements, while at the same time ensures the obtained $\mathbb{P}_0$ are amenable to Monte Carlo simulation.

From now on we will focus on χ^2-distance as our choice of ϕ for convenience (as will be seen). Suppose that $\mathbb{P}_\theta$ has density $p(y; \theta)$. We then set $\mathcal{P}_0$ to be the collection of distributions with densities in the form

$$p_0(y) = \int_\Theta p(y; \theta)\mu(d\theta), \tag{2.55}$$

for some probability measure μ on Θ. This class of distributions is easy to sample assuming $p(y; \theta)$

and μ are, as one can first sample $\theta \sim \mu(d\theta)$ and then $\xi \sim \mathbb{P}_\theta$ given θ.

Searching for the best $p_0(y)$ requires minimizing $\mathcal{D}_{data}(\mathbb{P}_0)$ over $\mathbb{P}_0 \in \mathcal{P}_0$ (where for convenience we denote $\mathcal{D}_{data}(\mathbb{P}_0)$ as $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ with ϕ representing the χ^2-distance). We first use (2.45) to write

$$\begin{aligned}\mathcal{D}_{data}(\mathbb{P}_0) &= \sup_{\theta \in \mathcal{U}_{data}} \int_{\mathcal{Y}} \left(\frac{p(y;\theta)}{p_0(y)} - 1\right)^2 p_0(y) dy \\ &= \sup_{\theta \in \mathcal{U}_{data}} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p_0(y)} dy - 1 \\ &= \sup_{\theta \in \mathcal{U}_{data}} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu(d\theta')} dy - 1. \end{aligned} \tag{2.56}$$

Denoting $\mathcal{P}(\Theta)$ as the space of probability measures on Θ, we define the function $L : \mathcal{P}(\Theta) \times \Theta \to \mathbb{R}$ to be

$$L(\mu, \theta) \triangleq \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu(d\theta')} dy, \tag{2.57}$$

assuming the integral is well-defined for $\mathcal{P}(\Theta) \times \Theta$ and further define

$$l(\mu) \triangleq \sup_{\theta \in \mathcal{U}_{data}} L(\mu, \theta). \tag{2.58}$$

Thus (2.56) can be written as $\mathcal{D}_{data}(\mathbb{P}_0) = l(\mu) - 1$, and minimizing $\mathcal{D}_{data}(\mathbb{P}_0)$ is equivalent to solving

$$\min_{\mu \in \mathcal{P}(\Theta)} l(\mu) = \min_{\mu \in \mathcal{P}(\Theta)} \max_{\theta \in \mathcal{U}_{data}} L(\mu, \theta). \tag{2.59}$$

Optimization (2.59) has the following convexity property:

Lemma 1. *The outer minimization in problem* (2.59) *is convex.*

Lemma 1 can be proved by direct verification, which is shown in Supplementary 2.10. Note also that, if μ is the point mass δ_θ for $\theta \in \Theta$, then the mixture distribution would recover the parametric distribution $\mathbb{P}_\theta$. Hence the proposed family $\mathcal{P}_0$ includes $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$, and in particular the original baseline distribution $\mathbb{P}_{\hat{\theta}}$. Although the outer minimization of (2.59) is a convex problem, computing $l(\mu)$ involves a non-convex optimization and is difficult in general. Our approach is

to search for a descent direction for the convex function $l(\cdot)$ from $\delta_{\hat{\theta}}$. In the following, we will study two types of search directions, each using its own version of Danskin's Theorem [74, 75]. To proceed, we introduce the following definition:

Definition 1. *Define $\Theta^*(\mu)$ to be the set of optimal points for the maximization problem in $l(\mu) = \sup_{\theta \in \mathcal{U}_{data}} L(\mu, \theta)$ given $\mu \in \mathcal{P}(\Theta)$:*

$$\Theta^*(\mu) = argmax_{\theta \in \mathcal{U}_{data}} L(\mu, \theta) \tag{2.60}$$

It can be shown that $\Theta^*(\mu)$ is non-empty and $\Theta^*(\mu) \subseteq \mathcal{U}_{data}$ because $\mathcal{U}_{data}$ is compact and $L(\mu, \theta)$ is continuous in θ.

2.4.3 Mixing with a Proposed Distribution

We consider mixing distributions in the form $(1-t)\delta_{\hat{\theta}} + t\mu_{prop}$ for some proposed distribution μ_{prop}, and look for a descent direction by varying t from 0 to 1. We have the following result that is a consequence of Danskin's Theorem that involves a one-sided derivative. We provide proofs both for this theorem and our following result in Supplementary 2.10.

Theorem 2.4.2. *Fix any $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$ and $\theta \in \Theta$. Under the assumptions that $\psi(t) = L((1-t)\mu_1 + t\mu_2, \theta)$ is well defined for $0 \leq t \leq 1$, we know that the function $g(y, t)$*

$$g(y,t) : \mathcal{Y} \times [0,1] \triangleq \frac{(p(y;\theta))^2}{(1-t)\int_\Theta p(y;\theta')\mu_1(d\theta') + t\int_\Theta p(y;\theta')\mu_2(d\theta')}$$

is integrable for $t \in [0,1]$. If we further assume that there exists a integrable function $g_0(y)$ such that

$$\left| \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\mu_1 - \mu_2)(d\theta')}{(\int_\Theta p(y;\theta')((1-t)\mu_1 + t\mu_2)(d\theta'))^2} \right| \leq g_0(y),$$

then we have the right derivative of $\psi(t)$ *at* $t = 0$ *given by*

$$\begin{aligned}\psi^+(0) &= \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \\ &= \sup_{\theta\in\Theta^*(\mu_1)} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\mu_1 - \mu_2)(d\theta')}{(\int_\Theta p(y;\theta')\mu_1(d\theta'))^2} dy. \end{aligned} \tag{2.61}$$

The quantity $\psi^+(0)$ is the directional derivative of $L(\mu_1)$ in the direction $\mu_2 - \mu_1$. Thus, to improve on $\mathcal{D}_{data}(\mathbb{P}_{\hat\theta})$, we can propose a mixing distribution $\mu_{prop}(d\theta')$, and substitute $\mu_1 = \delta_{\hat\theta}$ and $\mu_2 = \mu_{prop}$ in (2.61) to check if

$$\sup_{\theta\in\Theta^*(\delta_{\hat\theta})} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\delta_{\hat\theta} - \mu_{prop})(d\theta')}{p(y;\hat\theta)^2} dy < 0, \tag{2.62}$$

which indicates a strict descent for $l(\cdot)$ from $\delta_{\hat\theta}$ to μ_{prop}. In this case, it follows from the convexity of $l(\cdot)$ that we can find some $0 < t \leq 1$ such that $l((1-t)\delta_{\hat\theta} + t\mu_{prop}) < l(\delta_{\hat\theta})$, so that

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat\theta} + t\mu_{prop})(d\theta'), \tag{2.63}$$

gives rise to $\mathcal{D}_{data}(\mathbb{P}_0) < \mathcal{D}_{data}(\mathbb{P}_{\hat\theta})$. Finding such a t can be done by a bisection search or enumerating $\mathcal{D}_{data}(\mathbb{P}_0)$ on p_t over a grid of t. Note that the above can be implemented only if (2.62) can be verified and also if $\mathcal{D}_{data}(\mathbb{P}_0)$ is computable. We will show that both properties are satisfied for the case of multivariate Gaussian when μ_{prop} is properly chosen. In particular, we will identify general sufficient conditions for μ_{prop} to guarantee (2.62), and also find μ_{prop} such that the maximization involved in computing $\mathcal{D}_{data}(\mathbb{P}_0)$ in (2.56) can be reduced to a one-dimensional problem.

Consider a multivariate Gaussian distribution with unknown mean $\Theta \subset \mathbb{R}^D$ in an open convex set with density

$$p(y;\theta) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \cdot e^{-\frac{1}{2}(y-\theta)^\tau\Sigma^{-1}(y-\theta)}, \tag{2.64}$$

where Σ is a fixed positive semi-definite covariance matrix. Direct verification (in Supplementary

2.10) shows that

$$\mathcal{U}_{data} \triangleq \left\{\theta \in \Theta : \chi^2(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_\theta) \leq \frac{\chi^2_{1-\alpha,D}}{n}\right\} = \left\{\theta \in \Theta : e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} - 1 \leq \frac{\chi^2_{1-\alpha,D}}{n}\right\}$$
$$= \left\{\theta : \hat{\theta} + \Sigma^{\frac{1}{2}} v, \quad \text{for } \|v\|_2^2 \leq \log(1 + \frac{\chi^2_{1-\alpha,D}}{n})\right\}, \tag{2.65}$$

and thus

$$\Theta^*(\delta_{\hat{\theta}}) = \text{argmax}_{\theta \in \mathcal{U}_{data}} e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} = \left\{\theta : \hat{\theta} + \Sigma^{\frac{1}{2}} v, \quad \text{for } \|v\|_2^2 = \log(1 + \frac{\chi^2_{1-\alpha,D}}{n})\right\}. \tag{2.66}$$

We propose the following μ_{prop}. First, we call a distribution on Θ symmetrical around $\theta \in \Theta$ if its probability density or mass function has the same value for any $\theta_1, \theta_2 \in \Theta$ such that $\theta = \frac{\theta_1+\theta_2}{2}$.

Proposition 1. *Let $\mu_{prop}(d\theta')$ be any symmetrical distribution around $\hat{\theta}$. Given $\theta \in \Theta^*(\delta_{\hat{\theta}})$, we define $Y_\theta = (\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})$ with $\theta' \sim \mu_{prop}(d\theta')$. Suppose there exists an integrable random variable Y under the measure μ_{prop} such that $e^{2Y_\theta} \leq Y$ for all $\theta \in \Theta^*(\delta_{\hat{\theta}})$. If, for each $\theta \in \Theta^*(\delta_{\hat{\theta}})$, Y_θ does not equal to 0 with probability 1, then* (2.62) *holds and the mixture distribution produced by $\mu_{prop}(d\theta)$ would result in a descent direction on $\mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}})$.*

One can check that any Gaussian distribution with mean $\hat{\theta}$ satisfies the conditions of Proposition 1, and so does any $\mu_{prop}(d\theta')$ that is discrete, symmetrical around $\hat{\theta}$, whose outcome directions $\theta' - \theta$ constitute a basis of $\mathbb{R}^D$. Alternately, we also consider the following continuous μ_{prop}. We set $\theta' \sim \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$ where η is a random vector uniformly distributed on the surface of the D-dimension unit ball. Note that this θ' can be efficiently simulated by sampling D independent standard Gaussian random variables and scaling their norm to unit length to obtain η. While this μ_{prop} can be readily checked to satisfy the conditions in Proposition 1, we also provide an alternate proof on the validity of this μ_{prop} in achieving a descent direction in Lemma 5 in the Supplementary, as results proven therein provide important reference to calculations in numerical experiments regarding μ_{prop}.

Next, we discuss the computation of $\mathcal{D}_{data}(\mathbb{P}_0)$ for a given $\mathbb{P}_0$. First, we call a random variable Y on $\mathcal{Y} \subset \mathbb{R}^k$ rotationally invariant if $Y \overset{\mathcal{D}}{=} Q^\intercal Y$ for any rotational matrix $Q \in \mathbb{R}^{k\times k}$. Using this notion, the following shows how one can reduce the D-dimensional maximization problem in the definition of $\mathcal{D}_{data}(\mathbb{P}_0)$ into a one-dimensional problem.

Proposition 2. *Given a nominal distribution $Y \sim \mathbb{P}_0$ and a multivariate Gaussian family with known covariance Σ denoted $\mathbb{P}_\theta = \mathcal{N}(\theta, \Sigma)$. If the nominal distribution $Y \sim \mathbb{P}_0$ satisfies the condition that the random variable $Z = \Sigma^{-1/2}(Y - \hat{\theta})$ is rotationally invariant, then for any θ_1, θ_2 satisfying $(\theta_1 - \hat{\theta})^\intercal \Sigma^{-1}(\theta_1 - \hat{\theta}) = (\theta_2 - \hat{\theta})^\intercal \Sigma^{-1}(\theta_2 - \hat{\theta})$, we have*

$$\chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_1}) = \chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_2}). \tag{2.67}$$

Thus, for $\mathcal{D}_{data}(\mathbb{P}_0) = \max_{\theta\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{P}_\theta)$ with $\mathcal{U}_{data} = \{\theta \in \Theta : (\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta - \hat{\theta}) \le \lambda\}$ as in (6.79)*, we have*

$$\mathcal{D}_{data}(\mathbb{P}_0) = \max_{0\le t\le 1} \chi^2(\mathbb{P}_0, \mathbb{P}_{(1-t)\hat{\theta}+t\theta^\star}), \tag{2.68}$$

given any $\theta^\star$ satisfying $(\theta^\star - \hat{\theta})^\intercal \Sigma^{-1}(\theta^\star - \hat{\theta}) = \lambda$.

Proposition 3. *Given $0 \le t \le 1$ and $\mu_{prop}(d\theta) = \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$, where η is a random vector uniformly distributed on the surface of the D-dimension unit ball, the nominal measure $\mathbb{P}_t$ with density*

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat{\theta}} + t\mu_{prop})(d\theta') = (1-t)\mathbb{P}_{\hat{\theta}} + t\int_\Theta p(y;\theta')\mu_{prop}(d\theta'),$$

satisfies the conditions in Proposition 2.

Therefore, in computing $\mathcal{D}_{data}(\mathbb{P}_0)$ derived from the proposed distribution $\mu_{prop}(d\theta) = \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta$, using Propositions 2 and 3 we can change the domain of the involved maximization from $\Theta \subset \mathbb{R}^D$ into $\mathbb{R}$, leading to a substantial reduction in the search space and a tractable problem.

2.4.4 Enlarging Mixture Variability

Our next proposal is to consider a continuous mixing distribution $\mu_r(d\theta')$ on Θ where $r \geq 0$ controls the variability of the distribution, so that $r = 0$ corresponds to $\delta_{\hat{\theta}}$. Here, we can parametrize the density of the generating distribution as

$$p_r(y) = \int_{\Theta} p(y;\theta')\mu_r(d\theta'), \tag{2.69}$$

and our search direction is along r starting from $r = 0$. We propose two possible ways to define $\mu_r(d\theta')$. First is to let $\mu_r^1(d\theta')$ follow the distribution of $\theta' \sim \hat{\theta} + \Sigma^{\frac{1}{2}} \cdot \eta_{\sqrt{r}}$ where $\eta_{\sqrt{r}}$ is the uniform distribution inside the D-dimensional unit ball with radius $\sqrt{r}$. Second is to let $\mu_r^2(d\theta')$ follow $\mathcal{N}(\hat{\theta}, r\Sigma)$. The second approach in particular can be intuited as the posterior distribution of the parameter from a Bayesian perspective. In both cases, we notice that letting $r = 0$ would recover the original baseline distribution $p(y;\hat{\theta})$.

To analyze these schemes, we abuse notation slightly and now define $L : \mathbb{R}^+ \times \Theta \to \mathbb{R}$ to be

$$L(r,\theta) \triangleq \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p_r(y)} dy, \tag{2.70}$$

and

$$l(r) \triangleq \sup_{\theta \in \mathcal{U}_{data}} L(r,\theta). \tag{2.71}$$

We show that increasing r to positive values would produce a descent direction for $l(r)$ at $r = 0$, when the underlying distribution is Gaussian. Recall that in this case $\Theta^*(\delta_{\hat{\theta}})$ can be expressed by (2.66). As $l(r)$ is not necessarily convex in this situation, we use a generalized version of Danskin's Theorem [76] for non-convex problems to get the following result:

Theorem 2.4.3. *With $l(r)$ and $L(r,\theta)$ defined in* (2.70) *and* (2.71)*, and $p(y;\theta)$ multivariate Gaus-*

sian with mean θ and known positive definite covariance Σ, we have

$$l^+(0) = \lim_{r\downarrow 0} \frac{l(r) - l(0)}{r} = (1 + \frac{\chi^2_{1-\alpha,D}}{n}) \cdot \lim_{r\downarrow 0} \frac{1}{r}\Big(1 - \inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_r}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\Big) \tag{2.72}$$

The proof is in Supplementary 2.10. With Theorem 2.4.3, we can show that both μ_r^1 and μ_r^2 proposed above are valid choices to produce descent directions. Moreover, we can also show that they allow tractable computation of $\mathcal{D}_{data}(\mathbb{P}_0)$. These are depicted as follows.

Corollary 2.4.3.1. *Under the assumptions in Theorem 2.4.3, $l^+(0) < 0$ for both μ_r^1 and μ_r^2.*

Corollary 2.4.3.2. *Given $r \geq 0$ and μ_{prop} being $\mu_r^1(d\theta)$ or $\mu_r^2(d\theta)$, the nominal measure $\mathbb{P}_r$ with density given by* (2.69) *satisfies the conditions in Proposition 2.*

The proofs of Corollary 2.4.3.1 and Corollary 2.4.3.2 are in Supplementary 2.10.

2.4.5 Numerical Demonstrations

To confirm our findings in Section 2.4.3 and 2.4.4, we perform several numerical experiments. Consider $\mathbb{P}_\theta$ to be multivariate Gaussian $\mathcal{N}(\theta, I_D)$ with $k = D = 10$. We set $\epsilon = \alpha = 0.05$ while $\beta = 0.01$ and data size $n = 10$ or 5. Notice in this case, the dimension D is high but the available sample n is low and we would actually need $N_{exact} = 371$ data points to perform standard SO. Based on our discussion, we compare three choices of μ_{prop}:

- $\mu_1 = \delta_{\hat{\theta}}$, the point mass at $\hat{\theta}$.
- $\mu_2 \sim \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \eta$, where η is the uniform random vector on the surface of a D-dimension unit ball, discussed in Section 2.4.3.
- $\mu_3 \sim \mathcal{N}(\hat{\theta}, I_D/n)$, the Gaussian distribution with mean $\hat{\theta}$ and covariance matrix I_D/n, discussed in Section 2.4.4.

For μ_1, μ_2 and μ_3, the calculation of $\mathcal{D}_{data}(\mathbb{P}_0)$ is tractable. We leave the details in the Supplementary as remarks following Lemma 5 and summarize the results in Table 2.1 and 2.2. We use N to

denote the number of Monte Carlo samples needed. Moreover, we use both algorithms Extended SO and Extended FAST discussed in Section 5 for demonstration. As we can see, the decrease in N under a better sampling distribution can be considerable, down to less than a third compared to using the baseline in some cases. Mixing with a proposed uniform distribution (μ_2) appears to reduce N more than applying a Gaussian mixture (μ_3). As a side note, we also observe Extended FAST requires significantly less sample size than Extended SO in this example.

Table 2.1: Comparisons among choices of $\mathbb{P}_0$ for 10 dimensional multivariate Gaussian when $n = 5$.

	$\mathcal{D}_{data}(\mathbb{P}_0)$	δ_ϵ	N for Extended SO	N for Extended FAST
$\mu_1(\delta_{\hat{\theta}})$	37.9161	6.5766×10^{-5}	285601	70221
μ_2	11.0368	2.2454×10^{-4}	83649	20707
μ_3	14.7391	1.6850×10^{-4}	111465	27528

Table 2.2: Comparisons among choices of $\mathbb{P}_0$ for 10 dimensional multivariate Gaussian when $n = 10$.

	$\mathcal{D}_{data}(\mathbb{P}_0)$	δ_ϵ	N for Extended SO	N for Extended FAST
$\mu_1(\delta_{\hat{\theta}})$	5.2383	4.6857×10^{-4}	40081	10026
μ_2	3.3139	7.3298×10^{-4}	25621	6481
μ_3	3.7926	6.4275×10^{-4}	29219	7363

2.5 Procedural Description

This section presents our procedures to find solutions for CCP (2.1) using SO-based methods, when the direct use of data $\xi_1, ..., \xi_n$ from $\mathbb{P}$ is possibly insufficient to achieve feasibility with a given confidence. Algorithm 1, which we call "Extended SO", first presents the basic and most easily applicable procedure arising from Corollary 2.3.2.1. Notice that, given an overall target confidence level, say c, we have flexibility in choosing α and β such that $\alpha + \beta = c$. In our experiments, we simply choose $\alpha = \beta = \frac{c}{2}$. However, if the required confidence level is high, it is more beneficial to choose a relatively small β, since the required Monte Carlo sample size depends only logarithmically on β (i.e., the required sample size for SO is of order $\log \frac{1}{\beta}$) [22]. On the other

hand, as the confidence level $1-\alpha$ grows higher, the size of uncertainty set $\mathcal{U}_{data}$ would grow and cause the tolerance level ϵ for the SO (under the baseline $\mathbb{P}_0$) to decrease. Here, the dependence of Monte Carlo sample size on ϵ is less favorable, typically of order $\frac{1}{\epsilon}$ [22].

Algorithm 1 *Extended SO* to obtain a feasible solution $\hat{x}$ for (2.1) with asymptotic confidence $1-\alpha-\beta$

1: **Inputs:** data points $\xi_1, \ldots, \xi_n$, a ϕ-divergence, parametric information $\mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta\in\Theta\subset\mathbb{R}^D}$.

2: Find the MLE $\hat{\theta}$ from the data $\xi_1, \ldots, \xi_n$ for parameter θ.

3: Set $\lambda \leftarrow \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n}$ where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of a χ^2_D distribution.

4: Set $\delta_\epsilon \leftarrow \epsilon'(\epsilon, \lambda, \phi)$ where ϵ' is defined in (2.38).

5: Set $N \leftarrow N_{exact}(\delta_\epsilon, \beta, d)$ where N_{exact} is defined in (2.11).

6: Generate $\xi_1^{MC}, ..., \xi_N^{MC}$ from $\mathbb{P}_{\hat{\theta}}$ to construct (2.10) and obtain a solution $\hat{x}$.

There are several variants of Algorithm 1. First, we have discussed the use of plain SO and that the required sample size is (2.11), while on the other hand, as mentioned at the end of Section 2.2.2, we can use other variants of SO such as FAST that requires a smaller sample size for either the data or the Monte Carlo samples we generate. In the case of FAST, we would have $N_{exact}(\epsilon, \beta, d) = 20d + \frac{1}{\epsilon}\log\frac{1}{\beta}$, as suggested by [28]. Thus, a variant of Algorithm 1 is to replace N_{exact} with this latter quantity, and replace (2.10) with the FAST procedure in [28] for the last step of Algorithm 1 (we call this algorithm "Extended FAST" which will also be used in the next section).

The explicit expression for $\epsilon'(\epsilon, \lambda, \phi)$ for different ϕ, ϵ and λ can be found in [39]. For example, if we choose $\phi = (x-1)^2$ which corresponds to the χ^2-distance, then for $\epsilon < 1/2$, we have $\epsilon' = \max\{0, \epsilon - \frac{\sqrt{\lambda^2+4\lambda(\epsilon-\epsilon^2)}-(1-2\epsilon)\lambda}{2\lambda+2}\}$. We can also replace $\epsilon'(\epsilon, \lambda, \phi)$ by any δ_ϵ that achieves $M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta_\epsilon) \le \epsilon$. In Supplementary 2.9, we derive a self-contained easy upper bound for $M(\mathbb{P}_{\hat{\theta}}, \{\mathbb{Q} : d_\phi(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \le \lambda\}, \delta)$ in the case of χ^2-distance and use it to find such a δ_ϵ. This easy computation of δ_ϵ will also be used in our numerics in the next section.

Section 2.4.1 has investigated some proposals to improve the generating distributions. Algorithm 2 depicts these proposals in a general form. The main difference of Algorithm 2 compared to Algorithm 1 is the introduction of $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ that one can attempt to minimize over a class

of generating distribution $\mathbb{P}_0$ or evaluate for trial-and-error choices of $\mathbb{P}_0$, so that at the end we have $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \phi''(1)\chi^2_{1-\alpha,D}/(2n)$. As discussed in Section 2.4.1, using this $\mathbb{P}_0$ allows us to obtain a smaller Monte Carlo size requirement than simple relaxation of the parametric constraint. Sections 2.4.3 and 2.4.4 describe the possibilities of achieving such a reduction, in the case of Gaussian underlying distributions and using χ^2-distance. Note that, just like in Algorithm 1, we can consider other variants such as incorporating FAST and using alternate bounds for M instead of ϵ', by undertaking the same modifications as in Algorithm 1.

Algorithm 2 *Extended SO with improved generating distribution* to obtain a feasible solution $\hat{x}$ for (2.1) with asymptotic confidence $1 - \alpha - \beta$

1: **Inputs:** data points $\xi_1, \ldots, \xi_n$, a ϕ-divergence, parametric information $\mathcal{P}_{para} = \{\mathbb{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^D}$.
2: Find the MLE $\hat{\theta}$ from the data $\xi_1, \ldots, \xi_n$ for parameter θ.
3: Set $\lambda \leftarrow \frac{\phi''(1)\chi^2_{1-\alpha,D}}{2n}$ where $\chi^2_{1-\alpha,D}$ is the $1-\alpha$ quantile of a χ^2_D distribution.
4: Obtain $\mathbb{P}_0$ by minimizing $\mathcal{D}_{data}(\mathbb{P}_0, \phi)$ defined in (2.50) over a class of distributions or simple trial-and-error search so that $\mathcal{D}_{data}(\mathbb{P}_0, \phi) < \lambda$.
5: Set $\delta_\epsilon \leftarrow \epsilon'(\epsilon, \mathcal{D}_{data}(\mathbb{P}_0, \phi), \phi)$ where ϵ' is defined in (2.38).
6: Set $N \leftarrow N_{exact}(\delta_\epsilon, \beta, d)$ where N_{exact} is defined in (2.11).
7: Generate $\xi_1^{MC}, ..., \xi_N^{MC}$ from $\mathbb{P}_0$ to construct (2.10) and obtain a solution $\hat{x}$.

2.6 Numerical Experiments

This section presents some numerical examples to support our theoretical findings and illustrate the performance of our proposed procedures for data-driven CCPs. We focus on Algorithm 1 (Extended SO) and its FAST variant discussed in Section 2.5 (Extended FAST). We consider both single and joint CCPs (i.e., one and multiple inequalities respectively in the safety condition of the probability) as well as quadratic optimization problems. Moreover, we compare numerically with methods of robust optimization (RO) in [77, 78]. The experimental outputs that we report include:

- Under each setting, we repeat the experiment 1000 times with new data generated each time. For the solution $\hat{x}$ obtained in each trial from a given algorithm, we evaluate the

violation probability $V(\hat{x}, \mathbb{P})$ under the true probability measure $\mathbb{P}$ (under θ_{true}) either through exact calculation or Monte Carlo simulation with sample size 10000. Moreover, using the empirical distribution for the violation probabilities, we report $\hat{\epsilon}$ as the average violation probability $V(\hat{x}, \mathbb{P})$ as well as Q_{95}, the 95-percentile. Finally, we report and compare "f_{val}", the average objective value for the optimization problem across all 1000 runs.

- We fix $\alpha = 0.05$ and $\beta = 0.01$ across different values of ϵ and d. However, when we compare our methods with robust optimization approaches, we set $\alpha = 0.05$ and $\beta = 0.001$, since RO approaches essentially guarantee $\beta = 0$. On the other hand, the sample size chosen for FAST is taken with default values $N_1 = 20d$ in stage 1 and $N_2 = \frac{\log\beta - \log(B_\epsilon^{N_1,d})}{\log(1-\epsilon)}$ in stage 2 as discussed in [28].

- For given ϵ and d, we denote N_{exact} as the required sample size if we can directly sample from $\mathbb{P}$ and use standard SO. We denote n as the available data size ($n < N_{exact}$) and N as the Monte Carlo size needed for the our DRO-based methods. In DRO-based methods, we fix our generating distribution $\mathbb{P}_0$ as $\mathbb{P}_{\hat{\theta}}$ and use the χ^2-distance across the experiments.

2.6.1 Single Linear Chance Constraint Problem

We first consider a single linear CCP

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}((a+\xi)^T x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{2.73}$$

where $x \in \mathbb{R}^d$ is the decision variable, $a, c \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are fixed and $\xi \in \mathbb{R}^d$ is a random vector following some parametric distribution. We fix $a = [5, 5, ..., 5] \in \mathbb{R}^d$, $b = 5$ and $c = [-1, -1, ..., -1] \in \mathbb{R}^d$ and the problem would have a non-empty feasible region with high probability for ξ considered here. Moreover, a robustly feasible point for FAST [28] is chosen to be $\bar{x} = \mathbf{0} \in \mathbb{R}^d$ and an explicit $\mathcal{U}_{data}$ is constructed as (2.27) for our DRO.

Multivariate Gaussian

We conduct experiments when $\xi \sim \mathcal{N}(\theta, \Sigma)$ with fixed but a priori randomly generated positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and unknown $\theta \in \mathbb{R}^d$. Due to the normality of ξ, for any given θ, we can reformulate the chance constraint exactly as a second-order cone constraint, which can be robustified straightforwardly in the ambiguous chance constraint case. The underlying true parameter is taken to be $\theta_{true} = \mathbf{0} \in \mathbb{R}^d$ and the results are summarized in Table 2.3 and 2.4.

Table 2.3: Single linear CCP under Gaussian with unknown mean for different ϵ and d.

	$\epsilon = 0.1$ $d = 5$	$\epsilon = 0.1$ $d = 10$	$\epsilon = 0.1$ $d = 20$	$\epsilon = 0.05$ $d = 5$	$\epsilon = 0.05$ $d = 10$	$\epsilon = 0.05$ $d = 20$
n	50	80	200	50	80	200
N_{exact}	113	183	312	229	371	631
N	449	743	1016	1443	2349	3118
$\hat{\epsilon}$	0.0050	0.0041	0.0041	0.0015	0.0015	0.0014
Q_{95}	0.0136	0.0103	0.0088	0.0045	0.0037	0.0031
f_{val}	-0.7577	-0.7447	-0.7360	-0.7353	-0.7243	-0.7128

Table 2.4: Comparisons for single linear CCP under Gaussian: $\epsilon = 0.05$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	447	447
N	NA	2887	1079
$\hat{\epsilon}$	0.0180	0.0011	0.00069
Q_{95}	0.0272	0.0029	0.0019
f_{val}	-0.8008	-0.7212	-0.7093

Exponential Distribution

We conduct experiment when each coordinate ξ_i of $\xi \in \mathbb{R}^d$ independently follows exponential distribution with rate λ_i. Since ξ is no longer Gaussian and the domain of the moment generating

moment function for exponential distribution depends on $\lambda = (\lambda_1, \ldots, \lambda_d)$, for convenience we use RO constructed from a convex approximation using Chebyshev's inequality:

$$\mathbb{P}_\lambda\Big(\xi^T x - \sum_{i=1}^{d} \frac{x_i}{\lambda_i} > \epsilon^{-1/2}\sqrt{Var(\xi^T x)}\Big) \leq \epsilon$$

which, combined with $\mathcal{U}_{data}$ as in (2.27), reduces the ambiguous chance constraint into a robust conic quadratic constraint

$$\epsilon^{-1/2}\sqrt{\sum_{i=1}^{d}(\frac{x_i}{\lambda_i})^2} + a^T x + \epsilon^{-1/2}\sum_{i=1}^{d}\frac{x_i}{\lambda_i} - b \leq 0, \quad \forall \lambda : \sum_{i=1}^{d}(1 - \frac{\lambda_i}{\hat{\lambda}_i})^2 \leq \frac{\chi^2_{1-\alpha,d}}{n},$$

The above can be tractably reformulated as in Section 5 of [77] on problems in the form of 5(b), with $\Omega = (\min_i(\hat{\lambda}_i)(1 - \frac{\chi^2_{1-\alpha,d}}{n}))^{-1}$ where $\hat{\lambda}_i$ represents the MLE estimate of λ_i. Finally, the underlying true parameters are taken as $\lambda_i = 1, \forall i$, and results are summarized in Table 2.5.

Table 2.5: Comparisons for single linear CCP under Exponential: $\epsilon = 0.05$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	447	447
N	NA	2887	1079
$\hat{\epsilon}$	0.0045	0.0047	0.0016
Q_{95}	0.0094	0.0100	0.0050
f_{val}	-0.6978	-0.6981	-0.6701

From the results of the experiments, we can see the vast majority of solutions produced by three methods satisfy statistical feasibility. In fact, all methods are conservative with respect to the violation probability ϵ, although some are more conservative than the other. In particular, when ξ is Gaussian, RO takes advantage of an exact formulation to produce less conservative solution with lower objective value (closer to the optimal value). This can be seen in Table 4, where $\hat{\epsilon} = 0.018$ $f_{val} = -0.80$ for RO and $\hat{\epsilon} = 0.0011$ $f_{val} = -0.72$ only for Extended SO. When ξ is no longer Gaussian, RO appears to produce similar-quality solutions as Extended SO in terms

of feasibility or optimality. For example in Table 5, we have $\hat{\epsilon} = 0.0045$ $f_{val} = -0.6978$ for RO and $\hat{\epsilon} = 0.0047$ $f_{val} = -0.6981$ for Extended SO. Note that while the validity of RO depends crucially on the applicability and accuracy of convex approximation, the validity of Extended SO or Extended FAST is not restricted by the distributions of ξ, and they also do not require intensive, case-specific analysis as RO. In general, we observe consistent performances of our methods in both experiments.

2.6.2 Joint Linear Chance Constraint Problem

Next, we consider a joint chance-constrained linear problem:

$$\begin{aligned} \min_{x \in X \subseteq \mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}((A + \Xi)x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{2.74}$$

where $x \in \mathbb{R}^d$ is the decision variable, $A \in \mathbb{R}^{m \times d}, c \in \mathbb{R}^d$ and $b \in \mathbb{R}^m$ are fixed and $\Xi \in \mathbb{R}^{m \times d}$ is a random matrix following some parametric distribution. We set c, each row of A and b to be the same as in the single linear CCP. We treat $\Xi \in \mathbb{R}^{m \times d}$ as a matrix concatenated from a random vector $\xi \in \mathbb{R}^{md} \sim \mathcal{N}(\theta, \Sigma)$ with fixed but a priori randomly generated positive definite covariance matrix $\Sigma \in \mathbb{R}^{md \times md}$ and unknown $\theta \in \mathbb{R}^{md}$. To solve RO, we use Bonferroni's inequality as in [79] to first divide the violation probability ϵ uniformly across m individual chance constraints and then follow the procedure as in single linear CCP. The results are summarized in Table 2.6.

Table 2.6: Comparisons for Joint linear CCP under Gaussian: $\epsilon = 0.05$, $m = 3$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	80	80	80
N_{exact}	NA	291	291
N	NA	2388	1214
$\hat{\epsilon}$	0.0003	0.0012	0.0226
Q_{95}	0.0007	0.0033	0.0564
f_{val}	-0.6448	-0.6626	-0.6466

In this joint linear example, Extended FAST provides the least conservative solution in terms of the achieved tolerance level ($\hat{\epsilon} = 0.0226$, which is closer to 0.05, compared to 0.0003 in RO and 0.0012 in Extended SO), and Extended SO is the least conservative in terms of the objective value ($f_{val} = -0.6626$ compared to -0.6448 in RO and -0.6466 in Extended FAST). RO appears to be the most conservative in terms of both the achieved tolerance level and objective value. Note that this occurs even though the underlying randomness is Gaussian, which allows exact reformulation in the single chance constraint case. The conservative performance here is likely (and unsurprisingly) due to the crude Bonferroni's correction. Note that other alternatives to using Bonferroni, if one considers tractable reformulation, would be to use moment-based DRO where tractability can be achieved (e.g., [43]). However, it is unclear if using moment-based DRO would be more or less conservative than using Bonferroni correction along with exact reformulation for the individualized constraints, which could comprise an interesting comparison for a future study. Nonetheless, our Extended SO/FAST, being purely sampled-based, avoids the additional conservativeness coming from the Bonferroni correction. However, we note that a large number of Monte Carlo samples are required due to the large size of $\mathcal{U}_{data}$ in this high-dimensional problem.

2.6.3 Non-Linear Chance Constrained Problems

In this section, we conduct numerical experiments for non-linear CCP. We consider two examples. First is a quadratic objective with joint linear chance constraints, and second is a linear objective with a quadratic chance constraint, similar as the QM problem considered in [80].

Quadratic Objective with Joint Linear Chance Constraint

We adopt the same setup (thus the robust feasibility condition remains the same) as in (2.74) except we modify the objective with a quadratic term

$$\begin{aligned} \min_{x \in X \subseteq \mathbb{R}^d} \quad & \frac{1}{2} x^T H x + c^T x \\ \text{s.t.} \quad & \mathbb{P}((A + \Xi)x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{2.75}$$

for a fixed but a priori randomly generated positive definite matrix H. We use $\epsilon = 0.05$. Results are summarized in Table 2.7. As we can see, feasibility in terms of violation probability is satisfied by all methods, though RO suffers from higher conservativeness compared to Extended SO/FAST in terms of the objective value ($f_{val} = -0.48$ compared to -0.5547 and -0.5476 for Extended SO and FAST respectively). Like the previous example, this could be attributed to the Bonferroni correction used in the RO. Extended FAST gives the least conservative solution in terms of the tolerance level ($\hat{\epsilon} = 0.0096$), using only one third of the samples compared to Extended SO (3888 vs 1384). On the other hand, Extended SO gives the least conservative solution in terms of the objective value ($f_{val} = -0.5547$).

Table 2.7: Comparisons for quadratic objective with joint linear chance constraint under Gaussian: $\epsilon = 0.05$, $m = 5$, $d = 10$ and $\beta = 0.001$.

	RO	Extended SO	Extended FAST
n	200	200	200
N_{exact}	NA	447	447
N	NA	3888	1384
$\hat{\epsilon}$	0	0.0006	0.0096
Q_{95}	0	0.0017	0.0253
f_{val}	-0.4800	-0.5547	-0.5476

Linear Objective with Quadratic Chance Constraint

We consider the following setup:

$$\begin{aligned} \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \quad & c^T x \\ \text{s.t.} \quad & \mathbb{P}(x^T \Xi x + a^T x \leq b) \geq 1 - \epsilon, x \geq 0 \end{aligned} \tag{2.76}$$

We set $\Xi = \frac{1}{m} \sum_{i=1}^{m} \xi_i \xi_i^T$ and $\xi_i \in \mathbb{R}^d \sim \mathcal{N}(\theta, \Sigma)$ are i.i.d. with unknown θ. We set $\theta_{true} = 0 \in \mathbb{R}^d$ and consequently $m\Xi$ follows a Wishart distribution $\mathcal{W}(\Sigma, m)$ with m degrees of freedom and covariance matrix Σ under $\mathbb{P}$. We use $\epsilon = 0.05$. The RO formulation for this problem is not readily available while our sampling-based methods are still directly applicable. We thus focus on

evaluating the performance of Extended FAST under different hyper-parameters. The results are summarized in Table 2.8. As we can see, the high dimensions of the problem do not affect the sample size requirement of Extended FAST dramatically, as it increases moderately form $N = 154$ when $d = 5$ to $N = 334$ when $d = 10$ and to $N = 422$ when $d = 15$. Moreover, the average optimal value f_{val} is around -0.85 and feasibility is satisfied ($\hat{\epsilon}$ all within 0.05), showing the consistent effectiveness of our method.

Table 2.8: Linear objective with quadratic chance constraint for different ϵ, m and d.

	$\epsilon = 0.1, d = 5, m = 5$	$\epsilon = 0.05, d = 10, m = 10$	$\epsilon = 0.05, d = 15, m = 15$
n	80	200	300
N_{exact}	113	371	504
N	154	334	422
$\hat{\epsilon}$	0.0092	0.0050	0.0048
Q_{95}	0.0263	0.0133	0.0128
f_{val}	-0.8393	-0.8576	-0.8672

2.7 Conclusion

We consider data-driven chance constrained problems with limited data. In such situation, standard approaches in SO may not be able to generate statistically feasible solutions. We investigate an approach that uses divergence-based DRO to efficiently incorporate parametric information through a data-driven uncertainty set, and subsequently uses Monte Carlo sampling to generate enough samples to handle the distributionally robust chance constraint. In this way our framework translates the data size requirement in SO into a Monte Carlo requirement, the latter could be much more abundant thanks to cheap modern computational power.

To exploit the full capability of our framework, we have investigated the optimality of the generating distribution in drawing the Monte Carlo samples in the sense of minimizing its required sample size. We have shown that, while the optimal choice is the baseline distribution in the unambiguous and nonparametric DRO cases, this natural choice can be dominated by other distri-

butions in the parametric DRO case. We proved this by connecting the Neyman-Pearson lemma in statistical hypothesis testing to DRO and SO, which comprises the first such results of its kind as far as we know. We then studied several ways to find better generating distributions by searching for mixtures that enhance distributional variability. Lastly, we showed some numerical results to demonstrate how our approach can give rise to feasible solutions in a wide range of settings where other methods such as RO cannot be utilized directly or give more conservative solutions.

2.8 Supplementary A: Regularity Conditions to Verify Assumption 1

We consider the following conditions:

(C1) $p(x, \theta_1) = p(x, \theta_2)$ for all x implies $\theta_1 = \theta_2$.

(C2) θ_{true} is an inner point of $\Theta \subseteq \mathbb{R}^D$.

(C3) The support of distribution $\{x : p(x, \theta) > 0\}$ does not depend on θ.

(C4) There exists a measurable function $L_1(x)$ such that $\mathbb{E}_{\theta_{true}} L_1^2 < \infty$ and

$$|\log p(x, \theta_1) - \log p(x, \theta_2)| \leq L_1(x)\|\theta_1 - \theta_2\|_2 \tag{2.77}$$

for all θ_1, θ_2 in a neighborhood of θ_{true}.

(C5) $I(\theta_{true})$ is non-singular.

(C6) The density family $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is differentiable in quadratic mean at θ_{true}, i.e., there exists a measurable function $L_2(x) : \mathcal{X} \to R^D$ such that for any $h \in \mathbb{R}^D$ that converges to 0,

$$\int \left(\sqrt{p(x, \theta_{true} + h)} - \sqrt{p(x, \theta_{true})} - \frac{1}{2} h^T L_2(x) \sqrt{p(x, \theta_{true})}\right)^2 dx = o(\|h\|_2^2). \tag{2.78}$$

The consistency and asymptotic normality of MLE in Assumption 1 is guaranteed under conditions (C1)-(C6). See [69, 70].

2.9 Supplementary B: Alternate Bounds Using χ^2 Distance

Consider the χ^2-based uncertainty set $\mathcal{U}_{data} = \{\mathbb{Q} \in \mathcal{P}_{para} : \chi^2(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}) \leq \lambda\}$. Here we provide an alternate upper bound for the function $M(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$, which we call $\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$. That is, we find $\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta)$ that satisfies

$$\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) \leq \tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta), \quad \text{for all } A \text{ such that } \mathbb{P}_0(A) \leq \delta.$$

For any $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}_0$, we have

$$\begin{aligned}
\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) &= \mathbb{P}_0(\xi \in A) + \Big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \mathbb{Q}(\xi \in A) - \mathbb{P}_0(\xi \in A) \Big) \\
&= \mathbb{P}_0(\xi \in A) + \sup_{\mathbb{Q}\in\mathcal{U}_{data}} \int \mathbf{1}\{\xi \in A\} \big(\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1 \big) d\mathbb{P}_0(\xi) \\
&\leq \mathbb{P}_0(\xi \in A) + \sup_{\mathbb{Q}\in\mathcal{U}_{data}} \Big(\int \mathbf{1}\{\xi \in A\} d\mathbb{P}_0(\xi) \Big)^{1/2} \cdot \Big(\int \big(\frac{d\mathbb{Q}}{d\mathbb{P}_0} - 1 \big)^2 d\mathbb{P}_0(\xi) \Big)^{1/2} \\
&\leq \delta + \delta^{1/2} \cdot \big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{Q}) \big)^{1/2}, \qquad (2.79)
\end{aligned}$$

where the fourth line follows from the Cauchy-Schwarz inequality. Thus, we can set

$$\tilde{M}(\mathbb{P}_0, \mathcal{U}_{data}, \delta) = \delta + \delta^{1/2} \cdot \big(\sup_{\mathbb{Q}\in\mathcal{U}_{data}} \chi^2(\mathbb{P}_0, \mathbb{Q}) \big)^{1/2} = \delta + \delta^{1/2} \cdot \big(\mathcal{D}_{data}(\mathbb{P}_0) \big)^{1/2},$$

which is non-decreasing in δ. By (2.15), we can choose δ_ϵ such that $\delta_\epsilon + \delta_\epsilon^{1/2} \left(\mathcal{D}_{data}(\mathbb{P}_0) \right)^{1/2} \leq \epsilon$, or equivalently,

$$\delta_\epsilon \leq \epsilon + \frac{\mathcal{D}_{data}(\mathbb{P}_0)}{2} - \sqrt{\epsilon \cdot \mathcal{D}_{data}(\mathbb{P}_0) + \frac{1}{4} (\mathcal{D}_{data}(\mathbb{P}_0))^2}, \qquad (2.80)$$

by solving the quadratic equation. In the case where we relax the parametric constraint completely, we have $\mathcal{D}_{data}(\mathbb{P}_0) = \lambda$. Compared to the bound obtained from Theorem 2.3.2 and Corollary 2.3.2.1, (2.80) is less tight, but the gap can be shown to asymptotically vanish when $\epsilon, \frac{\chi^2_{1-\alpha,D}}{n} \to 0$.

2.10 Supplementary C: Proofs and Other Technical Results

Proof of Lemma 1. First, by definition $\mathcal{P}(\Theta)$ is a convex set and, for any $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$ and $0 < t < 1$, we have

$$(1-t)\mu_1 + t\mu_2 \in \mathcal{P}(\Theta).$$

Next, fixing $\theta \in \mathcal{U}_{data}$, the function $L(\cdot, \theta)$ is convex since:

$$\begin{aligned}
L((1-t)\mu_1 + t\mu_2, \theta) =& \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')((1-t)\mu_1 + t\mu_2)(d\theta')} dy \\
=& \int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{(1-t)\int_\Theta p(y;\theta')\mu_1(d\theta') + t\int_\Theta p(y;\theta')\mu_2(d\theta')} dy \\
\leq& (1-t)\int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu_1(d\theta')} dy + t\int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{\int_\Theta p(y;\theta')\mu_2(d\theta')} dy \\
=& (1-t)L(\mu_1, \theta) + tL(\mu_2, \theta)
\end{aligned}$$

for any $0 < t < 1$ where the inequality follows from the convexity of the function $1/x$ for $x > 0$. Thus, as the supremum of convex functions, $l(\mu) \triangleq \sup_{\theta \in \mathcal{U}_{data}} L(\mu, \theta)$ is also convex. □

We provide a version of Danskin' Theorem needed to prove Theorem 2.4.2. Alternately, one can also resort to a generalized version in [76] by verifying the conditions there. Here we opt for the former and provide a self-contained proof, which mostly relies on the techniques from Proposition 4.5.1 of [75] but with some slight modification to handle issues regarding the domain of the involved function. We have:

Lemma 2. *Fix probability measures $\mu_1, \mu_2 \in \mathcal{P}(\Theta)$. Suppose $t_k \downarrow 0$ is a positive sequence such that $(1-t_k)\mu_1 + t_k\mu_2 \in \mathcal{P}(\Theta)$ for all k and $\theta_k \in \Theta^*((1-t_k)\mu_1 + t_k\mu_2)$ is a sequence such that $\theta_k \to \theta_0$ for some $\theta_0 \in \mathcal{U}_{data}$, then we have*

$$\limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \leq \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t},$$

if we assume $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \leq t \leq 1$ and $\theta \in \Theta$.

Proof. It is known that if $f : \mathbb{I} \to \mathbb{R}$ is a convex function with $\mathbb{I}$ being an open interval containing some point x, we then have the following results [75]:

$$f^+(x) = \lim_{t \downarrow 0} \frac{f(x+t) - f(x)}{t} = \inf_{t>0} \frac{f(x+t) - f(x)}{t}, \tag{2.81}$$

$$f^-(x) = \lim_{t \downarrow 0} \frac{f(x) - f(x-t)}{t} = \sup_{t>0} \frac{f(x) - f(x-t)}{t}, \tag{2.82}$$

and

$$f^+(x) \geq f^-(x). \tag{2.83}$$

In other words, these limits exist and satisfy the above relations for convex functions. Thus, if we define $f_k(t) = L((1-t_k)\mu_1 + t_k\mu_2 + t(\mu_2 - \mu_1), \theta_k)$, it follows from the convexity of $\mathcal{P}(\Theta)$ and $L(\cdot, \theta_k)$ that $f_k(t)$ is convex and well-defined for $-t_k \leq t \leq 1 - t_k$. Using the above results in (2.81), (5.21) and (2.83), we then have

$$\begin{aligned} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} &= \frac{f_k(0) - f_k(-t_k)}{t_k} \\ &\leq \sup_{t>0} \frac{f_k(0) - f_k(-t)}{t} = f_k^-(0) \leq f_k^+(0) = \inf_{t>0} \frac{f_k(t) - f_k(0)}{t}. \end{aligned} \tag{2.84}$$

On the other hand, if we define $f_0(t) = L((1-t)\mu_1 + t\mu_2, \theta_0)$, it also follows that $f_0(t)$ is convex and well-defined for $0 \leq t \leq 1$. It follows from the convexity of $f_0(\cdot)$ as well as (2.81) that

$$\begin{aligned} \lim_{t \downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} &= \lim_{t \downarrow 0} \frac{f_0(t) - f_0(0)}{t} \\ &= \inf_{t>0} \frac{f_0(t) - f_0(0)}{t} = f_0^+(0). \end{aligned} \tag{2.85}$$

Then, it again follows from the convexity of $f_0(\cdot)$ that, given any $\tau > 0$, we can find some $\eta > 0$ such that

$$\frac{f_0(s) - f_0(0)}{s} \leq f_0^+(0) + \tau, \tag{2.86}$$

for all $0 < s < \eta$. It then follows from definitions and (2.86) that

$$\begin{aligned}\frac{L((1-s)\mu_1 + s\mu_2, \theta_0) - L(\mu_1, \theta_0)}{s} =& \frac{L((\mu_1 + s(\mu_2 - \mu_1), \theta_0) - L(\mu_1, \theta_0)}{s} \\ =& \frac{f_0(s) - f_0(0)}{s} \leq f_0^+(0) + \tau, \end{aligned} \tag{2.87}$$

for all $0 < s < \eta$. Fixing one such s, since the function $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \leq t \leq 1$ and $\theta \in \Theta$, and the sequence satisfies $\theta_k \to \theta_0$, we have

$$\begin{aligned}\lim_{k\to\infty} \frac{f_k(s) - f_k(0)}{s} =& \lim_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2 + s(\mu_2 - \mu_1), \theta_k) - L((1-t_k)\mu_1 + t_k\mu_2, \theta_k)}{s} \\ =& \frac{L((\mu_1 + s(\mu_2 - \mu_1), \theta_0) - L(\mu_1, \theta_0)}{s} = \frac{f_0(s) - f_0(0)}{s} \leq f_0^+(0) + 2\tau,\end{aligned}$$

as long as we make $\eta > s > 0$ small enough so that $\eta \leq 1 - t_k$ for all k. Then, for k large enough, we have

$$\inf_{t>0} \frac{f_k(t) - f_k(0)}{t} \leq \frac{f_k(s) - f_k(0)}{s} \leq f_0^+(0) + 2\tau. \tag{2.88}$$

Combining (2.84), (2.85) and (2.88), we have that, for k large enough,

$$\frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \leq \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} + 2\tau.$$

Finally, since τ is arbitrary, we conclude that

$$\limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \leq \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t}.$$

□

We now prove the following version of Danskins' Theorem:

Theorem 2.10.1. *Fix $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{U}_{data})$. Suppose $t_k \downarrow 0$ is a positive sequence such that $(1 - t_k)\mu_1 + t_k\mu_2 \in \mathcal{P}(\Theta)$ for all k and $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continuous in $0 \leq t \leq 1$ and $\theta \in \Theta$. Then, if we let $\psi(t) = l((1-t)\mu_1 + t\mu_2)$ for $0 \leq t \leq 1$ with $l(\cdot) = \sup_{\theta\in\mathcal{U}_{data}} L(\cdot, \theta)$ defined as*

(2.58)*, we have*

$$\psi^+(0) = \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \tag{2.89}$$

Proof. For any $\theta_0 \in \Theta^*(\mu_1)$ and $\theta_t \in \Theta^*((1-t)\mu_1 + t\mu_2)$, we have

$$\begin{aligned}\frac{\psi(t) - \psi(0)}{t} = \frac{l((1-t)\mu_1 + t\mu_2) - l(\mu_1)}{t} =& \frac{L((1-t)\mu_1 + t\mu_2, \theta_t) - \tilde{L}(\mu_1, \theta_0)}{t} \\ \geq& \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t}.\end{aligned}$$

Thus, by taking $t \downarrow 0$ and taking the supremum over all $\theta_0 \in \Theta^*(\mu_1)$, we have

$$\psi^+(0) \geq \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t}. \tag{2.90}$$

Notice that the existence of the several limits above follows from the convexity of related functions. To prove the reverse inequality, we consider a sequence $\{t_k\}$ with $0 < t_k < 1$ and $t_k \downarrow 0$. Then, we pick another sequence $\{\theta_k\} \subseteq \mathcal{U}_{data}$ with $\theta_k \in \Theta^*((1-t_k)\mu_1 + t_k\mu_2)$ for all k. Since $\mathcal{U}_{data}$ is compact, there exist a subsequence of $\{\theta_k\}$ converge to some $\theta_0 \in \mathcal{U}_{data}$. Without loss of generality, we drop the subsequence and simply assume $\theta_k \to \theta_0$. We first show $\theta_0 \in \Theta^*(\mu_1)$. To do this, pick any $\tilde{\theta}_0 \in \Theta^*(\mu_1)$. Since $L((1-t)\mu_1 + t\mu_2, \theta)$ is jointly continous in t and θ, we have

$$L(\mu_1, \theta_0) = \lim_{k\to\infty} L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) \geq \lim_{k\to\infty} L((1-t_k)\mu_1 + t_k\mu_2, \tilde{\theta}_0) = L(\mu_1, \tilde{\theta}_0),$$

where the inequality follows from the definition of θ_k. Now, since $\tilde{\theta}_0 \in \Theta^*(\mu_1)$ and $L(\mu_1, \theta_0) \geq L(\mu_1, \tilde{\theta}_0)$, we must have

$$L(\mu_1, \theta_0) = L(\mu_1, \tilde{\theta}_0) \text{ and } \theta_0 \in \Theta^*(\mu_1).$$

Now, using the definition of $\Theta^*(\mu_1)$, we can write

$$\begin{aligned}\psi^+(0) = \inf_{0<t} \frac{\psi(t)-\psi(0)}{t} \leq \frac{\psi(t_k)-\psi(0)}{t_k} =& \frac{l((1-t_k)\mu_1 + t_k\mu_2) - l(\mu_1)}{t_k} \\ =& \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_0)}{t_k} \\ \leq& \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k}. \end{aligned} \tag{2.91}$$

Now we use Lemma 2 to conclude that

$$\begin{aligned}\psi^+(0) &\leq \limsup_{k\to\infty} \frac{L((1-t_k)\mu_1 + t_k\mu_2, \theta_k) - L(\mu_1, \theta_k)}{t_k} \\ &\leq \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta_0) - L(\mu_1, \theta_0)}{t} \\ &\leq \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t} \end{aligned} \tag{2.92}$$

Finally, we combine (2.90) and (2.92) to conclude the proof

$$\psi^+(0) = \sup_{\theta\in\Theta^*(\mu_1)} \lim_{t\downarrow 0} \frac{L((1-t)\mu_1 + t\mu_2, \theta) - L(\mu_1, \theta)}{t}$$

□

Proof of Theorem 2.4.2. The result can be obtained from Leibniz's integral rule (i.e. differentiation under the integral sign). See, for example, Theorem 2.27 in [81]. □

Next we prove Proposition 1. For convenience, we note that (2.64) can be written in a compact form for exponential family [72]:

$$p(y;\theta) = e^{\langle t(y),\theta\rangle - F(\theta) + k(y)}, \tag{2.93}$$

where $\langle a, b\rangle = a^\top b$ represents the usual inner product in the Euclidean space, and $t(\cdot), F(\cdot)$ and

$k(\cdot)$ are known functions. In particular, we have

$$F(\theta) = \frac{\theta^\intercal \Sigma^{-1} \theta}{2} \tag{2.94}$$

To facilitate the calculation, we first introduce two lemmas involving the exponential parametric family based on [72].

Lemma 3. *Pick $\theta_1, \theta_2 \in \Theta$. If $2\theta_2 - \theta_1 \in \Theta$, then we have*

$$\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)} dy = e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))}.$$

In particular, if $F(\theta) = \frac{\theta^\intercal \Sigma^{-1}\theta}{2}$, then $\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)} dy = e^{(\theta_2-\theta_1)^\intercal \Sigma^{-1}(\theta_2-\theta_1)}$.

Proof. It follows from (2.93) that

$$\begin{aligned}
\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2}{p(y;\theta_1)} dy =& e^{<t(y),2\theta_2-\theta_1>-(2F(\theta_2)-F(\theta_1))+k(y)} dy \\
=& e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))} \cdot \int_{\mathcal{Y}} p(y; 2\theta_2 - \theta_1) dy \\
=& e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))}.
\end{aligned}$$

□

Lemma 4. *Pick θ_1,θ_2 and $\theta_3 \in \Theta$. If $2\theta_2 - 2\theta_1 + \theta_3 \in \Theta$, then we have*

$$\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2 p(y;\theta_3)}{(p(y;\theta_1))^2} dy = e^{F(2\theta_2-2\theta_1+\theta_3)-2F(\theta_2)+2F(\theta_1)-F(\theta_3)}.$$

In particular, if $F(\theta) = \frac{\theta^\intercal \Sigma^{-1}\theta}{2}$, then $\int_{\mathcal{Y}} \frac{(p(y;\theta_2))^2 p(y;\theta_3)}{(p(y;\theta_1))^2} dy = e^{(\theta_2-\theta_1)^\intercal \Sigma^{-1}(\theta_2-\theta_1)+2(\theta_2-\theta_1)\Sigma^{-1}(\theta_3-\theta_1)}$.

Proof. The proof follows from the same techniques as in Lemma 3. □

Then (6.79) follows from (2.26), (2.94) and Lemma 3 so that

$$\begin{aligned}\mathcal{U}_{data} \triangleq \left\{\theta \in \Theta : \chi^2(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_\theta) \leq \frac{\chi^2_{1-\alpha,D}}{n}\right\} &= \left\{\theta \in \Theta : e^{F(2\theta-\hat{\theta})-(2F(\theta)-F(\hat{\theta}))} - 1 \leq \frac{\chi^2_{1-\alpha,D}}{n}\right\} \\ &= \left\{\theta \in \Theta : e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} - 1 \leq \frac{\chi^2_{1-\alpha,D}}{n}\right\} \\ &= \left\{\theta : \hat{\theta} + \Sigma^{\frac{1}{2}} v, \quad \text{for all } \|v\|_2^2 \leq \log(1 + \frac{\chi^2_{1-\alpha,D}}{n})\right\},\end{aligned}$$

and (2.66) follows. We now prove Proposition 1:

Proof of Proposition 1. Following Theorem 2.4.2, Lemma 3 and Lemma 4, we have

$$\begin{aligned}&\sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\delta_{\hat{\theta}} - \mu_{prop})(d\theta')}{(\int_\Theta p(y;\theta')\delta_{\hat{\theta}}(d\theta'))^2} dy \\ =& \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \left(\int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p(y;\hat{\theta})} dy - \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot \int_\Theta p(y;\theta')(\mu_{prop})(d\theta')}{(p(y;\hat{\theta}))^2} dy \right) \\ =& \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \left(\int_{\mathcal{Y}} \frac{(p(y;\theta))^2}{p(y;\hat{\theta})} dy - \int_\Theta \int_{\mathcal{Y}} \frac{(p(y;\theta))^2 \cdot p(y;\theta')}{(p(y;\hat{\theta}))^2} dy \cdot \mu_{prop}(d\theta') \right) \\ =& \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \left(e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta})} - \int_\Theta e^{(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta-\hat{\theta}) + 2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})} \mu_{prop}(d\theta') \right) \\ =&(1 + \frac{\chi^2_{1-\alpha,D}}{n}) \cdot \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \left(1 - \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\right) \\ =&(1 + \frac{\chi^2_{1-\alpha,D}}{n}) \cdot \left(1 - \inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\right). \end{aligned} \tag{2.95}$$

Notice the second equality follows from Fubini's theorem. The third equality follows from Lemma 3 and Lemma 4. The fourth equality follows from (2.66). Now, following the last line (2.95), for the search of descent direction, it is sufficient to prove

$$\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta'\sim\mu_{prop}}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}] > 1.$$

However, since $\mu_{prop}(d\theta')$ is a symmetrical distribution around $\hat{\theta}$, we know that

$$\mathbb{E}_{\theta' \sim \mu_{prop}}[2(\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})] = 0.$$

for any $\theta \in \Theta^*(\delta_{\hat{\theta}})$. Then, it follows from Jensen's inequality that

$$\inf_{\theta \in \Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] \geq 1.$$

Now suppose for the sake of contradiction that

$$\inf_{\theta \in \Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] = 1.$$

Then, let $\{\theta_k\}_k \subseteq \Theta^*(\delta_{\hat{\theta}})$ be a subsequence such that $\mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta_k - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] \to 1$. Due to the compactness of $\Theta^*(\delta_{\hat{\theta}})$, we can find a subsequence of $\{\theta_k\}_k$ converging to some $\theta_0 \in \Theta^*(\delta_{\hat{\theta}})$. For convenience we drop the subsequence and suppose $\theta_k \to \theta_0$. Then the existence of Y allows us to use dominated convergence theorem:

$$\mathbb{E}[e^{2Y_{\theta_0}}] = \mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta_0 - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] = \lim_{k \to \infty} \mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta_k - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] = 1.$$

However, Jensen's inequality would indicate that $\mathbb{E}[e^{2Y_{\theta_0}}] = 1$ if and only $\mathbb{P}(Y_{\theta_0} = 0) = 1$, which contradicts our assumption. Thus, we know that

$$\inf_{\theta \in \Theta^*(\delta_{\hat{\theta}})} \mathbb{E}_{\theta' \sim \mu_{prop}}[e^{2(\theta - \hat{\theta})^\intercal \Sigma^{-1}(\theta' - \hat{\theta})}] > 1,$$

as claimed. □

Proof of Proposition 2. First we prove (2.67). Letting $c = \frac{1}{(2\pi)^D|\Sigma|}$, we know that

$$\begin{aligned}
\chi^2(\mathbb{P}_0, \mathbb{P}_\theta) =& \int \frac{p^2(y;\theta)}{p_0(y)} dy - 1 \\
=& c \int \frac{e^{-(y-\theta)^T\Sigma^{-1}(y-\theta)}}{p_0(y)} dy - 1 \\
=& c e^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2} \int \frac{e^{-(y-\hat{\theta})^T\Sigma^{-1}(y-\hat{\theta})} \cdot e^{-2(y-\hat{\theta})^T\Sigma^{-1}(\hat{\theta}-\theta)}}{p_0(y)} dy - 1 \\
=& c|\Sigma^{1/2}| e^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2} \int \frac{e^{-z^Tz} \cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta)}}{p_0(\Sigma^{1/2}z+\hat{\theta})} dz - 1 \\
=& c|\Sigma| e^{-\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2^2} \int \frac{e^{-z^Tz} \cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta)}}{p_Z(z)} dz - 1
\end{aligned} \tag{2.96}$$

where we denote $p_Z(\cdot)$ to be the density function of random variable $Z = \Sigma^{-1/2}(Y - \hat{\theta})$ with $Y \sim \mathbb{P}_0$ and the last two lines follow from a change of variable $z = \Sigma^{-1/2}(y - \hat{\theta})$. Now, since $\|\Sigma^{-1/2}(\theta_1 - \hat{\theta})\|_2^2 = \|\Sigma^{-1/2}(\theta_2 - \hat{\theta})\|_2^2 = r$ for some r by assumption, it follows from (2.96) that $\chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_1}) = \chi^2(\mathbb{P}_0, \mathbb{P}_{\theta_2})$ if we can show

$$\int \frac{e^{-z^Tz} \cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta_1)}}{p_Z(z)} dz = \int \frac{e^{-z^Tz} \cdot e^{-2z^T\Sigma^{-1/2}(\hat{\theta}-\theta_2)}}{p_Z(z)} dz.$$

However, since $p_Z(z)$ and e^{-z^Tz} are both rotationally invariant functions (i.e. $f(z) = f(Q^\intercal z)$ for all z and rotational matrix Q, with $|Q| = 1$), it can be shown that $\int \frac{e^{-z^Tz} \cdot e^{-2z^T\nu}}{p_Z(z)} dz$ holds the same value for any ν such that $\|\nu\|_2^2 = r$. Notice the rotational invariance of $p_Z(z)$ follows from the rotational invariance of Z. This proves (2.67). To prove (2.68), notice that for any $\theta \in \mathcal{U}_{data}$, we can find some $0 \le t \le 1$ such that

$$(((1-t)\hat{\theta} + t\theta^\star) - \hat{\theta})^\intercal \Sigma^{-1} (((1-t)\hat{\theta} + t\theta^\star) - \hat{\theta}) = (\theta - \hat{\theta})^\intercal \Sigma^{-1} (\theta - \hat{\theta})$$

and hence $\chi^2(\mathbb{P}_0, \mathbb{P}_{((1-t)\hat{\theta}+t\theta^*)}) = \chi^2(\mathbb{P}_0, \mathbb{P}_\theta)$ by (2.67). □

Proof of Proposition 3. To check that $Y \sim \mathbb{P}_t$ with density

$$p_t(y) = \int_\Theta p(y;\theta')((1-t)\delta_{\hat{\theta}} + t\mu_{prop})(d\theta') = (1-t)\mathbb{P}_{\hat{\theta}} + \int_\Theta p(y;\theta')\mu_{prop}(d\theta'),$$

leads to rotationally invariant $Z = \Sigma^{-1/2}(Y - \hat{\theta})$, simply notice that

$$Y \overset{\mathcal{D}}{=} (1-U_t)(\hat{\theta} + X_1) + U_t(\hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2}\eta + X_2),$$

where U_t is an independent Bernoulli variable with success rate t, η is a random vector uniformly distributed on the surface of the D-dimensional unit ball and X_1, X_2 are independent $\mathcal{N}(0,\Sigma)$. Then, it follows that

$$\Sigma^{-1/2}(Y - \hat{\theta}) \overset{\mathcal{D}}{=} (1-U_t)Z_1 + U_t(\sqrt{\frac{\chi^2_{1-\alpha,D}}{n}}\eta + Z_2)$$

where Z_1, Z_2 are now independent $\mathcal{N}(0, I_D)$. Consequently, the rotational invariance of Z now follows from the rotational invariance of Z_1, Z_2, η and their independence. □

Following the comments after Proposition 1, we show that $\theta \sim \mu_{prop}$ with $\theta \overset{\mathcal{D}}{=} \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2} \cdot \eta$ provides a descent direction, with an alternate proof using the following lemma and the last line of (2.95).

Lemma 5. *Fixing $\theta_1 \in \Theta^*(\delta_{\hat{\theta}})$, we have*

$$\mathbb{E}_{\theta_2 \sim \mu_{prop}}[e^{2(\theta_1 - \hat{\theta})^T \Sigma^{-1}(\theta_2 - \hat{\theta})}] > 1,$$

for $\theta_2 \sim \mu_{prop}(d\theta)$ where $\theta_2 \overset{\mathcal{D}}{=} \hat{\theta} + \sqrt{\frac{\chi^2_{1-\alpha,D}}{n}} \cdot \Sigma^{1/2} \cdot \eta$ with η following the uniform distribution on the surface of the D-dimensional unit ball.

Proof of Lemma 5. Let $u_1 \in \mathbb{R}^D$ denote an arbitary point on the surface of D dimensional unit ball ($\|u_1\|_2^2 = 1$) and let $\eta = [\eta_1, \eta_2, ..., \eta_D]$ be the random vector in $\mathbb{R}^D$ uniformly distributed on the

surface of D dimensional unit ball. Then we claim that $\frac{\mu_1^T\eta+1}{2} \sim Beta(\frac{D-1}{2}, \frac{D-1}{2})$.

To show this, assume without loss of generality that $u_1 = [1, 0, ..., 0] \in \mathbb{R}^D$. Then for any $t \in [-1, 1]$, it follows that $\mathbb{P}(u_1^T\eta \in dt)$ is proportional to the infinitesimal surface area on the ball corresponding to $\eta_1 \in dt$, which is in turn proportional to the product of the sub-dimension $D-2$ surface area on the belt $x_2^2+x_3^2+...+x_D^2 = 1-t^2$ with the infinitesimal width of this belt. Specifically, the sub-dimension $D-2$ surface area around the belt is proportional to $(\sqrt{1-t^2})^{D-2}$. This follows from the fact that points of the form $[0, \sqrt{1-t^2}, 0, ..., 0]$, $[0, 0, \sqrt{1-t^2}, 0, ..., 0]$, ..., $[0, 0, ..., 0, \sqrt{1-t^2}]$ are on this belt. Also, the width of this belt, according to the Pythagorean theorem, is $dt \cdot \sqrt{(\frac{d\sqrt{1-t^2}}{dt})^2 + 1} = \frac{dt}{\sqrt{1-t^2}}$. Thus,

$$\mathbb{P}(u_1^T\eta \in dt) \propto \frac{(\sqrt{1-t^2})^{D-2}}{\sqrt{1-t^2}}dt = (1-t^2)^{\frac{D-3}{2}}dt.$$

Now, we can substitute $\frac{t+1}{2} = s$ with $s \in [0, 1]$ to get

$$\mathbb{P}(\frac{u_1^T\eta+1}{2} \in ds) \propto (s)^{\frac{D-1}{2}-1}(1-s)^{\frac{D-1}{2}-1}ds,$$

which can only be the density function for $Beta(\frac{D-1}{2}, \frac{D-1}{2})$. It now follows from [82] that $\frac{u_1^T\eta+1}{2}$ has moment generating function

$$\begin{aligned} M(t) \triangleq& \mathbb{E}[e^{t\cdot\frac{u_1^T\eta+1}{2}}] \\ =&{}_1F_1(\frac{D-1}{2}, D-1, t) = e^{(t/2)}{}_0F_1(;\frac{D}{2}, \frac{t^2}{16}) \geq e^{t/2}(1+ct^2) > e^{(t/2)}. \end{aligned} \tag{2.97}$$

for some $c > 0$ where ${}_1F_1(\cdot,\cdot,\cdot)$ and ${}_0F_1(;,\cdot,\cdot)$ are the confluent hypergeometric function with identity ${}_1F_1(a, 2a, x) = e_0^{x/2}F_1(; a+1/2, x^2/16)$ (see [83]),

$${}_0F_1(;\alpha,t) \triangleq \sum_{k=0}^{\infty}\frac{t^k}{(\alpha)_k k!} \quad \text{and} \quad {}_1F_1(\alpha,\beta,t) \triangleq \sum_{k=0}^{\infty}\frac{(\alpha)_k t^k}{(\beta)_k k!},$$

with $(\gamma)_k = \frac{\Gamma(\gamma+k)}{\Gamma(\gamma)}$ being the Pochhammer symbol [82]. To conclude the proof, denote $\rho_n =$

$\sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}\cdot\sqrt{\frac{\chi^2_{1-\alpha,D}}{n}}$ and use (6.79), (2.66) and (2.97) to write

$$\begin{aligned}&\mathbb{E}_{\theta_2\sim\mu_{prop}}[e^{2(\theta_1-\hat{\theta})^T\Sigma^{-1}(\theta_2-\hat{\theta})}]\\ =&\mathbb{E}_{\nu\sim\eta}[e^{2\rho_n\cdot\mu_1^T\nu}]=\mathbb{E}_{X\sim Beta(\frac{D-1}{2},\frac{D-1}{2})}[e^{2\rho_n\cdot(2X-1)}]=M(4\rho_n)/e^{2\rho_n}\geq(1+16c\rho_n^2)>1.\end{aligned}$$

□

Remark 1. *Following Lemma 5, we discuss the numerical calculations of $\mathcal{D}(\mathbb{P}_0)$ following Proposition 3. We use $\mathcal{U}_{data}=\{\mathbb{P}_\theta:\|\theta-\hat{\theta}\|_2^2\leq\frac{\chi^2_{1-\alpha,D}}{n}\}$ where $p(y;\theta)=(2\pi)^{-\frac{D}{2}}e^{-\frac{1}{2}\|(y-\theta)\|_2^2}$. Then, for μ_1, the nominal $p_0(y)$ is simply $p(y;\hat{\theta})$ and $\mathcal{D}_{data}(\mathbb{P}_0)=\mathcal{D}_{data}(\mathbb{P}_{\hat{\theta}})=\max_{\theta\in\mathcal{U}}e^{\|\theta-\hat{\theta}\|_2^2}-1=e^{\frac{\chi^2_{1-\alpha,D}}{n}}-1$ according to* (6.79) *and Lemma 3. For μ_2, it can be shown that the nominal $\mathbb{P}_0$ follows $\mathcal{N}(\hat{\theta},(1+\frac{1}{n})\cdot I_D)$, and a direct computation would show that $\mathcal{D}_{data}(\mathbb{P}_0)=\max_{\theta\in\mathcal{U}}(\frac{(n+1)^2}{n(n+2)})^{\frac{d}{2}}e^{\frac{n}{n+2}\|\theta-\hat{\theta}\|_2^2}-1=(\frac{(n+1)^2}{n(n+2)})^{\frac{d}{2}}e^{\frac{n}{n+2}\frac{\chi^2_{1-\alpha,D}}{n}}-1$. Finally, for μ_3, assume w.l.o.g that $\hat{\theta}=0$. Then we use the derivation in Lemma 5 that $\frac{\mu_1^T\eta+1}{2}\sim Beta(\frac{D-1}{2},\frac{D-1}{2})$ for any u_1 on the D-dimensional unit ball surface to show that, for any $v\in\mathbb{R}^D$,*

$$\mathbb{E}_\eta[e^{\eta^T v}]=e^{-\|v\|_2}{}_1F_1(\frac{D-1}{2},D-1,2\|v\|_2),\tag{2.98}$$

and consequently

$$p_0(y)=(2\pi)_1^{-\frac{D}{2}}F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{1/2}\|y\|_2)e^{-\frac{1}{2}(\|y\|_2+\frac{\chi^2_{1-\alpha,D}}{n})^2}.$$

Then, to calculate $\mathcal{D}_{data}(\mathbb{P}_0)$, we note that

$$\begin{aligned}
\mathcal{D}_{data}(\mathbb{P}_0)+1 &= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \int \frac{p^2(y;\theta)}{p_0(y)}dy \\
&= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} (2\pi)^{-\frac{D}{2}} e^{-\frac{1}{2}\|y\|_2^2} \frac{e^{-\|\theta\|_2^2+2\theta^T y+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|y\|_2}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|y\|_2)} \\
&= \max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_{Y\sim\mathcal{N}(0,I_D)}\left[\frac{e^{-\|\theta\|_2^2+2\theta^T Y+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|Y\|_2}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}\|Y\|_2)}\right]. \qquad (2.99)
\end{aligned}$$

Furthermore, through either direct verification or analysis similar to those in Lemma 5, we note that $Y\sim\mathcal{N}(0,I_D)$ shares the same distribution of $L\eta$ where $L\in\mathbb{R}^+$ and $\eta\in\mathbb{R}^D$ are two independent random variables with L being the norm of $\mathcal{N}(0,I_D)$ bearing density $f_L(l)=1_{\{l\ge0\}}\frac{2^{1-\frac{D}{2}}}{\Gamma(\frac{D}{2})}l^{d-1}e^{-\frac{l^2}{2}}$ and η being the random vector on the D-dimensional unit ball surface. Thus, it follows from (2.99) *that* (2.99) *equals*

$$\begin{aligned}
&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[\mathbb{E}_\eta\left[\frac{e^{-\|\theta\|_2^2+2L\theta^T\eta+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L}}{{}_1F_1(\frac{d-1}{2},d-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L)}\middle|L\right]\right] \\
=&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[e^{-\|\theta\|_2^2+\frac{\chi^2_{1-\alpha,D}}{2n}+(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L-2L\|\theta\|_2}\frac{{}_1F_1(\frac{D-1}{2},D-1,4L\|\theta\|_2)}{{}_1F_1(\frac{D-1}{2},D-1,2(\frac{\chi^2_{1-\alpha,D}}{n})^{\frac{1}{2}}L)}\right] \\
=&\max_{\|\theta\|_2^2\le\frac{\chi^2_{1-\alpha,D}}{n}} \mathbb{E}_L\left[e^{-\|\theta\|_2^2+\frac{\chi^2_{1-\alpha,D}}{2n}}\frac{{}_0F_1(;\frac{D}{2},L^2\|\theta\|_2^2)}{{}_0F_1(;\frac{D}{2},L^2(\frac{\chi^2_{1-\alpha,D}}{4n})}\right] \\
=&\max_{t\le\frac{\chi^2_{1-\alpha,D}}{n}} e^{-t+\frac{\chi^2_{1-\alpha,D}}{2n}}\int_{l\ge0}\frac{{}_0F_1(;\frac{D}{2},l^2t)}{{}_0F_1(;\frac{D}{2},l^2(\frac{\chi^2_{1-\alpha,D}}{4n})}\frac{2^{1-\frac{D}{2}}}{\Gamma(\frac{D}{2})}l^{d-1}e^{-\frac{l^2}{2}}dl
\end{aligned}$$

which is numerically tractable.

Proof of Theorem 2.4.3. It follows from routine calculation that we can find a compact neighborhood of r around 0 such that $\nabla_r L(r,\theta)$ exists and is continuous. Thus we can use the main theorem

in [76] to show that

$$\begin{aligned}
\lim_{r\downarrow 0}\frac{l(r)-l(0)}{r} &= \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})}\int_{\mathcal{Y}} -\frac{(p(y;\theta))^2\cdot\lim_{r\downarrow 0}\frac{p_r(y)-p(y;\hat{\theta})}{r}}{(p(y;\hat{\theta}'))^2}dy \\
&= \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})}\lim_{r\downarrow 0}\frac{1}{r}\int_{\mathcal{Y}}\frac{(p(y;\theta))^2(p(y;\hat{\theta})-p_r(y))}{(p(y;\hat{\theta}'))^2}dy \\
&= \sup_{\theta\in\Theta^*(\delta_{\hat{\theta}})}\lim_{r\downarrow 0}\frac{1}{r}\int_{\mathcal{Y}}\frac{(p(y;\theta))^2}{p(y;\hat{\theta}')}-\frac{(p(y;\theta))^2\int_{\theta'\in\Theta}p(y;\theta')\mu_r(d\theta')}{(p(y;\hat{\theta}'))^2}dy \\
&= (1+\frac{\chi^2_{1-\alpha,D}}{n})\cdot\lim_{r\downarrow 0}\frac{1}{r}\Big(1-\inf_{\theta\in\Theta^*(\delta_{\hat{\theta}})}\mathbb{E}_{\theta'\sim\mu_r}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\Big).
\end{aligned}$$

□

To prove Corollary 2.4.3.1, we present two technical Lemmas 6 and 7.

Lemma 6. *For any $\theta\in\Theta^*(\delta_{\hat{\theta}})$, $\lim_{r\downarrow 0}\frac{1}{r}\Big(1-\mathbb{E}_{\theta'\sim\mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal\Sigma^{-1}(\theta'-\hat{\theta})}]\Big)$ is a fixed negative value.*

Proof of Lemma 6. For any $\theta\in\Theta^*(\delta_{\hat{\theta}})$, we have $\|\Sigma^{-1/2}(\theta-\hat{\theta})\|_2=\sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Denote $\rho_n=\sqrt{\log(1+\frac{\chi^2_{1-\alpha,D}}{n})}$. Furthermore, under $\theta'\sim\mu_r^1(d\theta')$, we have $\Sigma^{-1/2}(\theta'-\hat{\theta})\sim\eta_{\sqrt{r}}$, the uniform distribution inside the D-dimensional unit ball with radius $\sqrt{r}$, which can be viewed as the product of two independent random variables

$$\eta_{\sqrt{r}}\sim U\cdot R,$$

where U is the uniform distribution on the surface of the D-dimensional unit ball and R is the norm of the random vector ranged from 0 to $\sqrt{r}$. For any $0\le s\le\sqrt{r}$, since $\eta_{\sqrt{r}}$ follows a uniform distribution inside a D-dimensional unit ball, and the volume of a D-dimensional ball with radius s is proportional to s^D, then $f_R(s)$, the density of R, must satisfy

$$f_R(s)\sim\frac{ds^D}{ds}\sim s^{D-1},$$

which is equivalent to saying

$$f_R(s) = \frac{D}{(\sqrt{r})^D} s^{D-1}, \quad \text{for } 0 \le s \le \sqrt{r}.$$

Thus, we have that $\mathbb{E}[R^2] = c_1 r$ for some $c_1 > 0$. Now we let $u_1 = [1, 0, ..., 0] \in \mathbb{R}^D$. We utilize the proof in Lemma 5 as well as the independence of R, U to show that

$$\begin{aligned}
\mathbb{E}_{\theta' \sim \mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}] =& \mathbb{E}_{U,R}[e^{2\rho_n \cdot R \cdot u_1^\intercal U}] \\
=& \mathbb{E}_R[\mathbb{E}[e^{2\rho_n \cdot R \cdot u_1^\intercal U} | R]] \\
=& \mathbb{E}_R[M(4\rho_n R)/e^{2\rho_n R}] \\
\ge& \mathbb{E}[1 + 16c\rho_n^2 R^2] \ge 1 + 16c\rho_n^2 c_1 r.
\end{aligned}$$

Now it follows that

$$\lim_{r \downarrow 0} \frac{1}{r}\left(1 - \mathbb{E}_{\theta' \sim \mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\right) \le -16c\rho_n^2 c_1.$$

□

Lemma 7. *For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, $\lim_{r \downarrow 0} \frac{1}{r}\left(1 - \mathbb{E}_{\theta' \sim \mu_r^2}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\right)$ is a fixed negative value.*

Proof of Lemma 7. For any $\theta \in \Theta^*(\delta_{\hat{\theta}})$, we have $\|\Sigma^{-1/2}(\theta - \hat{\theta})\|_2 = \sqrt{\log(1 + \frac{\chi^2_{1-\alpha,D}}{n})}$. Denote $\rho_n = \sqrt{\log(1 + \frac{\chi^2_{1-\alpha,D}}{n})}$. Furthermore, under $\theta' \sim \mu_r^2(d\theta')$, we have $\Sigma^{-1/2}(\theta' - \hat{\theta}) \sim \mathcal{N}(0, rI_D)$. Using the moment generating function for Gaussian random variables, we have

$$\mathbb{E}_{\theta' \sim \mu_r^2}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}] = e^{(2r \cdot (\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta}))} = e^{2r\rho_n^2} \ge 1 + 2r\rho_n^2.$$

Now it follows that

$$\lim_{r \downarrow 0} \frac{1}{r}\left(1 - \mathbb{E}_{\theta' \sim \mu_r^1}[e^{2(\theta-\hat{\theta})^\intercal \Sigma^{-1}(\theta'-\hat{\theta})}]\right) \le -2\rho_n^2.$$

□

Proof of Corollary 2.4.3.1. Lemmas 6 and 7 combined with (2.72) indicate that increasing r to positive value would produce a descent direction for $l(r)$ at $r = 0$. □

Proof of Corollary 2.4.3.2. We proceed the proof as in Proposition 3. The proof for the case of $\mu_r^1(d\theta)$ is entirely similar. For the proof of the case $\mu_r^2(d\theta)$, we simply notice that if $Y \sim \mathbb{P}_t$, then

$$Y \overset{\mathcal{D}}{=} (1 - U_t)(\hat{\theta} + X_1) + U_t(\hat{\theta} + \sqrt{r}X_2 + X_3),$$

where U_t is an independent Bernoulli variable with success rate t and X_1, X_2, X_3 are independent $\mathcal{N}(0, \Sigma)$. Then, it follows that

$$\Sigma^{-1/2}(Y - \hat{\theta}) \overset{\mathcal{D}}{=} (1 - U_t)Z_1 + U_t(\sqrt{r}Z_2 + Z_3)$$

where Z_1, Z_2, Z_3 are now independent $\mathcal{N}(0, I_D)$. Consequently, the rotational invariance of Z now follows from the rotational invariance of Z_1, Z_2, Z_3 and their independence.

□

Chapter 3: General Feasibility Bounds for Sample Average Approximation via Vapnik-Chervonenkis Dimension

We investigate the feasibility of sample average approximation (SAA) for general stochastic optimization problems, including two-stage stochastic programming without the relatively complete recourse assumption. Instead of analyzing problems with specific structures, we utilize results from the *Vapnik-Chervonenkis* (VC) dimension and *Probably Approximately Correct* learning to provide a general framework that offers explicit feasibility bounds for SAA solutions under minimal structural or distributional assumption. We show that, as long as the hypothesis class formed by the feasbible region has a finite VC dimension, the infeasibility of SAA solutions decreases exponentially with computable rates and explicitly identifiable accompanying constants. We demonstrate how our bounds apply more generally and competitively compared to existing results.

The results here presented are new within the SAA feasibility domain. But similar results using VC-dimension have been applied to different contexts such as [67].

3.1 Introduction

Consider the stochastic optimization problem

$$\inf_{x\in\mathcal{X}} F(x) \triangleq \mathbb{E}[f(\xi, x)], \tag{3.1}$$

where $\mathcal{X}$ (typically $\mathcal{X} \subseteq \mathbb{R}^n$ or $\mathbb{R}^{n-p} \times \mathbb{Z}^p$ for mixed-integer decision sets) is a non-empty set for decision variable and $\xi : \Omega \to \Xi \subseteq R^r$ is some random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each realization of $\xi \in \Xi$, $f(\xi, \cdot) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a function taking values on the extended real line. Assume for each $x \in \mathcal{X}$, $f(\cdot, x) : \Xi \to \mathbb{R} \cup \{+\infty\}$ is measurable. We also assume the set

$\{x : x \in \mathcal{X} \text{ and } F(x) < +\infty\}$ is non-empty.

The class of problems under (3.1) are difficult to evaluate in general, especially for high-dimensional ξ. As a popular tractable approximation, the sample average approximation (SAA) method [84] solves the sampling-based counterpart of (3.1):

$$\inf_{x\in\mathcal{X}} \hat{F}_N(x) \triangleq \frac{1}{N}\sum_{i=1}^{N} f(\xi_i, x), \tag{3.2}$$

where $\xi^{[N]} \triangleq (\xi_1, \xi_2, ..., \xi_N)$ are IID samples drawn from $\mathbb{P}$. The optimal solution of SAA depends on the realization of $\xi^{[N]}$ and shall be denoted $x^\star(\xi^{[N]})$. Theoretical properties and numerical performances of SAA have been extensively studied in, e.g., [85, 84, 86], and its applications in stochastic optimization and chance-constrained programming can be found in, e.g., [87, 88, 89]. Most of these studies assume the condition $F(x) < +\infty$ for $x \in \mathcal{X}$, which is referred to as the relatively complete recourse condition in the context of two-stage stochastic programming. As an important class of (3.1), two-stage stochastic programing has applications in transportation planning [90, 91], disaster management [92], water recourse management [93] and inventory management [94]. However, in many real-word applications, relatively complete recourse assumption becomes restrictive and there has been a growing literature studying two-stage stochastic programming without this assumption, i.e. $F(x) = \infty$ for some $x \in \mathcal{X}$ (see [9, 95, 8]). In such a situation, the solution of SAA $x^\star(\xi^{[N]})$ may not be feasible for the original problem (3.1) and it would be desirable to quantify the level of feasibility of the SAA solution.

Indeed, as nicely discussed in [9, 8], the feasibility issue of SAA arises when $f(\xi, \cdot)$ maps to the extended real line. Following the notation of [8], let dom $f_\xi = \{x : f(\xi, x) < +\infty\}$. Then by solving (3.2) we would obtain an optimal solution $x^\star(\xi^{[N]}) \in$ dom $\hat{F}_N \triangleq \bigcap_{1\le i\le N}$ dom f_{ξ_i} where dom $\hat{F}_N$ is the feasible region for SAA. However, $x^\star(\xi^{[N]})$ might not be feasible for the original problem (3.1), i.e., $x^\star(\xi^{[N]}) \notin$ dom $F \triangleq \{x : F(x) < +\infty\}$, meaning it has a positive violation

probability $V(x^\star(\xi_{[N]})) > 0$ where

$$V(x) \triangleq \mathbb{P}(\xi : x \notin \text{dom } f_\xi). \tag{3.3}$$

We can also extend the definition of $V(\cdot)$ to include set input instead of point input, by letting

$$V(\mathcal{X}) \triangleq \mathbb{P}(\xi : \mathcal{X} \not\subseteq \text{dom } f_\xi). \tag{3.4}$$

In this chapter, we introduce a new framework based on the Vapnik-Chervonenkis (VC) dimension to study the feasibility of SAA solutions which includes, but is not limited to two-stage stochastic programming. Following [8, 9], we focus on showing the exponential decrease of $V(x^\star(\xi^{[N]}))$ as N grows. Specifically, letting $\mathbb{P}^N$ denote the IID sampling measure governing the generation of vector $\xi^{[N]}$ (notice the feasibility of $x^\star(\xi^{[N]})$ is random depending on $\xi^{[N]}$), we derive exponential bounds for $V(x^\star(\xi^{[N]}))$ under $\mathbb{P}^N$. As a key contribution, we show how our framework produces feasibility bounds that are both general and explicit. In particular, for solutions of SAA, we provide feasibility bounds *with explicit and computable* constants, *with no requirement on the geometric or distributional properties of* (3.1) (i.e., whether it is convex or linear, its optimal set has intersection with the boundary of dom F, $\mathcal{X}$ is finite or functions $\{f(\xi, \cdot)\}_{\xi \in \Xi}$ has a chain-constrained domain, as utilized in [9, 8]), and *with no specific regularity conditions on* $f(\xi, x)$ (i.e., Lipschitz continuity or the existence of certain moment generating function as in [9, 8]). Moreover, the analysis itself also does not hinge on the specific type of the problem (i.e., not limited to two-stage stochastic programming) and is widely applicable in both scenarios where some of the best-known results on SAA feasibility have been presented, and other scenarios where no similar results have been established. Furthermore, the feasibility result under this framework is not restricted to the optimal solution of SAA, but any generic point within the SAA feasible region with probability 1. Consequently, when the SAA problem is non-convex and solvable only up to local optimum, or when approximate algorithms are required, our results on feasibility guarantee would still hold. Finally, we show that the generality of this framework does not come at a cost of

worse sample complexity since the bounds under our framework are comparable to, if not better than the known ones.

The chapter is organized as follows. In Section 3.2, we review the recent papers with closely related results [9, 8]. In Section 3.3, we present our framework and general results. In Section 3.4, we specialize to examples of practical interests including two-stage stochastic programming. Moreover, we compare with known results to demonstrate the strengths of our framework.

3.2 Review of Related Results

We discuss the existing results on SAA feasibility in [9, 8]. A considerable part of [9] discusses how to solve a so-called "padded", modified version of SAA to obtain a *complete feasible* solution (i.e. $V(x) = 0$) with high confidence, which is somewhat different from the perspectives of this chapter and [8]. In particular, we consider the feasibility for SAA in its original form and do not restrict our attention to *complete feasible* solutions. However, [9] also discusses several results of

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta, \tag{3.5}$$

referred to as *high recourse likeihood solution* by the authors. These results are of the same type as ours and [8]. In particular, [9] presents these bounds in two cases, one of them being when $\mathcal{X}$ is finite and another being under the context of two-stage stochastic programming. We shall discuss in detail in Section 4 when we compare different results. On the other hand, the feasibility results in [8] are more general but can be summarized into three different scenarios.

- Scenario 1: In the presence of the so-called chain-constrained domain of order m (to be explained later) on dom f_ξ, [8] shows

$$\begin{aligned}\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) &\leq \sum_{k=0}^{m-1} \binom{N}{k} \alpha^k (1-\alpha)^{N-k} \\ &\leq \exp\left\{ - \frac{(N\alpha - m + 1)^2}{2N\alpha} \right\},\end{aligned}$$

where the second inequality is shown in both [20] and [8].

- Scenario 2: In the context of convexity, meaning $\mathcal{X}$ is closed and convex and the set of optimal solutions $\mathcal{X}^\star$ is non-empty and convex, and $f(\xi, \cdot)$ is convex for all $\xi \in \Xi$, along with additional regularity conditions on $f(\xi, \cdot)$ and $\mathcal{X}$, [8] shows that for $\mathcal{X}^\star$ in the interior of dom F,

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > 0) \leq Ce^{-N\beta},$$

where C and β are unknown constants.

- Scenario 3: In the context of convexity, if dom f_ξ is a chain-constrained domain as in Scenario 1, along with the additional regularity conditions, [8] shows that for $\mathcal{X}^\star$ which may have non-empty intersection with the boundary of dom F,

$$\begin{aligned}\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) &\leq Ce^{-N\beta} + \sum_{k=0}^{|\mathcal{J}|-1} \binom{N}{k} \alpha^k (1-\alpha)^{N-k} \\ &\leq Ce^{-N\beta} + \exp\left\{ - \frac{(N\alpha - |\mathcal{J}| + 1)^2}{2N\alpha} \right\}\end{aligned}$$

where C and β are again unknown constants as in Scenario 2 and J is the index set of active constraints at $\mathcal{X}^\star$ with the boundary of dom F. Notice it is shown in [22] that $|\mathcal{J}|$ is bounded by n, the dimension of the decision variable, which yields a useful upper bound regardless of the behavior of $\mathcal{J}$ (Also note that in this case the order of the chain-constrained domain does not play an explicit role in the bound).

In all scenarios, a desirable exponential decrease of $V(\cdot)$ as N grows can be shown. However, there are several potential limitations. First, there exist hidden constants in the feasibility bound: In Scenarios 2 and 3, which are of importance in stochastic convex programming, the rates of exponential decrease are governed by unknown constants β and C. Second, the dependence of the bound on m, the order of the chain-constrained structure which, as also mentioned in [9], can become potentially restrictive as m can be large in many cases. Furthermore, even though it is

motivated from practical examples in [8], the chain-constrained structure can be difficult to verify in general. The feasibility bound in Scenario 3 is less dependent on the chain order m, where the optimal solution of (3.1) intersects the boundary of dom F, but the chain-constrained structure is still required for analysis. It is thus desirable to generalize the feasibility results beyond the chain-constrained domain. Finally, note that while an explicit bound is presented in Scenario 1, it is a feasibility bound on the entire dom $\hat{F}_N$ instead of just $x^\star(\xi^{[N]})$, and is under the chain-constrained domain assumption.

3.3 Framework and Main Results

In this section we introduce our framework and main results. In particular, our framework is based on the Vapnik-Chervonenkis (VC) dimension of a collection of subsets on Ξ. This approach gives bounds for any generic point in dom $\hat{F}_N$, the feasible region of the SAA, which in particular implies bounds for $x^\star(\xi^{[N]})$. Note that our guarantee is still for a point, not for the entire set dom $\hat{F}_N$ which could lead to conservative estimates at an unnecessary cost, since we are interested in the feasibility of the SAA solution, not the entire region. In particular, instead of looking at dom $f_\xi = \{x : f(\xi, x) < +\infty\} \subseteq \mathcal{X}$, we investigate

$$H_x = \{\xi : f(\xi, x) < +\infty\} \subseteq \Xi \text{ for } x \in \mathcal{X} \tag{3.6}$$

and

$$H \triangleq \{H_x\}_{x\in\mathcal{X}}.$$

We consider the VC dimension of the class of subsets H. The VC dimension is commonly used to describe the complexity of a collection of sets or functions [96, 97, 98], which is also known as the “hypothesis space" in machine learning. The concept applies to a class of subsets H (see [99]), and can be generalized to binary functions and beyond. To define the VC dimension of a class of subsets H in $\mathbb{R}^r$, first note that a set of points $\{x_1, ..., x_d\} \subseteq \mathbb{R}^r$ is shattered by H if any subset of $\{x_1, ..., x_d\}$ can be picked out by some subset $C \in H$ (i.e., for any subset $D \subseteq \{x_1, ..., x_d\}$, there

is some $C \in H$ such that $D \subseteq C$ and $(\{x_1, ..., x_d\} \setminus D) \cap C = \emptyset$). The VC dimension of H is the maximal cardinality of the sets it can shatter, denoted by $V(H)$. For example, some well-known results on the VC dimensions of classes of sets are

- Positive intervals: if $H = \left\{\left\{x \in \mathbb{R} : x \in [a, b] \text{ for some } 0 \leq a \leq b\right\} | 0 \leq b \leq a\right\}$, we have $V(H) = 2$.
- Affine hyperplanes (Perceptrons): if $H = \left\{\left\{x \in \mathbb{R}^d : a^T x + b \geq 0 \text{ for } a \in \mathbb{R}^d \text{ and } b \in \mathbb{R}\right\} | a \in \mathbb{R}^d, b \in \mathbb{R}\right\}$, we have $V(H) = d + 1$.
- Convex sets: if $H = \left\{C : C \subseteq \mathbb{R}^d \text{ and } C \text{ is convex } \right\}$, we have $VC(H) = +\infty$.

An important concept in computational learning theory tightly related to the VC dimension is *Probably Approximately Correct* (PAC) learning. In this context, the VC dimension of H can be used in PAC learning to derive bounds on the sample complexity needed to achieve a desired level of accuracy between "in-sample-error" and "generalization error" within class H (see, e.g. [96, 100, 97]). As it turns out, these types of result directly transfer towards the sample complexity needed for desired level of feasibility for any generic point in dom $\hat{F}_N$.

We note, as we shall see in later sections, the Ξ in (3.6) can be reparametrized and does not have to be viewed in $\mathbb{R}^r$ for fixed r defined in (3.1). For illustration, consider the following example.

Suppose $x \in \mathcal{X} \subseteq \mathbb{R}$ and ξ is a random vector defined on $\mathbb{R}^r$ for some $r > 0$. Let $g(\cdot) : \mathbb{R}^r \to \mathbb{R}$ be a given function. Then, suppose $f(\xi, x) < \infty$ if and only if $g(\xi) \cdot x \geq 1$. Then, $\{H_x\}_x$ in (3.6) could be defined as

$$H_x = \{\xi : g(\xi) \cdot x \geq 1\} \subseteq \mathbb{R}^r, \forall x \in \mathcal{X}.$$

On the other hand, if we define random variable $\xi' = g(\xi)$ on $\mathbb{R}$, then we can alternatively define

$$H'_x = \{\xi' : \xi' \cdot x \geq 1\} \subseteq \mathbb{R}, \forall x \in \mathcal{X}.$$

Typically $\{H_x\}_x$ and $\{H'_x\}_x$ would have different VC dimensions, even though for any x, H_x and

H'_x are equivalent with probability 1:

$$\mathbb{P}\big(\{\omega \in \Omega : \xi(\omega) \in H_x\} \triangle \{\omega \in \Omega : \xi'(\omega) \in H'_x\}\big) = 0,$$

where $A \triangle B = (A \setminus B) \cup (B \setminus A)$ is the symmetric difference operator on sets. Consequently, instead of fixing a canonical representation of ξ in (3.1), we sometimes utilize this flexibility to change representations for the convenience of our analysis.

3.3.1 Main Result

We now present our main theorem on SAA feasibility and its proof.

Theorem 3.3.1. *Let $H \triangleq \{H_x\}_{x\in\mathcal{X}}$ be the class of subsets defined in* (3.6) *and suppose H has finite VC dimension d, i.e., $V(H) = d < +\infty$. Moreover, let $\xi^{[N]} = \{\xi_1, ..., \xi_N\}$ be IID samples from $\mathbb{P}$ (consequently $\xi^{[N]} \sim \mathbb{P}^N$). Then, if*

$$N \geq \frac{4}{\alpha}\left(d \log\left(\frac{12}{\alpha}\right) + \log\left(\frac{2}{\delta}\right)\right), \tag{3.7}$$

we have

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta$$

for any $0 < \delta, \alpha < 1$.

Proof. Under the assumption $d < +\infty$, it follows from Theorem 8.4.1 of [96] that, when $N \geq \frac{4}{\alpha}\left(d \log\left(\frac{12}{\alpha}\right) + \log\left(\frac{2}{\delta}\right)\right)$,

$$\sup_{x\in\operatorname{dom}\hat{F}_N} \mathbb{P}(f(\xi, x) = +\infty) \leq \alpha. \tag{3.8}$$

with probability at least $1 - \delta$ under $\mathbb{P}^N$. Thus, if we let $\mathcal{X}_\alpha = \{x \in \mathcal{X} : V(x) \leq \alpha\}$, then from (3.8) we know that $\operatorname{dom}\hat{F}_N \subseteq \mathcal{X}_\alpha$ with probability (under $\mathbb{P}^N$) is at least $1 - \delta$, which can be translated to

$$\mathbb{P}^N\left(\sup_{x\in\operatorname{dom}\hat{F}_N} V(x) > \alpha\right) \leq \delta.$$

Since $x^\star(\xi^{[N]}) \in \operatorname{dom} \hat{F}_N$ by definition, we have $V(x^\star(\xi^{[N]})) \leq \sup_{x\in\operatorname{dom} \hat{F}_N} V(x)$ and consequently

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta.$$

□

Remark 2. *First, in the proof of Theorem 5.1, the sample complexity in* (3.7) *comes from Theorem 8.4.1 of [96] and provides a* $O(\frac{d}{\epsilon}\log(\frac{1}{\epsilon}) + \frac{1}{\epsilon}\log(\frac{1}{\delta}))$ *bound. It is worth noting that a better sample complexity of* $O(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log(\frac{1}{\delta}))$ *can be achieved by recent breakthroughs of [101, 102]. We choose to present the result from Theorem 8.4.1 in [96] because it is more concise and explicit. However, under our framework, a better bound is indeed obtainable. Second, as shown in the proof, the feasibility result of the theorem holds not just for the solution of SAA, but also for any generic point within the feasible region of SAA. In other words, Theorem 5.1 holds for any algorithm that can output a solution* $x^\star(\xi^{[N]})$ *(not necessarily the optimal solution of the SAA) in the feasible region of SAA with probability 1. This observation is particularly important when the considered SAA problem is non-convex and solvable only up to local optimum, or when approximate algorithms are required.*

There are several advantages when applying Theorem 5.1 to bounding the feasibility of SAA solutions: 1) It does not rely on any strong assumptions on the structures of (3.1) and (3.2). As we shall see in an example later, even when the chain-constrained domain condition in [8] becomes restrictive, analysis based on VC dimension would remain effective. 2) Our bound is explicit and computable with no hidden constants. 3) One might argue the generality of Theorem 5.1 would come at a cost of higher sample complexity compared to analyses with more specific conditions. However, as we shall see, this is not necessarily the case when we compare bounds even within the chain-constrained context.

Next, while Theorem 5.1 is a result on sample complexity, it is straightforward to convert it into an asymptotic rate of convergence of feasibility with sample size N. The portion of infeasible SAA solutions still decreases exponentially as in [8], and the rate of which can now also be made

explicit. We summarize it into a corollary.

Corollary 3.3.1.1. *Under the same condition of Theorem 5.1,*

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq 2\exp\left(-\frac{N\alpha}{4}\right)\left(\frac{12}{\alpha}\right)^d.$$

We also note that direct comparisons on sample complexity in special cases are also possible (only when the rate of convergence is known, which only applies to Scenario 1 in [8]), because it is shown in [28] that a relatively tight sufficient condition for

$$\sum_{k=0}^{m-1}\binom{N}{k}\alpha^i(1-\alpha)^{N-i} \leq \delta,$$

is

$$N \geq \frac{e}{e-1}\frac{1}{\alpha}\left(m - 1 + \log\left(\frac{1}{\delta}\right)\right), \tag{3.9}$$

which provides a tight bound on sample complexity and whom we shall make use of later. Finally, we note that the VC dimension has also been used in [23] in analyzing constraint sampling, but in the context of solving Markov decision problems.

3.4 Examples and Special Structures

In this section we apply Theorem 5.1 in several problems of considerable practical interests and compare with established results. Throughout the proofs, we use the following definitions and Theorem 1.1 from [99]:

Theorem 3.4.1 (Theorem 1.1 from [99])**.** *Given classes of subsets C_1, C_2, ..., C_m with $V_j = V(C_j) < \infty$, define*

$$\sqcap_{j=1}^m C_j \triangleq \{\cap_{j=1}^m C_j : C_j \in C_j, j = 1, ..., m\}$$
$$\sqcup_{j=1}^m C_j \triangleq \{\cup_{j=1}^m C_j : C_j \in C_j, j = 1, ..., m\},$$

and let $V = \sum_{j=1}^{m} V_j$. *Then,*

$$\max(V(\sqcap_{j=1}^{m} C_j), V(\sqcup_{j=1}^{m} C_j)) \leq \frac{e}{(e-1)\log 2} V \log(\frac{e}{\log 2} m).$$

We also use a key result from [103] on the upper bound of VC dimension for sets determined by finite-dimensional function spaces. For a concise proof, one can also see Lemma 2.6.15 in [104].

Theorem 3.4.2. *Given arbitrary space* $\mathcal{S}$, *let* $\mathcal{G}$ *be a finite-dimensional vector space of functions* $g(\cdot) : \mathcal{S} \to \mathbb{R}$. *Then, the classes of sets:*

$$H = \{\{s \in \mathcal{S} : g(s) \geq 0\}\}_{g \in \mathcal{G}}$$

has VC dimension at most $\dim \mathcal{G}$.

According to Theorem 3.4.2, $\mathcal{U} = \{\{(y, z) : y^T x \leq z\}\}_{x \in \mathbb{R}^d}$ has VC dimension at most d (in fact, it is equal to d; see [23] or [103]).

Finally, we use the notation $[\cdot]$ in the following way. For a positive integer q, $[q]$ denotes the set $\{1, ..., q\}$. Moreover, given a vector $v \in \mathbb{R}^q$, $[v]_j$ denotes the j-th component of v, for $j \in [q]$.

3.4.1 Two-Stage Stochastic Programming

One of the main motivating examples in studying SAA feasibility, mentioned in both [9, 8], is the two-stage stochastic programming problem without relatively complete recourse. In [8], the form of $f(\xi, x)$ in (3.1) is defined as follows:

$$\begin{aligned} f(\xi, x) \triangleq \inf_{y} \quad & g(\xi, y) \\ \text{s.t.} \quad & W_\xi y + T_\xi x = h_\xi, \\ & y \geq 0, \end{aligned} \tag{3.10}$$

where $g(\xi, \cdot)$ is convex, finite everywhere $\forall \xi$, almost surely. Furthermore, [8] assumes that there are only finitely many distinct values for W_ξ or T_ξ , i.e., $|\{W_\xi\}| = p$ and $|\{T_\xi\}| = q$ where $\{p, q\} \subseteq \mathbb{Z}^+$. By Farkas' lemma, $\{y \geq 0 : W_\xi y + T_\xi x = h_\xi\}$ is non-empty if and only if $a^T(h_\xi - T_\xi x) \geq 0$ for all a such that $a^T W \geq 0$. Consequently, as shown in [8], we have

$$\text{dom } f_\xi = \{x : a_{ij}^T T_k x \leq a_{ij}^T h_\xi, W_\xi = W_i, j \in J_i, T_\xi = T_k\}, \tag{3.11}$$

where $\{a_{ij}\}_{j \in J_i}$ is the set of non-equivalent extreme rays of polyhedral cone $C_i = \{a : a^T W_i \geq 0\}$ and J_i is the index set for these extreme rays of C_i. This allows [8] to use the chain-constrained structure. Here, a chain-constrained domain is defined as follows:

Definition 1. *A collection of functions $\{f(\xi, \cdot)\}_{\xi \in \Xi}$ has chain-constrained domain of order m if there exist m chains $\{U_k^\xi\}_{\xi \in \Xi}$ and*

$$dom\ f_\xi = \bigcap_{k=1}^{m} U_k^\xi$$

where a collection of sets $\{U^\omega\}_{\omega \in I}$ is a chain if for any $\omega_1, \omega_2 \in I$, we have either $U_{\omega_1} \subseteq U_{\omega_2}$ or $U_{\omega_2} \subseteq U_{\omega_1}$.

It is shown in [8] that dom f_ξ in (3.11) is a chain-constrained domain of order $m = q \sum_{i=1}^{p} |J_i|$. Consequently, Scenario 1 in [8] can be applied to show that

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > 1 - \alpha) \leq \sum_{k=0}^{m-1} \binom{N}{k} \alpha^k (1 - \alpha)^{N-k},$$

which has a sample complexity

$$\frac{e}{e-1} \frac{1}{\alpha} \left(m - 1 + \log\left(\frac{1}{\delta}\right)\right) \tag{3.12}$$

for achieving $\mathbb{P}^N(V(x^\star(\xi^{[N]})) > 1 - \alpha) \leq \delta$ according to (3.9).

Notice a necessary assumption made in [8] is that only finitely many distinct values for W_ξ or T_ξ are allowed, i.e., $|\{W_\xi\}| = p$ and $|\{T_\xi\}| = q$ where $\{p, q\} \subseteq \mathbb{Z}^+$. However, using Theorem 5.1,

we can get a different sample complexity and concentration bounds, even when cardinalities of $|\{W_\xi\}|$ and $|\{T_\xi\}|$ are infinite. We first address (3.10) in its original form.

Corollary 3.4.2.1. *Consider* (3.10). *Let* $\xi^{[N]} = \{\xi_1, ..., \xi_N\}$ *be IID samples from* $\mathbb{P}$ *(consequently* $\xi^{[N]} \sim \mathbb{P}^N$*), and* $x^\star(\xi^{[N]})$ *be the SAA solution. Then, if*

$$N \geq \frac{4}{\alpha}\left(\left(\frac{e}{(e-1)\log 2}|J|(n+1)\log\left(\frac{e}{\log 2}\cdot|J|\right)\right)\log\left(\frac{12}{\alpha}\right)+\log\left(\frac{2}{\delta}\right)\right), \tag{3.13}$$

where $|J| = \max_{i\in[q]}|J_i|$*, we have*

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta.$$

for any $0 < \delta, \alpha < 1$*. Equivalently, in terms of convergence rate, we have*

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq 2\exp\left(-\frac{N\alpha}{4}\right)\left(\frac{12}{\alpha}\right)^{\left(\frac{e}{(e-1)\log 2}|J|(n+1)\log\left(\frac{e}{\log 2}\cdot|J|\right)\right)}. \tag{3.14}$$

Proof. Define $I(\cdot) : \Xi \to [p]$ as the indexing function such that $I(\xi) = i$ when $W_\xi = W_i$. We then observe $H \triangleq \{H_x\}_{x\in\mathcal{X}}$ defined in (3.6) becomes

$$H_x = \{\xi : a_{I(\xi)j}^T T_\xi x \leq a_{I(\xi)j}^T h_\xi, \forall j \in J_{I(\xi)}\}$$

where $\{a_{ij}\}_{j\in J_i}$ is the set of non-equivalent extreme rays of polyhedral cone $\{a : a^T W_i \geq 0\}$. Define $|J| = \max_{i\in[q]}|J_i|$ and for all $\xi \in \Xi$, let $\{(y_{\xi j}, z_{\xi j})\}_{j\in|J|}$ be

$$y_{\xi j}^T = \begin{cases} a_{I(\xi)j}^T T_\xi, & \text{for } 1 \leq j \leq |J_{I(\xi)}| \\ \mathbf{0}, & \text{for } |J_{I(\xi)}| < j \leq |J| \end{cases}$$

$$z_{\xi j} = \begin{cases} a_{I(\xi)j}^T h_\xi, & \text{for } 1 \leq j \leq |J_{I(\xi)}| \\ 0, & \text{for } |J_{I(\xi)}| < j \leq |J|. \end{cases}$$

Then, define $y_\xi^T = (y_{\xi 1}^T, y_{\xi 2}^T, ..., y_{\xi |J|}^T) \in \mathbb{R}^{|J|n}$ and $z_\xi = (z_{\xi 1}, z_{\xi 2}, ..., z_{\xi |J|}) \in \mathbb{R}^{|J|}$. Moreover, for $j \in [|J|]$, define $v_j(\cdot) : \mathbb{R}^n \to \mathbb{R}^{|J|n}$ to be

$$v_j(x) = \begin{cases} [x]_i, & \text{for } (j-1)n + 1 \le i \le jn \\ 0, & \text{otherwise.} \end{cases}$$

Then, we can redefine

$$H_x = \bigcap_{j=1}^{|J|} \{(y_\xi, z_\xi) : y_\xi^T v_j(x) \le [z_\xi]_j\}.$$

Given $j \in [|J|]$, let $e_j \in \mathbb{R}^{|J|}$ be the vector with 1 in the j-th component and 0 otherwise. Define a class of function $\mathcal{G} = \{g_{(x,c)}(\cdot)\}_{(x,c)\in\mathbb{R}^n\times\mathbb{R}}$ on $\mathbb{R}^{|J|(n+1)}$ such that, given $(y, z) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|}$,

$$g_{(x,c)}((y, z)) = [y, z]^T \begin{bmatrix} -v_j(x) \\ c \cdot e_j \end{bmatrix}.$$

It is straightforward to check $\mathcal{G}$ is a finite-dimensional vector space of functions with $\dim \mathcal{G} \le n+1$. Then, according to Theorem 3.4.2, the VC dimension of

$$\{\{(y, z) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|} : g_{x,c}((y, z)) \ge 0\}\}_{(x,c)\in\mathbb{R}^n\times\mathbb{R}}$$

is at most $n + 1$. Consequently, as a smaller collection of sets, the VC dimension of

$$\{\{(y, z) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|} : g_{x,1}((y, z)) \ge 0\}\}_{x\in\mathcal{X}}$$

is also at most $n + 1$. Thus, for each $j \in [|J|]$, the VC dimension of $\mathcal{U}_j = \{\{(y_\xi, z_\xi) : y_\xi^T v_j(x) \le [z_\xi]_j\}\}_{x\in\mathcal{X}}$ is at most $n + 1$. Finally, it follows from Theorem 3.4.1 that

$$V(H) \le V(\sqcap_{j=1}^{|J|} \mathcal{U}_j) \le \frac{e}{(e-1)\log 2} |J|(n+1) \log\left(\frac{e}{\log 2} \cdot |J|\right).$$

The corresponding sample complexity and convergence rate follow from Theorem 5.1. □

Note that 3.4.2.1 does not require any convexity assumption on g_ξ nor distributional assumptions on the random variables W_ξ and T_ξ. Furthermore, if $|\{W_\xi\}|$ and $|\{T_\xi\}|$ are infinite, our result still holds. This is because the same proof can be applied as long as $|J| = \max_{\xi\in\Xi}\{$ # of extreme rays for the co $a^T W_\xi \geq 0\}\}$ is finite. However, it is known that the number of non-equivalent extreme rays of a polyhedral cone $\{a : a^T W \geq 0\}$ is finite and can be bounded by a term of $O(m_1^{\lfloor \frac{n_1}{2} \rfloor})$ involving only m_1 and n_1 for $W \in \mathbb{R}^{m_1\times n_1}$ (see [105, 106, 107]). Thus, $|J| < +\infty$ regardless of the cardinalities of $\{W_\xi\} \subseteq \mathbb{R}^{m_1\times n_1}$. Notice we can view m_1 to be deterministic, as long as m_1 is bounded almost surely. We summarize this into another Corollary.

Corollary 3.4.2.2. *Consider* (3.10). *Let* $\xi^{[N]} = \{\xi_1, ..., \xi_N\}$ *be IID samples from* $\mathbb{P}$ *(consequently* $\xi^{[N]} \sim \mathbb{P}^N$*), and* $x^\star(\xi^{[N]})$ *be the SAA solution. If* $|\{W_\xi\}|$ *and* $|\{T_\xi\}|$ *are infinite but* m_1 *is bounded where* $\{W_\xi\} \subseteq \mathbb{R}^{m_1\times n_1}$*, then the result of 3.4.2.1 still holds.*

Proof. For $\{W_\xi\} \subseteq \mathbb{R}^{m_1\times n_1}$, it is known that $|J| < +\infty$ where

$$|J| = \max_{\xi\in\Xi}\{ \text{ \# of extreme rays for the cone } \{a : a^T W_\xi \geq 0\}\}.$$

Then, let $\mathcal{A}_\xi$ be the set of non-equivalent extreme rays of polyhedral cone $\{a : a^T W_\xi \geq 0\}$. Observe $H \triangleq \{H_x\}_{x\in\mathcal{X}}$ defined in (3.6) becomes

$$H_x = \{\xi : a^T T_\xi x \leq a^T h_\xi, \forall a \in \mathcal{A}_\xi\}.$$

For all $\xi \in \Xi$, since $|\mathcal{A}_\xi| \leq |J|$, we can label the elements in $\mathcal{A}_\xi$ by $\{a_{\xi j}\}_{j\in[|A_\xi|]}$. Then, define $\{(y_{\xi j}, z_{\xi j})\}_{j\in|J|}$ as

$$y_{\xi j}^T = \begin{cases} a_{\xi j}^T T_\xi, & \text{for } 1 \leq j \leq |\mathcal{A}_\xi| \\ \mathbf{0}, & \text{for } |\mathcal{A}_\xi| < j \leq |J| \end{cases}$$

$$z_{\xi j} = \begin{cases} a_{\xi j}^T h_\xi, & \text{for } 1 \le j \le |\mathcal{A}_\xi| \\ 0, & \text{for } |\mathcal{A}_\xi| < j \le |J|. \end{cases}$$

The rest of proof follows exactly as in Corollary 3.4.2.1. □

Compared with our bound (3.13), the chain-constrained bound (3.12) relies on the order of the chain $m = q\sum_{i=1}^{p} |J_i|$. If the cardinality of the support of W_ξ or T_ξ gets large (i.e., $qp \gg n$), or potentially infinite (for continuous random variable), then the bound in (3.12) with a sample complexity of $O(\frac{qp|J|}{\alpha} + \frac{1}{\alpha}\log(\frac{1}{\delta}))$ becomes loose or even inapplicable due to the term qp. On the other hand, our VC bound (3.13) with a sample complexity $O(\frac{|J|n}{\alpha}\log|J|\log(\frac{1}{\alpha}) + \frac{1}{\alpha}\log(\frac{1}{\delta}))$ maintains the same dependence on the dimension n regardless of the cardinality of the support for W_ξ or T_ξ. Moreover, if we use the PAC bound from [101, 102] as mentioned in Remark 2, the bound would be improved to $O(\frac{|J|n}{\alpha}\log|J| + \frac{1}{\alpha}\log(\frac{1}{\delta}))$. Finally, in both bounds, the term $|J|$ appears. However, as mentioned previously, an explicit bound for $|J|$ of $O(m_1^{\lfloor \frac{n_1}{2} \rfloor})$ can be obtained by m_1, n_1 where $\{W_\xi\} \subseteq \mathbb{R}^{m_1 \times n_1}$ but we omit it here as it is not essential for our comparison. Finally, the bound in Scenario 3 of [8] also applies to (3.10) and is not limited by the order of the chain-structure. However, the bound there is not explicitly computable since the β term is hidden.

The dependence on the order of the chain m is also addressed in [9]. Using ideas similar to the scenario approximation of chance-constrained problems in [108, 24, 22] as well as properties of linear programming (e.g. existence of basic optimal solutions), [9] is able to provide a sample complexity for two-stage stochastic linear programming independent of the cardinalities of $\{W_\xi\}$ or the order of the chain. However, the derivation of our bound in (3.13) does not depend on the linearity of the optimization problem and hence is not limited to two-stage stochastic programming with linear recourse. In particular, in [9], the first stage $\mathcal{X}$ is defined by linear constraints $Ax = b$ for some $A \in \mathbb{R}^{m \times n}$ and the second stage problem bears a linear objective $q(\xi)^T y$. In contrast, our bound is valid for general $\mathcal{X}$ in the first stage and $g(\xi, y)$ in the second-stage problem in (3.10). That being said, the bound derived in [9] has notable strengths in the linear case, in terms of the dependence on problem parameters, gained via a more efficient exploitation of the linear structure.

Specifically, the sample complexity in [9] is (adapted to the notation in this chapter)

$$O\left(\frac{1}{\alpha}\left(nn_1\big(\log(\frac{m_1}{n_1+1})+1\big)+n\big(\log(\frac{m}{n}+2)+\log(\frac{1}{\alpha})+1\big)+\log(\frac{1}{\delta})\right)\right), \tag{3.15}$$

which has better dependence on m_1, n_1, as the dependence on $|J|$ in (3.13) is $O(m_1^{\lfloor \frac{n_1}{2} \rfloor})$ in the worst case. Nonetheless, (3.13) has a similar dependence on n to the bound in (3.15), and does not depend on m in (3.15) at all. Omitting the dependence on these problem size parameters (e.g., constants based on n, m, m_1, n_1 and $|J|$), the bound derived in [9] is of order $O(\frac{1}{\alpha}\log(\frac{1}{\delta})+\frac{1}{\alpha}\log(\frac{1}{\alpha}))$, which of the same order as the bound (3.13). Moreover, (3.13) can be slightly improved to be of order $O(\frac{1}{\alpha}\log(\frac{1}{\delta})+\frac{1}{\alpha})$ bound based on Remark 2.

3.4.2 Two-Stage Stochastic Integer Programming

The SAA method has also been applied in two-stage stochastic programming with (mixed) integer recourse [109, 110, 111, 112]. We consider the following two-stage stochastic integer programming where $\mathcal{X} \subseteq \mathbb{R}^{n-p} \times \mathbb{Z}^p$ contains integer components in the first stage (3.1) and the second stage is a mixed integer program (MIP):

$$\begin{aligned} f(\xi, x) \triangleq \inf_{y} \quad & g(\xi, y, y_0) \\ \text{s.t.} \quad & W_\xi y + W^0_\xi y_0 + T_\xi x = h_\xi, \\ & y \in \mathbb{R}^{n'}_+, y_0 \in \mathcal{Z} \subseteq \mathbb{Z}^{p'}_+, \end{aligned} \tag{3.16}$$

for some $n', p' \in \mathbb{Z}_+$. Here $g(\xi, y, y_0)$ can be a general function as in (3.10), although for much of the theoretical and practical interest (also applicability), it is assumed to be in linear programs where $g(\xi, y, y_0) = q(\xi)^T y + q_0(\xi)^T y_0$. Moreover, most literature also assumes relatively complete recourse by fixing a deterministic recourse matrix (i.e., $W_\xi = W$ and $W^0_\xi = W^0$ with probability 1) such that $\{y \in \mathbb{R}^{n'}_+ \times \mathbb{Z}^{p'}_+ : [W|W^0]y = t\}$ is non-empty for all t. Consequently, the feasibility of SAA solution for two-stage stochastic integer programming without relatively complete recourse has rarely been considered. In fact, due to the general non-convex and discontinuous nature of MIP,

specialized approximate/iterative algorithms are usually required and the solutions are no longer guaranteed to be optimal. However, even without relatively complete recourse, as mentioned in Remark 2, as long as the solutions output from such algorithms are within the SAA feasible region with probability 1, the feasibility result from Theorem 5.1 still holds. Notice we have assumed the set $\{x : x \in \mathcal{X} \text{ and } F(x) < +\infty\}$ is non-empty throughout the chapter (see the beginning of the introduction) and the SAA feasible region is non-empty with probability 1 under this assumption.

Under this setting, it is possible to provide a feasibility bound for (3.16) when $|\mathcal{Z}| < \infty$. This condition is satisfied when y_0 is restricted to be binary as in [112] (i.e., $y_0 \in \{0,1\}^{p'}$). On the other hand, if the solutions are polynomially bounded by the size of data (e.g., integer linear programming [113]), then it is also possible to only consider solving (3.16) in a finite, although possibly large bounded set $\mathcal{Z} \subseteq \mathbb{Z}_+^{p'}$.

Corollary 3.4.2.3. *Consider* (3.16). *Suppose* $|\mathcal{Z}| < \infty$ *and* $|J| < \infty$ *where*

$$|J| = \max_{\xi \in \Xi} \{ \text{\# of extreme rays for the cone } \{a : a^T W_\xi \geq 0\}\}.$$

Then, let $\xi^{[N]} = \{\xi_1, ..., \xi_N\}$ *be IID samples from* $\mathbb{P}$ *(consequently* $\xi^{[N]} \sim \mathbb{P}^N$*), and* $x^\star(\xi^{[N]})$ *be the SAA solution, or any output within the SAA feasibility region with probability 1. Then, if*

$$N \geq \frac{4}{\alpha}\left(d \log\left(\frac{12}{\alpha}\right) + \log\left(\frac{2}{\delta}\right)\right), \tag{3.17}$$

where

$$d = \left(\frac{e}{(e-1)\log 2}\right)^2 |\mathcal{Z}||J|(n+2)\log\left(\frac{e|J|}{\log 2}\right)\log\left(\frac{e|\mathcal{Z}|}{\log 2}\right),$$

then we have $\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta$*, for any* $0 < \delta, \alpha < 1$*. Equivalently, in terms of convergence rate, we have*

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq 2\exp\left(-\frac{N\alpha}{4}\right)\left(\frac{12}{\alpha}\right)^d. \tag{3.18}$$

Proof. Let $\mathcal{A}_\xi$ be the set of non-equivalent extreme rays of polyhedral cone $\{a : a^T W_\xi \geq 0\}$.

Using Farkas' lemma, we can construct $H \triangleq \{H_x\}_{x\in\mathcal{X}}$ defined in (3.6) as

$$H_x = \bigcup_{y_0\in\mathcal{Z}} \{\xi : a^T(T_\xi x + W_\xi^0 y_0) \le a^T h_\xi, \forall a \in \mathcal{A}_\xi\},$$

for $(x, y_0) \in \mathcal{X}\times\mathcal{Z}$. For all $\xi \in \Xi$, since $|\mathcal{A}_\xi| \le |J|$, we can label the elements in $\mathcal{A}_\xi$ by $\{a_{\xi j}\}_{j\in[|A_\xi|]}$. Then, define $\{(y_{\xi j}, z_{\xi j}, w_{\xi j})\}_{j\in|J|}$ as

$$y_{\xi j}^T = \begin{cases} a_{\xi j}^T T_\xi, & \text{for } 1 \le j \le |\mathcal{A}_\xi| \\ \mathbf{0}, & \text{for } |\mathcal{A}_\xi| < j \le |J| \end{cases}$$

$$z_{\xi j} = \begin{cases} a_{\xi j}^T h_\xi, & \text{for } 1 \le j \le |\mathcal{A}_\xi| \\ 0, & \text{for } |\mathcal{A}_\xi| < j \le |J| \end{cases}$$

$$w_{\xi j}^T = \begin{cases} a_{\xi j}^T W_\xi^0, & \text{for } 1 \le j \le |\mathcal{A}_\xi| \\ \mathbf{0}, & \text{for } |\mathcal{A}_\xi| < j \le |J|. \end{cases}$$

Define $y_\xi^T = (y_{\xi 1}^T, y_{\xi 2}^T, ..., y_{\xi|J|}^T) \in \mathbb{R}^{|J|n}$, $z_\xi = (z_{\xi 1}, z_{\xi 2}, ..., z_{\xi|J|}) \in \mathbb{R}^{|J|}$ and $w_\xi^T = (w_{\xi 1}^T, w_{\xi 2}^T, ..., w_{\xi|J|}^T) \in \mathbb{R}^{|J|p'}$. Moreover, for $j \in [|J|]$, define $v_j(\cdot) : \mathbb{R}^n \to \mathbb{R}^{|J|n}$, $u_j : \mathbb{Z}^{p'} \to \mathbb{Z}^{|J|p'}$ to be

$$v_j(x) = \begin{cases} [x]_i, & \text{for } (j-1)n + 1 \le i \le jn \\ 0, & \text{otherwise} \end{cases}$$

$$u_j(x) = \begin{cases} [y_0]_i, & \text{for } (j-1)p' + 1 \le i \le jp' \\ 0, & \text{otherwise.} \end{cases}$$

Then, we can redefine

$$H_x = \bigcup_{y_0\in\mathcal{Z}} \bigcap_{j=1}^{|J|} \{(y_\xi, z_\xi, w_\xi) : y_\xi^T v_j(x) + w_\xi^T u_j(y_0) \le [z_\xi]_j\}. \tag{3.19}$$

Given $j \in [|J|]$ and $y_0 \in \mathcal{Z}$, let $e_j \in \mathbb{R}^{|J|}$ be the vector with 1 in the j-th component and 0 otherwise. Define a class of function $\mathcal{G} = \{g_{(x,c_1,c_2)}(\cdot)\}_{(x,c_1,c_2)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}}$ on $\mathbb{R}^{|J|(n+1+p')}$ such that, given $(y, z, w) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|} \times \mathbb{R}^{|J|p'}$,

$$g_{(x,c_1,c_2)}((y,z,w)) = [y,z,w]^T \begin{bmatrix} -v_j(x) \\ c_1 \cdot e_j \\ -c_2 \cdot u_j(y_0) \end{bmatrix}.$$

It is straightforward to check $\mathcal{G}$ is a finite-dimensional vector space of functions with $\dim \mathcal{G} \leq n+2$. Then, according to Theorem 3.4.2, the VC dimension of

$$\{\{(y,z,w) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|} \times \mathbb{R}^{|J|p'} : g_{(x,c_1,c_2)}((y,z,w)) \geq 0\}\}_{(x,c_1,c_2)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}}$$

is at most $n + 2$. Consequently, as a smaller collection of sets, the VC dimension of

$$\{\{(y,z,w) \in \mathbb{R}^{|J|n} \times \mathbb{R}^{|J|} \times \mathbb{R}^{|J|p'} : g_{(x,1,1)}((y,z,w)) \geq 0\}\}_{x\in\mathcal{X}}$$

is also at most $n + 2$. Thus, for each $j \in [|J|]$, the VC dimension of

$$\mathcal{U}_j^{y_0} = \{\{(y_\xi, z_\xi, w_\xi) : y_\xi^T v_j(x) + w_\xi^T u_j(y_0) \leq [z_\xi]_j\}\}_{x\in\mathcal{X}}$$

is at most $n + 2$. Consequently, given $y_0 \in \mathcal{Z}$, it follows from Theorem 3.4.1 that

$$V(\sqcap_{j=1}^{|J|} \mathcal{U}_j^{y_0}) \leq \frac{e}{(e-1)\log 2}|J|(n+2)\log(\frac{e}{\log 2}|J|)$$

and

$$V\Big(\sqcup_{y_0\in\mathcal{Z}}\Big(\sqcap_{j=1}^{|J|}\mathcal{U}_j^{y_0}\Big)\Big) \leq \Big(\frac{e}{(e-1)\log 2}\Big)^2 |\mathcal{Z}||J|(n+2)\log(\frac{e|J|}{\log 2})\log(\frac{e|\mathcal{Z}|}{\log 2}).$$

Thus, for H_x defined in (3.19), we have $V(\{H_x\}_{x\in\mathcal{X}}) \le d$ where

$$d = \Big(\frac{e}{(e-1)\log 2}\Big)^2 |\mathcal{Z}||J|(n+2)\log(\frac{e|J|}{\log 2})\log(\frac{e|\mathcal{Z}|}{\log 2}).$$

The rest of the proof follows as in 3.4.2.1. □

As we can see, the portion of infeasible SAA solutions (not necessarily optimal) still decreases exponentially as the sample size N increases, although it is worth noting that the rate now depends on $|\mathcal{Z}|$ as well.

3.4.3 Chain-Constrained Domain

We have seen that Theorem 5.1 can be used to analyze example (3.10) without using the chain-constrained structure. However, it is worth noting that Theorem 5.1 still offers an explicit bound on the feasibility of $x^\star(\xi^{[N]})$ based solely on the chain-constrained structure, although at a slightly worse sample complexity than [8]. In particular, the VC dimension of any chain-constrained domain can be directly bounded as follows.

Lemma 8. *If dom f_ξ has a chain-constrained domain of order m, then the VC dimension of $H = \{H_x\}_{x\in\mathcal{X}}$ in* (3.6) *satisfies*

$$V(H) \le \frac{e}{(e-1)\log 2} m \log\big(\frac{e}{\log 2}\cdot m\big) \sim O(m\log m).$$

Proof. Under the assumption we have dom $f_\xi = \bigcap_{k=1}^m U_k^\xi$ where each $U_k^\xi \in \{U_k^{\xi'}\}_{\xi'\in\Xi}$ is a chain living on $\mathcal{X} \subseteq \mathbb{R}^n$ indexed by $\xi \in \mathbb{R}^r$. Now, for $k \in [m]$, define $W_k^x = \{\xi : x \in U_k^\xi\}$, we have from (3.6) that $H_x = \{\xi : x \in \text{dom} f_\xi\} = \{\xi : x \in \bigcap_{k=1}^m U_k^\xi\} = \bigcap_{k=1}^m W_k^x$. We show, for each $k \in [m]$, $\{W_k^x\}_{x\in\mathcal{X}}$ is a chain as well. Suppose this is not the case, then there exist $x_1, x_2 \in \mathcal{X}$ such that $W_k^{x_1} \not\subseteq W_k^{x_2}$ and $W_k^{x_2} \not\subseteq W_k^{x_1}$. This implies there exist $\xi_1 \in W_k^{x_1}$ and $\xi_2 \in W_k^{x_2}$ such that $\xi_1 \notin W_k^{x_2}$ and $\xi_2 \notin W_k^{x_1}$. This further implies $x_1 \in U_k^{\xi_1}, x_2 \notin U_k^{\xi_1}$ and $x_2 \in U_k^{\xi_2}, x_1 \notin U_k^{\xi_2}$. Consequently, neither $U_k^{\xi_1} \subseteq U_k^{\xi_2}$ nor $U_k^{\xi_2} \subseteq U_k^{\xi_1}$ can be true, contradicting the assumption that $\{U_k^\xi\}_{\xi\in\Xi}$ is a chain.

Thus, $\{W_k^x\}_{x\in\mathcal{X}}$ is a chain on Ξ for each $k \in [m]$ and H_x is a chain-constrained domain of order m.

On the other hand, the VC dimension of a class of sets which are a chain $\{U^\omega\}_{\omega\in I}$ is at most 1 because it cannot shatter any two points. In particular, if $\{x_1, x_2\}$ are two points living on the same space as $\{U^\omega\}_{\omega\in I}$, the shattering of $\{x_1, x_2\}$ requires $x_1 \in U^{\omega_1}, x_2 \notin U^{\omega_1}$ and $x_2 \in U^{\omega_2}, x_1 \notin U^{\omega_2}$ for some $U^{\omega_1}, U^{\omega_2} \in \{U^\omega\}_{\omega\in I}$. If this were to happen, then neither $U^{\omega_1} \subseteq U^{\omega_2}$ nor $U^{\omega_2} \subseteq U^{\omega_1}$ could be true, contradicting the definition of a chain. Then, if $\{\mathcal{U}_k\}_{k\in[m]}$ are the m chains consisting of a chain-constrained domain $\mathcal{U}$ of order m where each $U \in \mathcal{U}$ is of the form $U = \bigcap_{k=1}^m U_k$ for some $U_k \in \mathcal{U}_k$, it again follows from Theorem 1.1 in [99] that

$$V(\mathcal{U}) \leq V(\sqcap_{k=1}^m \mathcal{U}_k) \leq \frac{e}{(e-1)\log 2} m \log(\frac{e}{\log 2} \cdot m),$$

where $\sqcap_{k=1}^m \mathcal{U}_k \triangleq \left\{ \bigcap_{k=1}^m U_k : U_k \in \mathcal{U}_k, k \in [m] \right\}$. The result follows now from the fact that H_x is a chain of order m. □

Lemma 8 combined with Theorem 5.1 can provide an explicit sample complexity for feasibility.

Corollary 3.4.2.4. *If dom f_ξ has a chain-constrained domain of order m, then Theorem 5.1 guarantees that for*

$$N \geq \frac{4}{\alpha}\left(\left(\frac{e}{(e-1)\log 2} m \log\left(\frac{e}{\log 2} \cdot m\right)\right) \log\left(\frac{12}{\alpha}\right) + \log\left(\frac{2}{\delta}\right)\right), \tag{3.20}$$

we have

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta.$$

for any $0 < \delta, \alpha < 1$.

Corollary 3.4.2.4 provides a sample complexity $O(\frac{m}{\alpha}\log m \log(\frac{1}{\alpha}) + \frac{1}{\alpha}\log(\frac{1}{\delta}))$ for chain-constrained domains, or $O(\frac{m}{\alpha}\log m + \frac{1}{\alpha}\log(\frac{1}{\delta}))$ using PAC bounds from [101, 102], while Scenario 1 in [8] provides a $O(\frac{m}{\alpha} + \frac{1}{\alpha}\log(\frac{1}{\delta}))$ bound according to (3.9). As we can see, the more refined analysis on the chain-constrained structure in [8] leads to a better rate over Corollary 3.4.2.4 by log factors. How-

ever, the generality offered by Theorem 5.1 is still evident, since its applicability in most situations either does not hinge on the chain-constrained domain or can be improved by reparametrizations of ξ.

3.4.4 Finite Feasible Region

In this subsection, we apply Theorem 5.1 in the case where the decision set $\mathcal{X}$ is finite.

Corollary 3.4.2.5. *Suppose $|\mathcal{X}| < +\infty$ and let $\xi^{[N]} = \{\xi_1, ..., \xi_N\}$ be IID samples from $\mathbb{P}$ (consequently $\xi^{[N]} \sim \mathbb{P}^N$). Then, if*

$$N \geq \frac{4}{\alpha}\Big(\log_2 |\mathcal{X}| \cdot \log\big(\frac{12}{\alpha}\big) + \log\big(\frac{2}{\delta}\big)\Big), \tag{3.21}$$

we have

$$\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \alpha) \leq \delta$$

for any $0 < \delta, \alpha < 1$.

Proof. Let $H \triangleq \{H_x\}_{x\in\mathcal{X}}$ be the class of subsets defined in (3.6). It follows that $|H| \leq |\mathcal{X}| < +\infty$. It is known that if $|H| < +\infty$, then $VC(H) \leq \log_2 |H|$ (by definition of VC dimension or see [96]). The result then follows from Theorem 5.1. □

Note that since the VC dimension of a finite hypothesis class is bounded by the logarithm of its cardinality, we get the results in Corollary 3.4.2.5 for free. Section 4 of [9] also discusses the case of finite feasible region $\mathcal{X}$, with a slightly different focus. In particular, with assumptions on the moment generating functions, [9] proves exponential convergence of a δ-optimal set towards an ϵ-optimal set using large deviations (LD) theory. The rate of convergence also depends on constants from the LD analysis. A more direct analysis on the feasibility of SAA solution $x^\star(\xi^{[N]})$ which does not rely on distributional assumptions of $f(\xi, x)$ is also availble from Lemma 9 of [9] which states:

$$\mathbb{P}^N(\hat{F}_N(x) < +\infty) \leq (1-\eta)^N, \text{ for } x \in \mathcal{X}^{Infea} \tag{3.22}$$

where $\mathcal{X}^{Infea} = \{x : x \in \mathcal{X} \text{ and } V(x) > 0\}$ and $\eta = \min\{V(x) : x \in \mathcal{X}^{Infea}\}$. Building on (3.22), we can deduce the following direct bound regarding $x^\star(\xi^{[N]})$:

$$\begin{aligned}
\mathbb{P}^N(V(x^\star(\xi^{[N]})) > \eta) &= \mathbb{P}^N(x^\star(\xi^{[N]}) \in \mathcal{X}^{Infea}) \\
&\leq \mathbb{P}^N\Big(\bigcup_{x \in \mathcal{X}^{Infea}} \{\hat{F}_N(x) < +\infty\}\Big) \\
&\leq \sum_{x \in \mathcal{X}^{Infea}} \mathbb{P}^N(\hat{F}_N(x) < +\infty) \leq |\mathcal{X}^{Infea}|(1-\eta)^N, \qquad (3.23)
\end{aligned}$$

which leads to a $O(\frac{1}{\eta}\log(|\mathcal{X}|)+\frac{1}{\eta}\log(\frac{1}{\beta}))$ sample complexity, comparable to the $O(\frac{1}{\eta}\log(\frac{1}{\eta})\log_2(|\mathcal{X}|)+\frac{1}{\eta}\log(\frac{1}{\beta}))$ complexity in (3.21). Moreover, if we utilize the PAC bound from Remark 2, the bound in (3.21) could be improved to $O(\frac{1}{\eta}\log_2(|\mathcal{X}|)+\frac{1}{\eta}\log(\frac{1}{\beta}))$ which is of the same order as (3.23).

Chapter 4: Robust Importance Weighting for Covariate Shift

In many learning problems, the training and testing data follow different distributions and a particularly common situation is the *covariate shift*. To correct for sampling biases, most approaches, including the popular kernel mean matching (KMM), focus on estimating the importance weights between the two distributions. Reweighting-based methods, however, are exposed to high variance when the distributional discrepancy is large and the weights are poorly estimated. On the other hand, the alternate approach of using nonparametric regression (NR) incurs high bias when the training size is limited. In this Chapter, we propose and analyze a new estimator that systematically integrates the residuals of NR with KMM reweighting, based on a control-variate perspective. The proposed estimator can be shown to either strictly outperform or match the best-known existing rates for both KMM and NR, and thus is a robust combination of both estimators. The experiments shows the estimator works well in practice.

4.1 Introduction

Traditional machine learning implicitly assumes training and test data are drawn from the same distribution. However, mismatches between training and test distributions occur frequently in reality. For example, in clinical trials the patients used for prognostic factor identification may not come from the target population due to sample selection bias [114, 115]; incoming signals used for natural language and image processing, bioinformatics or econometric analyses change in distribution over time and seasonality [116, 117, 118, 119, 120, 121, 122]; patterns for engineering controls fluctuate due to the non-stationarity of environments [123, 124].

Many such problems are investigated under the *covariate shift* assumption [125]. Namely, in a supervised learning setting with covariate X and label Y, the marginal distribution of X in

the training set $P_{tr}(x)$, shifts away from the marginal distribution of the test set $P_{te}(x)$, while the conditional distribution $P(y|x)$ remains invariant in both sets. Because test labels are either too costly to obtain or unobserved, it could be uneconomical or impossible to build predictive models only on the test set. In this case, one is obliged to utilize the invariance of conditional probability to adapt or transfer knowledge from the training set, termed as transfer learning [126] or domain adaptation [121, 127]. Intuitively, to correct for covariate shift (i.e., cancel the bias from the training set), one can reweight the training data by assigning more weights to observations where the test data locate more often. Indeed, the key to many approaches addressing covariate shift is the estimation of importance sampling weights, or the Radon-Nikodym derivative (RND) of dP_{te}/dP_{tr} between P_{te} and P_{tr} [128, 129, 130, 131, 132, 133, 134, 119, 123]. Among them is the popular kernel mean matching (KMM) [114, 119], which estimates the importance weights by matching means in a reproducing kernel Hilbert space (RKHS) and can be implemented efficiently by quadratic programming (QP).

Despite the demonstrated efficiency in many covariate shift problems [128, 119, 115], KMM can suffer from high variance, due to several reasons. The first one regards the RKHS assumption. As pointed out in [135], under a more realistic assumption from learning theory [136], when the true regression function does not lie in the RKHS but a general range space indexed by a smoothness parameter $\theta > 0$, KMM degrades to sub-canonical rate $O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}})$ from the parametric rate $O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$. Second, if the discrepancy between the training and testing distributions is large (e.g., test samples concentrate on regions where few training samples are located), the RND becomes unstable and leads to high resulting variance [137], partially due to an induced sparsity as most weights shrink towards zero while the non-zero ones surge to huge values. This is an intrinsic challenge for reweighting methods that occurs even if the RND is known in closed-form. One way to bypass it is to identify model misspecification [138], but as mentioned in [139], the cross-validation for model selection needed in many related methods often requires the importance weights to cancel biases and the necessity for reweighting remains.

In this Chapter we propose a method to reduce the variance of KMM in covariate shift prob-

lems. Our method relies on an estimated regression function and the application of the importance weighting on the *residuals* of the regression. Intuitively, the residuals have smaller magnitudes than the original loss values, and the resulting reweighted estimator is thus less sensitive to the variances of weights. Then, we cancel the bias incurred by the use of residuals by a judicious compensation through the estimated regression function evaluated on the test set.

Our method shares similarities with the Doubly Robust (DR) estimator in causal inference problems [140]. However, different from DR, we do not require semi-parametric estimates of the baseline prediction (corresponding to our regression function g) and conditional probability (corresponding to our importance weight) to both converge at rates $O(n^{\alpha})$ for $\alpha > 1/4$. In particular, we specialize our method by using a nonparametric regression (NR) function constructed from regularized least square in RKHS [136, 141, 142], also known as the Tikhonov regularized learning algorithm [143]. We show that our new estimator achieves the rate $O(n_{tr}^{-\frac{\theta}{2\theta+2}} + n_{te}^{-\frac{\theta}{2\theta+2}})$, which is superior to the best-known rate of KMM in [135], with the same computational complexity of KMM. Although the gap to the parametric rate is yet to be closed, the new estimator certainly seems to be a step towards the right direction. To put into perspective, we also compare with an alternate approach in [135] which constructs an NR function using the training set and then predicts by evaluating on the test set. Such an approach leads to a better dependence on the test size but worse dependence on the training size than KMM. Our estimator, which can be viewed as an ensemble of KMM and NR, achieves a convergence rate that is either superior or matches both of these methods, thus in a sense robust against both estimators. In fact, we show our estimator can be motivated both from a variance reduction perspective on KMM using control variates [144, 145] and a bias reduction perspective on NR.

Another noticable feature of the new estimator relates to data aggregation in empirical risk minimization (ERM). Specifically, when KMM is applied in learning algorithms or ERMs, the resulting optimal solution is typically a finite-dimensional span of the training data mapped into feature space [146]. The optimal solution of our estimator, on the other hand, depends on both the training and testing data, thus highlighting a different and more efficient information leveraging

that utilizes both data sets simultaneously.

The Chapter is organized as follows. Section 2 reviews the background on KMM and NR that motivates our estimator. Section 3 presents the details of our estimator and studies its convergence property. Section 4 generalizes our method to ERM. Section 5 demonstrates experimental results.

4.2 Background and Motivation

Denote P_{tr} to be the probability measure for training variables X^{tr} and P_{te} for test variables X^{te}.

Assumption 2. $P_{tr}(dy|\boldsymbol{x}) = P_{te}(dy|\boldsymbol{x})$.

Assumption 3. *The Radon-Nikodym derivative* $\beta(\boldsymbol{x}) \triangleq \frac{dP_{te}}{dP_{tr}}(\boldsymbol{x})$ *exists and is bounded by* $B < \infty$.

Assumption 4. *The covariate space* $\mathcal{X}$ *is compact and the label space* $\mathcal{Y} \subseteq [0, 1]$. *Furthermore, there exists a kernel* $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *which induces an RKHS* $\mathcal{H}$ *and a canonical feature map* $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ *such that* $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathcal{H}}$ *and* $\|\Phi(\boldsymbol{x})\|_{\mathcal{H}} \leq R$ *for some* $0 < R < \infty$.

Assumption 2 is the covariate shift assumption which states the conditional distribution $P(dy|\boldsymbol{x})$ remains invariant while the marginal $P_{tr}(\boldsymbol{x})$ and $P_{te}(\boldsymbol{x})$ differ. Assumptions 3 and 4 are common for establishing theoretical results. Specifically, Assumption 3 can be satisfied by restricting the support of P_{te} and P_{tr} on a compact set, although B could be potentially large.

4.2.1 Preliminaries and Existing Approaches

Given n_{tr} labelled training data $\{(\boldsymbol{x}_j^{tr}, \boldsymbol{y}_j^{tr})\}_{j=1}^{n_{tr}}$ and n_{te} unlabelled test data $\{\boldsymbol{x}_i^{te}\}_{i=1}^{n_{te}}$ (i.e., $\{y_i^{te}\}_{i=1}^{n_{te}}$ are unavailable), the goal is to estimate $\nu = \mathbb{E}[Y^{te}]$. The KMM estimator [114, 115] is

$$V_{KMM} = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}(\boldsymbol{x}_j^{tr}) y_j^{tr},$$

where $\hat{\beta}(\boldsymbol{x}_j^{tr})$ are solutions of a QP that attempts to match the means of training and test sets in the feature space using weights $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} &\min_{\hat{\boldsymbol{\beta}}}\Big\{\hat{L}(\hat{\boldsymbol{\beta}}) \triangleq \Big\|\frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}}\hat{\beta}_j\Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\Phi(\boldsymbol{x}_i^{te})\Big\|_{\mathcal{H}}^2\Big\} \\ &\text{s.t. } 0 \le \hat{\beta}_j \le B, \forall 1 \le j \le n_{tr}. \end{aligned} \tag{4.1}$$

Notice we write $\hat{\beta}_j$ as $\hat{\beta}(\boldsymbol{x}_j^{tr})$ in V_{KMM} informally to highlight $\hat{\beta}_j$ as estimates of $\beta(\boldsymbol{x}_j^{tr})$. The fact that (4.1) is a QP can be verified by the kernel trick, as in [115]. Indeed, define matrix $K_{ij} = K(\boldsymbol{x}_i^{tr}, \boldsymbol{x}_j^{tr})$ and $\kappa_j \triangleq \frac{n_{tr}}{n_{te}}\sum_{i=1}^{n_{te}} K(\boldsymbol{x}_j^{tr}, \boldsymbol{x}_i^{te})$, optimization (4.1) is equivalent to

$$\begin{aligned} &\min_{\hat{\boldsymbol{\beta}}} \quad \frac{1}{n_{tr}^2}\hat{\boldsymbol{\beta}}^T\boldsymbol{K}\hat{\boldsymbol{\beta}} - \frac{2}{n_{tr}^2}\boldsymbol{\kappa}^T\hat{\boldsymbol{\beta}}, \\ &\text{s.t. } 0 \le \hat{\beta}_j \le B, \forall 1 \le j \le n_{tr}. \end{aligned} \tag{4.2}$$

In practice, a constraint $\left|\frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}}\hat{\beta}_j - 1\right| \le \epsilon$ for a tolerance $\epsilon > 0$ is included to regularize the $\hat{\boldsymbol{\beta}}$ towards the RND. As in [135], we omit them to simplify analysis. On the other hand, the NR estimator

$$V_{NR} = \frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{g}(\boldsymbol{x}_i^{te}),$$

is based on $\hat{g}(\cdot)$, some estimate of the regression function $g(\boldsymbol{x}) \triangleq \mathbb{E}[Y|\boldsymbol{x}]$. Notice the conditional expectation is taken regardless of $\boldsymbol{x} \sim P_{tr}$ or P_{te}. Here, we consider a $\hat{g}(\cdot)$ that is estimated nonparametrically by regularized least square in RKHS:

$$\hat{g}_{\gamma,data}(\cdot) = \operatorname*{argmin}_{f\in\mathcal{H}}\Big\{\frac{1}{m}\sum_{j=1}^{m}(f(\boldsymbol{x}_j^{tr}) - y_j^{tr})^2 + \gamma\|f\|_{\mathcal{H}}^2\Big\}, \tag{4.3}$$

where γ is a regularization term to be chosen and the subscript $data$ represents $\{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^m$. Using the representation theorem [146], optimization problem (4.3) can be solved in closed form

with $\hat{g}_{\gamma,data}(\boldsymbol{x}) = \sum_{j=1}^{m} \alpha_j^{reg} K(\boldsymbol{x}_j^{tr}, \boldsymbol{x})$ where

$$\boldsymbol{\alpha}^{reg} = (\boldsymbol{K} + \gamma \boldsymbol{I})^{-1} \boldsymbol{y}^{tr}, \tag{4.4}$$

and $\boldsymbol{y}^{tr} = [y_1^{tr}, ..., y_m^{tr}]$.

4.2.2 Motivation

Depending on properties of $g(\cdot)$, [135] proves different rates of KMM. The most notable case is when $g \notin \mathcal{H}$ but rather $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, where $\mathcal{T}_K$ is the integral operator $(\mathcal{T}_K f)(x') = \int_{\mathcal{X}} K(x', x) f(x) P_{tr}(dx)$ on $\mathscr{L}^2_{P_{tr}}$. Here, one can imagine θ as a smoothness parameter in measuring the space of functions $g(\cdot)$ lives in. The higher θ is, the more smooth g is. In the extreme cases that $\theta \to \infty$, we know that $Range(\mathcal{T}_K^{1/2} \subseteq \mathcal{H}$. In this case, [135] characterize g with the approximation error

$$\mathcal{A}_2(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq F} \|g - f\|_{\mathscr{L}^2_{P_{tr}}} \leq C F^{-\frac{\theta}{2}}, \tag{4.5}$$

and the rates of KMM drops to sub-canonical $|V_{KMM} - \nu| = O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}})$, as opposed to $O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$ when $g \in \mathcal{H}$. As shown in Lemma 4 in the Supplementary and Theorem 4.1 of [136]), (4.5) is almost equivalent to $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$: $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ implies (4.5) while (4.5) leads to $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}-\epsilon})$ for any $\epsilon > 0$. We adopt the characterization $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ as our analysis is based on related learning theory estimates. In particular, our proofs rely on these estimates and are different from [135]. For example, in (4.3), γ is used as a free parameter for controlling $\|f\|_{\mathcal{H}}$, whereas [135] uses the parameter F in (4.5). Although the two approaches are equivalent from an optimization viewpoint, with γ being the Lagrange dual variable, the former approach turns out to be more suitable to our analysis.

Correspondingly, the convergence rate for V_{NR} when $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ is also shown in [135] as $|V_{NR} - \nu| = O(n_{te}^{-\frac{1}{2}} + n_{tr}^{-\frac{3\theta}{12\theta+16}})$, with $\hat{g}$ taken as $\hat{g}_{\gamma,data}$ in (4.3) and γ chosen optimally. The rate of V_{KMM} is usually better than V_{NR} due to labelling cost (i.e. $n_{tr} < n_{te}$). However, in practice the performance of V_{KMM} is not always better than V_{NR}. This could be partially explained

by the hidden dependence of V_{KMM} on potentially large B, but more importantly, without variance reduction, KMM is subject to the negative effects of unstable importance sampling weights (i.e. the $\hat{\boldsymbol{\beta}}$). On the other hand, the training of $\hat{g}$ requires labels hence can only be done on training set. Consequently, without reweighting, when estimating the test quantity ν, the rate of V_{NR} suffers from the bias.

This motivates the search for a robust estimator which does not require prior knowledge on the performance of V_{KMM} or V_{NR} and can, through a combination, reach or even surpass the best performance among both. For simplicity, we use the mean squared error (MSE) criterion $\mathrm{MSE}(V) = \mathrm{Var}(V) + (\mathrm{Bias}(V))^2$ and assume an additive model $Y = g(X) + \mathcal{E}$ where $\mathcal{E} \sim \mathcal{N}(0, \sigma^2)$ is independent with X and other errors. Under this framework, we motivate a remedy from two perspectives:

Variance Reduction for KMM: Consider an idealized KMM with $V_{KMM} \triangleq \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr}) y_j^{tr}$ and $\beta(\cdot)$ being the true RND. Since

$$\mathbb{E}[\beta(X^{tr})Y^{tr}] = \mathbb{E}_{\boldsymbol{x}\sim P_{tr}}(\beta(\boldsymbol{x})g(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})] = \nu,$$

V_{KMM} is unbiased and the only source of MSE becomes the variance. It then follows from standard control variates that, given an estimator V and a zero-mean random variable W, we can set $t^\star = \frac{\mathrm{Cov}(V,W)}{\mathrm{Var}(W)}$ and use $V - t^\star W$ to obtain

$$\min_t \mathrm{Var}(V - tW) = (1 - \mathrm{corr}^2(V, W))\mathrm{Var}(V) \leq \mathrm{Var}(V),$$

without altering the mean of V. Thus we can use

$$W = \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(\hat{g}(\boldsymbol{x}_j^{tr})) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$$

with $t^\star = \frac{\mathrm{Cov}(V_{KMM}, W)}{\mathrm{Var}(W)}$. To calculate $t^\star$, suppose X^{te} and X^{tr} are independent, then we have

$$\begin{aligned}\mathrm{Cov}(V_{KMM}, W) =& \frac{1}{n_{tr}}\mathrm{Cov}(\beta(X^{tr})Y^{tr}, \beta(X^{tr})\hat{g}(X^{tr})) \\ =& \frac{1}{n_{tr}}\mathrm{Cov}(\beta(X^{tr})g(X^{tr}), \beta(X^{tr})\hat{g}(X^{tr})) \\ \approx& \frac{1}{n_{tr}}\mathrm{Var}(\beta(X^{tr})\hat{g}(X^{tr})),\end{aligned}$$

if $\hat{g}$ is close enough to g. On the other hand, in the usual case where $n_{te} \gg n_{tr}$,

$$\begin{aligned}\mathrm{Var}(W) =& \frac{1}{n_{tr}}\mathrm{Var}(\beta(X^{tr})\hat{g}(X^{tr})) + \frac{1}{n_{te}}\mathrm{Var}(\hat{g}(X^{te})) \\ \approx& \frac{1}{n_{tr}}\mathrm{Var}(\beta(X^{tr})\hat{g}(X^{tr})).\end{aligned}$$

Thus, $t^\star \approx 1$ which gives our estimator

$$V_R = \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}).$$

Bias Reduction for NR: Consider the NR estimator $V_{NR} \triangleq \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te})$. Assuming again the common case where $n_{te} \gg n_{tr}$, we have

$$\mathrm{Var}(V_{NR}) = \frac{1}{n_{te}}\mathrm{Var}(\hat{g}(X^{te})) \approx 0,$$

and the main source of MSE is bias $\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x})]$. If we add $W = \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))$ to V_{NR}, we eliminate the bias which gives the same estimator

$$V_R = \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}).$$

4.3 Robust Estimator

We construct a new estimator $V_R(\rho)$ that can be shown to perform robustly against both KMM and NR estimators discussed above. In our construction, we split the training set with a proportion $\rho \in [0,1]$, i.e., divide $\{\boldsymbol{X}^{tr}, \boldsymbol{Y}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^{n_{tr}}$ into

$$\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{Y}_{KMM}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=1}^{\lfloor \rho n_{tr} \rfloor},$$

and

$$\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}_{data} \triangleq \{(\boldsymbol{x}_j^{tr}, y_j^{tr})\}_{j=\lfloor \rho n_{tr} \rfloor + 1}^{n_{tr}},$$

where $\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{X}^{te}\}_{data} \triangleq \{\{\boldsymbol{x}_j^{tr}\}_{j=1}^{\lfloor \rho n_{tr} \rfloor}, \{\boldsymbol{x}_i^{te}\}_{i=1}^{n_{te}}\}$ is used to solve for the weight $\hat{\boldsymbol{\beta}}$ in (4.1) and $\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}_{data}$ is used to train an NR function $\hat{g}(\cdot) = \hat{g}_{\gamma,data}(\cdot)$ for some γ as in (4.3). Finally, we define our estimator $V_R(\rho)$ as

$$\begin{aligned} V_R(\rho) \triangleq & \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr})) \\ & + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}). \end{aligned} \tag{4.6}$$

First, we remark the parameter ρ controlling the splitting of data serves mainly for theoretical considerations. In practice, the data can be used for both purposes simultaneously. Second, as mentioned, many $\hat{g}$ other than (4.3) could be considered for control variate. However, aside from the availability of closed-form expression (4.4), $\hat{g}_{\gamma,data}$ is connected to the learning theory estimates [136]. Thus, for establishing a theoretical bound, we focus on $\hat{g} = \hat{g}_{\gamma,data}$ for now.

Our main result is the convergence analysis with respect to n_{tr} and n_{te} which rigorously justified the previous intuition. In particular, we show that V_R either surpasses or achieves the better rate between V_{KMM} and V_{NR}. In all theorems that follow, the big-O notations can be interpreted either as $1-\delta$ high probability bound or a bound on expectation. The proofs are left in the Supplementary.

Theorem 4.3.1. *Under Assumptions 2-4, if we assume $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, the convergence rate*

of $V_R(\rho)$ satisfies

$$|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{2\theta+2}} + n_{te}^{-\frac{\theta}{2\theta+2}}), \tag{4.7}$$

when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ in (4.6) with $\gamma = n^{-\frac{\theta+2}{\theta+1}}$ and $n \triangleq \min(n_{tr}, n_{te})$.

Under the same setting of Theorem 4.3.1, if we choose $\gamma = n^{-1}$, we have

$$|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}}) \tag{4.8}$$

and if we choose $\gamma = n_{tr}^{-1}$,

$$|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}}). \tag{4.9}$$

We remark several implications. First, although not achieving canonical, (4.7) is an improvement over the best-known $O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{\theta}{2\theta+4}})$ rate of V_{KMM} when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, especially for small θ, suggesting that V_R is more suitable than V_{KMM} when g is irregular. Indeed, θ is a smoothness parameter that measures the regularity of g. When θ increases, functions in $Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ get smoother and $Range(\mathcal{T}_K^{\frac{\theta_2}{2\theta_2+4}}) \subseteq Range(\mathcal{T}_K^{\frac{\theta_1}{2\theta_1+4}})$ for $0 < \theta_1 < \theta_2$, with the limiting case that $\theta \to \infty$, $\frac{\theta}{2\theta+4} \to 1/2$ and $Range(\mathcal{T}_K^{\frac{1}{2}}) \subseteq \mathcal{H}$ (i.e. $g \in \mathcal{H}$) for universal kernels by Mercer's theorem.

Second, as in Theorem 4 of [135], the optimal tuning of γ that leads to (4.7) depends on the unknown parameter θ, which may not be adaptive in practice. However, if one simply choose $\gamma = n^{-1}$, V_R still achieves a rate no worse than V_{KMM} as depicted in (4.8).

Third, also in Theorem 4 of [135], the rate of V_{NR} is $O(n_{te}^{-\frac{1}{2}} + n_{tr}^{-\frac{3\theta}{12\theta+16}})$ when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, which is better on n_{te} but not n_{tr}. Since usually $n_{tr} < n_{te}$, the rate of V_{KMM} generally excels. Indeed, in this case the rate of V_{NR} beats V_{KMM} only if $\lim_{n\to\infty} n_{te}^{\frac{6\theta+8}{3\theta+6}}/n_{tr} \to 0$. However, if so, V_R can still achieve $O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}})$ rate in (4.9) which is better than V_{NR}, by simply taking $\gamma = n_{tr}^{-1}$, i.e., regularizing the training process more when the test set is small. Moreover, as $\theta \to \infty$, our estimator V_R recovers the canonical rate $n_{tr}^{-\frac{1}{2}}$ as opposed to $n_{tr}^{-\frac{1}{4}}$ in V_{NR}.

Thus, in summary, when $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, our estimator V_R outperforms both V_{KMM} and V_{NR} across the relative sizes of n_{tr} and n_{te}. The outperformance over V_{KMM} is strict when γ is chosen dependent on θ, and the performance is matched when γ is chosen robustly without knowledge of θ.

For completeness, we consider two other characterizations of g discussed in [135]: one is $g \in \mathcal{H}$ and the other is $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq F} \|g - f\| \leq C(\log F)^{-s}$ for some $C, s > 0$ (e.g., $g \in H^s(\mathcal{X})$ with $K(\cdot, \cdot)$ being the Gaussian kernel, where H^s is the Sobolev space with integer s). The two assumptions are, in a sense, more extreme (being optimistic or pessimistic). The next two results show that the rates of V_R in these situations match the existing ones for V_{KMM} (the rates for V_{NR} are not discussed in [135] under these assumptions).

Proposition 4. *Under Assumptions 2-4, if $g \in \mathcal{H}$, the convergence rate of $V_R(\rho)$ satisfies $|V_R(\rho) - \nu| = O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}})$, when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ for $\gamma > 0$ in (4.6).*

Proposition 5. *Under Assumptions 2-4, if $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq F} \|g - f\| \leq C(\log F)^{-s}$ for some $C, s > 0$, the convergence rate of $V_R(\rho)$ satisfies $|V_R(\rho) - \nu| = O\left(\log \frac{n_{tr} n_{te}}{n_{tr}+n_{te}}\right)^{-s}$, when $\hat{g}$ is taken to be $\hat{g}_{\gamma,data}$ for $\gamma > 0$ in (4.6).*

4.4 Empirical Risk Minimization

The robust estimator can handle empirical risk minimization (ERM). Given loss function $l'(x, y; \theta) : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ given θ in $\mathcal{D}$, we optimize over

$$\min_{\theta \in \mathcal{D}} \mathbb{E}[l'(X^{te}, Y^{te}; \theta)] = \min_{\theta \in \mathcal{D}} \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[l(\boldsymbol{x}; \theta)],$$

where $l(\boldsymbol{x}; \theta) \triangleq \mathbb{E}_{Y|\boldsymbol{x}}[l'(\boldsymbol{x}, Y; \theta)]$ to find

$$\theta^\star \triangleq \operatorname*{argmin}_{\theta \in \mathcal{D}} \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[l(X^{te}; \theta)].$$

In practice, usually a regularization term $\Omega[\theta]$ on θ is added. For example, the KMM in [114] considers

$$\min_{\theta\in\mathcal{D}} \frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} \hat{\beta}(\boldsymbol{x}_j^{tr}) l'(\boldsymbol{x}_j^{tr}, y_j^{tr}; \theta) + \lambda\Omega[\theta]. \tag{4.10}$$

We can carry out a similar modification for V_R:

$$\begin{aligned}\min_{\theta\in\mathcal{D}} \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(l'(\boldsymbol{x}_j^{tr}, y_j^{tr}; \theta) - \hat{l}(\boldsymbol{x}_j^{tr}; \theta)) \\ + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te}; \theta) + \lambda\Omega[\theta],\end{aligned} \tag{4.11}$$

with $\hat{\boldsymbol{\beta}}$ based on $\{\boldsymbol{X}_{KMM}^{tr}, \boldsymbol{X}^{te}\}$ and $\hat{l}(x;\theta)$ being an estimate of $l(x;\theta)$ based on $\{\boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}\}$. For later reference, we note that a similar modification can also be used on V_{NR}:

$$\min_{\theta\in\mathcal{D}} \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te}; \theta) + \lambda\Omega[\theta]. \tag{4.12}$$

We discuss two classical learning problems by (4.11).

Penalized Least Square Regression: Consider a regression problem with $l'(\boldsymbol{x}, y; \theta) = (y - \langle\theta, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}})^2$, $\Omega[\theta] = \|\theta\|_{\mathcal{H}}^2$ and $y \in [0, 1]$. We have

$$l(\boldsymbol{x};\theta) = \mathbb{E}[Y^2|\boldsymbol{x}] - 2g(\boldsymbol{x})\langle\theta, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}} + \langle\theta, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}}^2,$$

and a candidate for $\hat{l}(\boldsymbol{x}, \theta)$ is to substitute g with $\hat{g}_{\gamma,data}$. Then, (4.11) becomes

$$\begin{aligned}\min_{\theta\in\mathcal{D}} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} -\frac{2\beta(\boldsymbol{x}_j^{tr})}{\lfloor \rho n_{tr} \rfloor}(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))\langle\theta, \Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}} \\ + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} (\hat{g}(\boldsymbol{x}_i^{te}) - \langle\theta, \Phi(\boldsymbol{x})\rangle_{\mathcal{H}})^2 + \lambda\|\theta\|_{\mathcal{H}}^2,\end{aligned}$$

by adding and removing the components not involving θ. Furthermore, it simplifies to the QP:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\lfloor\rho n_{tr}\rfloor+n_{te}}} \frac{-2_1^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{\lfloor\rho n_{tr}\rfloor} + \lambda\boldsymbol{\alpha}^T\boldsymbol{K}_{tot}\boldsymbol{\alpha} + \frac{(2-\boldsymbol{K}_{tot}\boldsymbol{\alpha})_3^T(2-\boldsymbol{K}_{tot}\boldsymbol{\alpha})}{n_{te}}, \tag{4.13}$$

by the representation theorem [146]. Here $(\boldsymbol{K}_{tot})_{ij} = K(\boldsymbol{x}_i^{tot}, \boldsymbol{x}_j^{tot})$ and $_3 = \text{diag}(_3)$ where $\boldsymbol{x}_i^{tot} = \boldsymbol{x}_i^{tr}$, $(w_1)_i = \beta(\boldsymbol{x}_i^{tr})(y_i^{tr} - \hat{g}(\boldsymbol{x}_i^{tr}))$, $(w_2)_i = 0$, $(w_3)_i = 0$ for $1 \le i \le \lfloor\rho n_{tr}\rfloor$ and $\boldsymbol{x}_i^{tot} = \boldsymbol{x}_{i-\lfloor\rho n_{tr}\rfloor}^{te}$, $(w_1)_i = 0$, $(w_2)_i = \hat{g}(\boldsymbol{x}_{i-\lfloor\rho n_{tr}\rfloor}^{te})$, $(w_3)_i = 1$ for $\lfloor\rho n_{tr}\rfloor + 1 \le i \le \lfloor\rho n_{tr}\rfloor + n_{te}$. Notice (4.13) has a closed-form solution

$$\hat{\boldsymbol{\alpha}} = (_3\boldsymbol{K}_{tot} + \lambda n_{te}\boldsymbol{I})^{-1}(\frac{n_{te}}{\lfloor\rho n_{tr}\rfloor}{}_1 + _2).$$

Penalized Logistic Regression: Consider a binary classification problem with $y \in \{0, 1\}$, $\Omega[\theta] = \|\theta\|_{\mathcal{H}}^2$ and $-l'(\boldsymbol{x}, y; \theta) = y\log(\frac{1}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}) + (1-y)\log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}})$. Thus, we have

$$-l(\boldsymbol{x};\theta) = -g(\boldsymbol{x})\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}} + \log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x})\rangle_{\mathcal{H}}}),$$

and we can again substitute g with $\hat{g}_{\gamma,data}$. Then, (4.11) becomes

$$\begin{aligned}\min_{\theta\in\mathcal{D}} &\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\frac{\beta(\boldsymbol{x}_j^{tr})}{\lfloor\rho n_{tr}\rfloor}(y_j^{tr} - \hat{g}(\boldsymbol{x}_j^{tr}))\langle\theta,\Phi(\boldsymbol{x}_j^{tr})\rangle_{\mathcal{H}}\\ &+\frac{1}{n_{te}}\sum_{i=1}^{n_{te}} -\hat{g}(\boldsymbol{x}_i^{te})\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}} + \lambda\|\theta\|_{\mathcal{H}}^2\\ &+\log(\frac{\exp\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}}}{1+\exp\langle\theta,\Phi(\boldsymbol{x}_i^{te})\rangle_{\mathcal{H}}}).\end{aligned}$$

which again simplifies to, by [146], the convex program:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{\lfloor\rho n_{tr}\rfloor+n_{te}}} \frac{{}_1^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{\lfloor\rho n_{tr}\rfloor} - \frac{{}_2^T\boldsymbol{K}_{tot}\boldsymbol{\alpha}}{n_{te}} + \lambda\boldsymbol{\alpha}^T\boldsymbol{K}_{tot}\boldsymbol{\alpha} + \frac{\sum_{i=1}^{n_{te}}\log(\frac{\exp(\boldsymbol{K}_{tot}\boldsymbol{\alpha})_{\lfloor\rho n_{tr}\rfloor+i}}{1+\exp(\boldsymbol{K}_{tot}\boldsymbol{\alpha})_{\lfloor\rho n_{tr}\rfloor+i}})}{n_{te}}. \tag{4.14}$$

Both (4.13) and (4.14) can be optimized efficiently by standard solvers. Notably, derived from (4.11), an optimal solution is in the form $\hat{\theta} = \sum_{i=1} \hat{\alpha}_i K(\boldsymbol{x}_i^{tot}, \boldsymbol{x})$ which spans on both training and test data. In contrast, the solution of (4.10) or (4.12) only spans on one of them. For example, as shown in [114], the penalized least square solution for (4.10) is $\hat{\theta} = \sum_{i=1} \hat{\alpha}_i K(\boldsymbol{x}_i^{tr}, \boldsymbol{x})$ where

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + n_{te}\lambda \operatorname{diag}(\hat{\boldsymbol{\beta}})^{-1})^{-1}\boldsymbol{y}^{tr}$$

(we use $\hat{\boldsymbol{\alpha}} = (\operatorname{diag}(\hat{\boldsymbol{\beta}})\boldsymbol{K} + n_{te}\lambda \boldsymbol{I})^{-1} \operatorname{diag}(\hat{\boldsymbol{\beta}})\boldsymbol{y}^{tr}$ in experiments to avoid invertibility issues caused by the sparsity of $\hat{\boldsymbol{\beta}}$), so only the training data are in the span of the feature space that constitutes $\hat{\theta}$. The aggregation of both sets suggests a more effective utilization of data . We conclude with a theorem on ERM similar to Corollary 8.9 in [115], which guarantees the convergence of the solution of (4.11) in a simple setting.

Theorem 4.4.1. *Assume $l(x;\theta)$ and $\hat{l}(x;\theta) \in \mathcal{H}$ can be expressed as $\langle \Phi(x), \theta \rangle_{\mathcal{H}} + f(x;\theta)$ with $||\theta||_{\mathcal{H}} \leq C$ and $l'(x,y;\theta) \in \mathcal{H}$ as $\langle \Upsilon(x,y), \Lambda \rangle_{\mathcal{H}} + f(x;\theta)$ with $||\Lambda||_{\mathcal{H}} \leq C$. Denote this class of loss functions $\mathcal{G}$ and further assume $l(x;\theta)$ are continuous, bounded by D and L-Lipschitz on θ uniformly over x for (θ, x) in a compact set $\mathcal{D} \times \mathcal{X}$. Then, the ERM with*

$$\begin{aligned} V_R(\theta) \triangleq & \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(l'(\boldsymbol{x}_j^{tr}, y_j^{tr}; \theta) - \hat{l}(\boldsymbol{x}_j^{tr}; \theta)) \\ & + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{l}(\boldsymbol{x}_i^{te}; \theta) \end{aligned}$$

and $\hat{\theta}_R \triangleq \operatorname{argmin}_{\theta \in \mathcal{D}} V_R(\theta)$ satisfies

$$\mathbb{E}[l'(X_{te}, Y_{te}; \hat{\theta}_R)] \leq \mathbb{E}[l'(X_{te}, Y_{te}; \theta^\star)] + O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}).$$

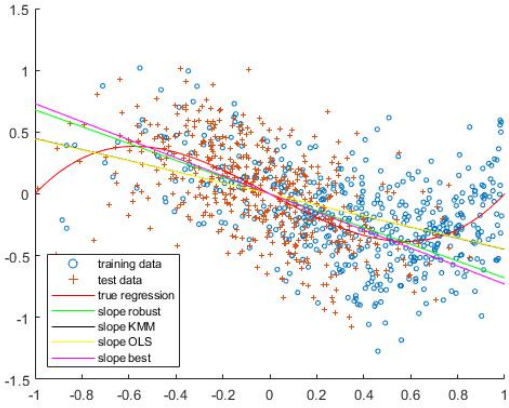
4.5 Experiments

4.5.1 Toy Dataset Regression

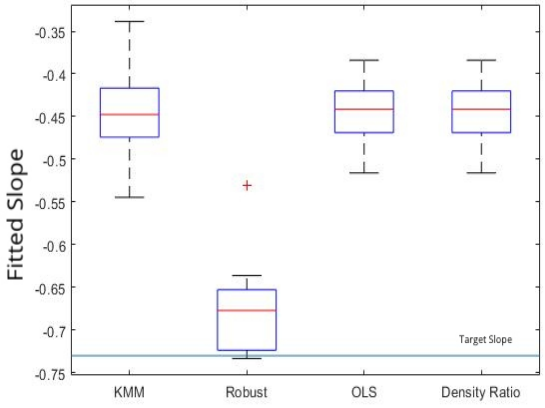
We first present a toy example to provide comparison with KMM. The data is generated as the polynomial regression example in [125, 114], where $P_{tr} \sim \mathcal{N}(0.5, 0.5^2)$, $P_{te} \sim \mathcal{N}(0, 0.3^2)$ are Gaussian distributions. The labels are generated according to $y = -x + x^3$ and observed with Gaussian noise $\mathcal{N}(0, 0.3^2)$. We sample 500 points in both training and test data and fit a linear model using ordinary least square (OLS), KMM and our robust estimator, respectively. On the population level, the best linear fit is $y = -0.73x$ (i.e. $_{\alpha_0,\beta_0}\mathbb{E}_{x\sim P_{te}}(Y - (\alpha_0 x + \beta_0))^2$ is $\alpha_0 = -0.73, \beta_0 = 0$). For simplicity, we set the intercept $\beta_0 = 0$ as known and compare the fitted slopes for different estimators. We use a degree-3 polynomial kernel and set γ in $\hat{g}_{\gamma,data}$ to the default value n_{tr}^{-1}. The tolerance ϵ for $\hat{\boldsymbol{\beta}}$ is set similarly as in [114] with a slight tuning to avoid an overly sparse solution. The slope is fitted without regularization. In Figure 1(a), the red curve is the true polynomial regression function and the purple line is the best linear fit. The blue circle is the training data and the orange cross is the test data. For three different approaches, as well as an additional density-ratio-based method in [125], the fitted slope over 20 trials are summarized in Figure 1(b). The average value is plotted in Figure 1(a) with black (KMM), green (robust) and yellow (OLS) respectively. As we see, the robust estimator outperforms the two other methods, achieving higher accuracy than KMM and unweighted OLS and recovering the slope closest to the best one in the vast majority of trials.

4.5.2 Real World Dataset for ERM

Next, we test our approach in ERM on a real world dataset, the breast cancer dataset from the UCI Archive. We consider the second biased sampling scheme in [114] where the sampling bias operates jointly across multiple features. In particular, after randomly splitting the training and test sets based on different proportions, the training set is further subsampled with probability of selecting $\boldsymbol{x}_i$ in the training set proportional to $\exp(-\sigma_1 \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|)$ for some $\sigma_1 > 0$ and the



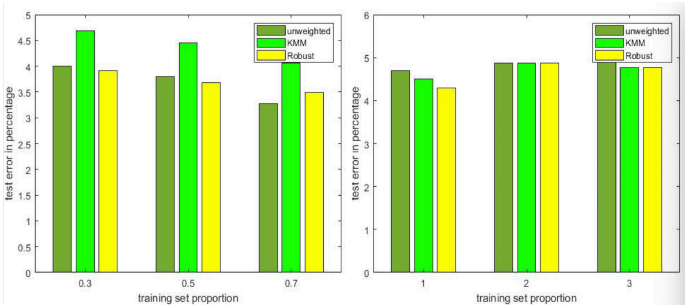
(a)



(b)

Figure 4.1: (a): Linear fit with OLS, KMM and robust estimator; (b): Boxplot on slope estimation

training sample mean $\bar{\boldsymbol{x}}$. Since this is a binary classification problem and we are interested in comparing different approaches, we experiment with both the penalized least square regression and the penalized logistic regression for training sets of several sizes, i.e., the proportions of the training data are 0.3, 0.5, and 0.7 respectively, with respect to the total data. We used a Gaussian kernel $\exp(-\sigma_2\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$ for some $\sigma_2 > 0$. The tolerance ϵ for $\hat{\boldsymbol{\beta}}$ is set exactly as in [114]. For both experiments, we choose parameters $\gamma = n_{tr}^{-1}$ as default, $\lambda = 5$ by cross-validation and $\sigma_1 = -1/100$, $\sigma_2 = \sqrt{0.5}$. Finally, we used the fitted parameters (i.e., optimal solution $\hat{\theta}$ in ERM) to predict the labels on the test set and compare with the hidden real ones. The summary of test error comparison is shown in Figure 2 where we use the term *unweighted* to denote the case for (4.12), *KMM* for (4.10) and *Robust* for (4.11). The robust estimator gives the lowest test error in



(a) (b)

Figure 4.2: Classification performance for (a): penalized least square regression; (b) penalized logistic regression

5 cases out of 6 and follows KMM closely in the exceptional case, confirming our finding on its improvement over the traditional methods.

4.5.3 Simulated Dataset for Estimation

To test the performace of robust estimator on an estimation problem, we simulate data from two ten-dimensional Gaussian distributions with different, randomly generated means and covariance matrices as training and test sets. The target value is $\nu = \mathbb{E}_{\boldsymbol{x}\sim P_{te}}[g(\boldsymbol{x})]$ for an artificially constructed regression function $g(x) = \sin(c_1\|\boldsymbol{x}\|_2^2)+(1+\exp(\frac{T}{2}\boldsymbol{x}))^{-1}$ with random $c_{1,2}$ and labels are observed with Gaussian noise. The Gaussian kernel $\exp(-\sigma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$ for $\sigma > 0$ and a tolerance ϵ for $\hat{\boldsymbol{\beta}}$ are set with exactly the same parameters as in [115] with $\sigma = \sqrt{5}$, $B = 1000$ and $\epsilon = \frac{\sqrt{n_{tr}}-1}{\sqrt{n_{tr}}}$. We also experiment with a different $\hat{g}$ by substituting $\hat{g}_{\gamma,data}$ for a naive linear OLS fit with a lasso regularization term $\lambda > 0$. At each iteration, we use the sample mean from 10^6 data points (without adding noise) as the true mean and calculate the average MSE over 100 estimations for V_R, V_{KMM} and V_{NR} respectively. As shown in Table 1, the performances of V_R are again consistently on par with the best case scenarios, even when the form of $\hat{g}_{\gamma,data}$ is replaced with a naive OLS fit, suggesting the robust estimator still works well under other forms of control variate functions. Moreover, we see that the robust estimator exhibits satisfactory performance even when the usual assumption $n_{tr} < n_{te}$ is violated.

Table 4.1: Average MSE for Estimation

Hyperparameters	**MSE**		
$(\lambda, n_{tr}, n_{te})$	V_{NR}	V_{KMM}	V_R
$(0.1, 50, 500)$	0.9970	0.9489	**0.9134**
$(0.1, 500, 500)$	1.0006	**0.9294**	0.9340
$(0.1, 500, 50)$	1.0021	0.9245	**0.9242**
$(10, 50, 500)$	0.9962	0.9493	**0.9467**
$(10, 500, 500)$	0.9964	0.9294	**0.9288**
$(10, 500, 50)$	0.9965	**0.9245**	0.9293

4.6 Conclusion

Motivated from variance and bias reduction, we introduced a new robust estimator for covariate shift problems which leads to improved accuracy over both KMM and NR in different settings. From a practical standpoint, the control variates and data aggregation enable the estimation/training process to be more stable and data-efficient at no expense of significant computational complexity increase. From an analytical standpoint, when the regression function lies in range spaces outside of RKHS, a promising progress is made to improve upon the well-known rate gap of KMM towards the parametric. For future work, note the canonical rate is still not achieved and it remains unclear the suitable tools for further improvement, if possible at all. Moreover, outside the KMM context with the regularized empirical regression function in RKHS, establishing the eligibility and effectiveness of other reweighting method coupled with different regression functions from learning schemes requires rigorous analysis.

4.7 Supplementary

Throughout the proofs, $h(\cdot) \in \mathcal{H}$ is assumed to be an unspecified function in the RKHS. Also, we use $\mathbb{E}_X[\cdot]$ to denote expectation over the randomness of X while fixing others and $\mathbb{E}_{|X}[\cdot]$ as the conditional expectation $\mathbb{E}[\cdot|X]$. Moreover we remark that all results involving $\hat{g}_{\gamma,data}$ can be interpreted either as a high probability bound or a bound on expectation over $\mathbb{E}_{data}$ (i.e., if we train

$\hat{g}_{\gamma,\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$ using $\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}$, then $\mathbb{E}_{data}$ means $\mathbb{E}_{\boldsymbol{X}^{tr}_{NR},\boldsymbol{Y}^{tr}_{NR}}$). The same interpretation applies for the results with Big-O notations. Finally, constants C_2, C_2', C_3, C_3' and C_3'' as well as similar constants introduced later which depend on $R, g(\cdot)$ or δ (for $1-\delta$ high probability bound) will sometimes be denoted by a common C during the proofs for ease of presentation.

4.7.1 Preliminaries

Lemma 9. *Under Assumption 3, for any $f \in \mathcal{H}$, we have*

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |\langle f(\cdot), \Phi(\cdot, x)\rangle_{\mathcal{H}}| \leq R\|f\|_{\mathcal{H}}. \tag{4.15}$$

and consequently $\|f\|_{\mathscr{L}^2_{P_{tr}}} \leq R\|f\|_{\mathcal{H}}$ as well.

Lemma 10 (Azuma-Hoeffding)**.** *Let $X_1, ..., X_n$ be independent and identically distributed random variables with $0 \leq X \leq B$, then*

$$P(|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i - \mathbb{E}[X]| > \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{B^2}}. \tag{4.16}$$

Under the same assumption of Lemma 10, with probability at least $1-\delta$,

$$|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i - \mathbb{E}[X]| \leq B\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}. \tag{4.17}$$

Moreover, an important $(1-\delta)$-probability bound we shall use later for $\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}^{tr}_1,...,\boldsymbol{x}^{tr}_{n_{tr}}})$ follows from [135] (see also [115] and [147]):

$$\begin{aligned}\hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}^{tr}_1,...,\boldsymbol{x}^{tr}_{n_{tr}}}) =& \left\| \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} \beta(\boldsymbol{x}^{tr}_j)\Phi(\boldsymbol{x}^{tr}_j) - \frac{1}{n_{te}}\sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}^{te}_i) \right\|_{\mathcal{H}} \\ \leq& \sqrt{2\log\frac{2}{\delta}} R \sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}}\right)}. \end{aligned} \tag{4.18}$$

4.7.2 Learning Theory Estimates

To adopt the more realistic assumption as in [135, 136] that the true regression function $g(\cdot) \notin \mathcal{H}$ but rather $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, we need results from learning theory.

First, define $\zeta \triangleq \frac{\theta}{2\theta+4}$ for some $\theta > 0$ so that $0 < \zeta < 1/2$. Given $g(\cdot) \in Range(\mathcal{T}_K^{\zeta})$ and m training sample $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^m$ (sampled from P_{tr})), we define $g_\gamma(\cdot) \in \mathcal{H} : \mathcal{X} \to \mathbb{R}$ to be

$$g_\gamma(\cdot) = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \|f - g\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|f\|^2_{\mathcal{H}} \right\} \tag{4.19}$$

where $\|f - g\|_{\mathscr{L}^2_{P_{tr}}} = \sqrt{\mathbb{E}_{\boldsymbol{x} \sim P_{tr}} (f(\boldsymbol{x}) - g(\boldsymbol{x}))^2}$ denotes the $\mathscr{L}^2$ norm under P_{tr}. On the other hand, $\hat{g}_{\gamma,data}(\cdot) \in \mathcal{H}$ is defined in (3)

$$\hat{g}_{\gamma,data}(\cdot) = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left\{ \frac{1}{m} \sum_{j=1}^{m} (f(\boldsymbol{x}_j) - y_j)^2 + \gamma \|f\|^2_{\mathcal{H}} \right\}.$$

Moreover, following the notations in Section 4.5 of [136], given Banach space $(\mathscr{L}^2_{P_{tr}}, \|\cdot\|_{\mathscr{L}^2_{P_{tr}}})$ and our kernel-induced Hilbert subspace $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, we define a $\tilde{\mathbb{K}}$-functional: $\mathscr{L}^2_{P_{tr}} \times (0, \infty) \to \mathbb{R}$ to be

$$\tilde{\mathbb{K}}(l, \gamma) \triangleq \inf_{f \in \mathcal{H}} \{ \|l - f\|_{\mathscr{L}^2_{P_{tr}}} + \gamma \|f\|_{\mathcal{H}} \}$$

for $l(\cdot) \in \mathscr{L}^2_{P_{tr}}$ and $t > 0$. For $0 < r < 1$, the interpolation space $(\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$ consists of all the elements $l(\cdot) \in \mathscr{L}^2_{P_{tr}}$ such that

$$\|l\|_r \triangleq \sup_{\gamma > 0} \frac{\tilde{\mathbb{K}}(l, \gamma)}{\gamma^r} < \infty. \tag{4.20}$$

Lemma 11. *Define* $\mathbb{K} : \mathscr{L}^2_{P_{tr}} \times (0, \infty) \to \mathbb{R}$ *to be*

$$\mathbb{K}(l, \gamma) \triangleq \inf_{f \in \mathcal{H}} \{ \|l - f\|^2_{\mathscr{L}^2_{P_{tr}}} + \gamma \|f\|^2_{\mathcal{H}} \}. \tag{4.21}$$

Then for any $l(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$, we have

$$\sup_{\gamma>0} \frac{\mathbb{K}(l,\gamma)}{\gamma^r} \leq \left(\sup_{\gamma>0} \frac{\tilde{\mathbb{K}}(l,\sqrt{\gamma})}{(\sqrt{\gamma})^r} \right)^2 = \|l\|_r^2 < \infty. \tag{4.22}$$

Proof. It follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}, \quad \forall a, b \geq 0$ that

$$\sqrt{\mathbb{K}(l,\gamma)} \leq \tilde{\mathbb{K}}(l,\sqrt{\gamma}). \tag{4.23}$$

Thus, for any $l(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_r$, we have

$$\sup_{\gamma>0} \frac{\mathbb{K}(l,\gamma)}{\gamma^r} \leq \left(\sup_{\gamma>0} \frac{\tilde{\mathbb{K}}(l,\sqrt{\gamma})}{(\sqrt{\gamma})^r} \right)^2 = \|l\|_r^2 < \infty. \tag{4.24}$$

□

On the other hand, assuming $g(\cdot) \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, it follows from the proof of Theorem 4.1 in [136] that

$$g(\cdot) \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H}^+)_{\frac{\theta}{\theta+2}} \tag{4.25}$$

where $\mathcal{H}^+$ is a closed subspace of $\mathcal{H}$ spanned by eigenfunctions of the kernel K (e.g., $\mathcal{H}^+ = \mathcal{H}$ when P_{tr} is non-degenerate, see Remark 4.18 of [136]). Indeed, the next lemma shows we can measure smoothness through interpolation space just as range space.

Lemma 12. *Assuming P_{tr} is non-degenerate on $\mathcal{X}$. Then if $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$, we have $g \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_{\frac{\theta}{\theta+2}}$. On the other hand, if $g \in (\mathscr{L}^2_{P_{tr}}, \mathcal{H})_{\frac{\theta}{\theta+2}}$, then $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}-\epsilon})$ for all $\epsilon > 0$.*

Proof. The proof follows from Theorem 4.1, Corollary 4.17 and Remark 4.18 of [136]. □

Now we are ready to adopt some common assumptions and theoretical results from learning theory in RKHS. They can be found in [136, 142, 141, 135]. First, given $g(\cdot) \in Range(\mathcal{T}_K^{\zeta})$ and m training sample $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^m$ (sampled from P_{tr})), it follows from Lemma 3 of [141] (see as well

Remark 3.3 and Corollary 3.2 in [142]) that

$$\|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2 \gamma^\zeta. \tag{4.26}$$

Second, it follows from Theorem 3.1 in [142] as well as [141, 148] that

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_2'(\gamma^{-1/2} m^{-1/2} + \gamma^{-1} m^{-3/4}), \tag{4.27}$$

and, by the triangle inequality,

$$\|g - \hat{g}_{\gamma,data}\|_{\mathscr{L}^2_{P_{tr}}} \leq C_3(\gamma^\zeta + \gamma^{-1/2} m^{-1/2} + \gamma^{-1} m^{-3/4}). \tag{4.28}$$

Notice here that by choosing $\gamma = m^{-\frac{3}{4(1+\zeta)}}$, we recover Corollary 3.2 of [142]. Finally it follows from Theorem 1 of [141], we have

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3' \gamma^{-1} m^{-1/2}, \tag{4.29}$$

with $C_3' = 6R \log \frac{2}{\delta}$. In fact, if we define $\sigma^2 \triangleq \mathbb{E}_{\boldsymbol{x} \sim P_{tr}} \mathbb{E}_{Y|\boldsymbol{x}} (g(\boldsymbol{x}) - Y)^2$, then Theorem 3 of [141] stated that

$$\|g_\gamma - \hat{g}_{\gamma,data}\|_{\mathcal{H}} \leq C_3''((\sqrt{\sigma^2} + \|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}})\gamma^{-1} m^{-1/2} + \gamma^{-1} m^{-1}). \tag{4.30}$$

4.7.3 Main Proofs

Proof of Theorem 1 and Corollary 1. If $g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ (i.e. $\zeta = \frac{\theta}{2\theta+4}$) and we set $h(\cdot) = g_\gamma(\cdot)$ and $\hat{g} = \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}$ for some $\gamma > 0$, then

$$
\begin{aligned}
&V_R(\rho) - \nu \\
=&\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr})) + \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr}) - h(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) \\
&+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \beta(\boldsymbol{x}_j^{tr})(g(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) - \nu. \qquad (4.31)
\end{aligned}
$$

To bound terms in (4.31), we first use Corollary 4.7.1 to conclude that with probability at least $1 - \delta$,

$$
|\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr})(y_j^{tr} - g(\boldsymbol{x}_j^{tr}))| \leq B \sqrt{\frac{1}{\lfloor \rho n_{tr} \rfloor} \log \frac{2}{\delta}} = O(n_{tr}^{-1/2}). \qquad (4.32)
$$

We hold on our discussion for the second term. For the third term, since $h, \hat{g} \in \mathcal{H}$,

$$
\begin{aligned}
&\left| \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))(h(\boldsymbol{x}_j^{tr}) - \hat{g}(\boldsymbol{x}_j^{tr})) \right| \\
=&\left| \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\left\langle h - \hat{g}, \Phi(\boldsymbol{x}_j^{tr}) \right\rangle_{\mathcal{H}} \right| \\
=&\left| \left\langle h - \hat{g}, \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}_j^{tr}) - \beta(\boldsymbol{x}_j^{tr}))\Phi(\boldsymbol{x}_j^{tr}) \right\rangle_{\mathcal{H}} \right| \\
\leq&\|h - \hat{g}\|_{\mathcal{H}} (\hat{L}(\hat{\boldsymbol{\beta}}) + \hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr}, \ldots, \boldsymbol{x}_{\lfloor \rho n_{tr} \rfloor}^{tr}})) \leq 2\|h - \hat{g}\|_{\mathcal{H}} \hat{L}(\boldsymbol{\beta}_{|\boldsymbol{x}_1^{tr}, \ldots, \boldsymbol{x}_{\lfloor \rho n_{tr} \rfloor}^{tr}}), \qquad (4.33)
\end{aligned}
$$

by definition of (1). Thus, when taking $h = g_\gamma$ and $\hat{g} = \hat{g}_{\gamma, \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}$ for some γ, we can combine (4.18) and (4.29) to guarantee, with probability $1 - 2\delta$,

$$\begin{aligned}
&\left| \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} (\hat{\beta}(\boldsymbol{x}^{tr}_j) - \beta(\boldsymbol{x}^{tr}_j))(h(\boldsymbol{x}^{tr}_j) - \hat{g}(\boldsymbol{x}^{tr}_j)) \right| \\
\leq & \sqrt{8 \log \frac{2}{\delta}} RC(1-\rho)^{-1/2} (\gamma^{-1} n_{tr}^{-1/2}) \cdot \sqrt{\left(\frac{B^2}{n_{tr}} + \frac{1}{n_{te}} \right)} \\
= & O(\gamma^{-1} n_{tr}^{-1/2} (n_{tr}^{-1} + n_{te}^{-1})^{\frac{1}{2}}).
\end{aligned} \tag{4.34}$$

For the last term $\tau \triangleq \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \beta(\boldsymbol{x}^{tr}_j)(g(\boldsymbol{x}^{tr}_j) - \hat{g}(\boldsymbol{x}^{tr}_j)) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}^{te}_i) - \nu$, the analysis relies the splitting of data, as we notice that

$$\begin{aligned}
&\mathbb{E}_{| \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}} \left[\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \beta(\boldsymbol{x}^{tr}_j)(g(\boldsymbol{x}^{tr}_j) - \hat{g}(\boldsymbol{x}^{tr}_j)) + \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(X^{te}_i) - \nu \right] \\
= & \mathbb{E}_{\boldsymbol{x} \sim P_{tr}}[\beta(\boldsymbol{x}) g(\boldsymbol{x})] - \nu - \mathbb{E}_{\boldsymbol{x} \sim P_{tr}}[\beta(\boldsymbol{x}) \hat{g}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[\hat{g}(\boldsymbol{x})] \\
= & \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[g(\boldsymbol{x})] - \nu - \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[\hat{g}(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim P_{te}}[\hat{g}(\boldsymbol{x})] \\
= & 0.
\end{aligned} \tag{4.35}$$

Notice the second line follows since $\hat{g}(\cdot)$ is determined by $\{\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}\}$ and thus is independent of $\{\boldsymbol{X}^{tr}_{KMM}, \boldsymbol{Y}^{tr}_{KMM}\}$ or $\{\boldsymbol{X}^{te}\}$. Thus, we have

$$\begin{aligned}
\mathrm{Var}(\tau) = & \mathrm{Var}(\mathbb{E}_{| \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}(\tau)) + \mathbb{E}[\mathrm{Var}_{| \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}(\tau)] \\
= & \mathbb{E}[\mathrm{Var}_{| \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}(\tau)] \\
= & \frac{1}{\lfloor \rho n_{tr} \rfloor} \mathbb{E}[\mathrm{Var}_{\boldsymbol{x} \sim P_{tr} | \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}(\beta(\boldsymbol{x})(g(\boldsymbol{x}) - \hat{g}(\boldsymbol{x})))] + \frac{1}{n_{te}} \mathbb{E}[\mathrm{Var}_{\boldsymbol{x} \sim P_{te} | \boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}}(\hat{g}(\boldsymbol{x}))] \\
\leq & \frac{B^2}{\lfloor \rho n_{tr} \rfloor} \mathbb{E}_{\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}} \| g - \hat{g} \|^2_{\mathscr{L}^2_{P_{tr}}} + \frac{1}{n_{te}} \mathbb{E}_{\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}} \| \hat{g} \|^2_{\mathscr{L}^2_{P_{te}}} \\
\leq & \frac{B^2}{\lfloor \rho n_{tr} \rfloor} \mathbb{E}_{\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}} \| g - \hat{g} \|^2_{\mathscr{L}^2_{P_{tr}}} + \frac{B}{n_{te}} \mathbb{E}_{\boldsymbol{X}^{tr}_{NR}, \boldsymbol{Y}^{tr}_{NR}} \| \hat{g} \|^2_{\mathscr{L}^2_{P_{tr}}},
\end{aligned} \tag{4.36}$$

and we can use the Chebyshev inequality and Lemma 9 to conclude, with probability at least $1-\delta$,

$$|\tau| \leq \sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr} \rfloor}\mathbb{E}_{\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}\|g-\hat{g}\|_{\mathscr{L}_{P_{tr}}^2}^2 + \frac{BR^2}{n_{te}}}, \tag{4.37}$$

which becomes, by (4.28), with probability $1-2\delta$,

$$\begin{aligned}|\tau| \leq& \sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\lfloor \rho n_{tr} \rfloor}C(1-\rho)^{-3/4}(\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4}) + \frac{BR^2}{n_{te}}} \\ =& O((\gamma^{\zeta}+\gamma^{-1/2}n_{tr}^{-1/2}+\gamma^{-1}n_{tr}^{-3/4})n_{tr}^{-1/2}+n_{te}^{-1/2})\end{aligned} \tag{4.38}$$

with $\zeta = \frac{\theta}{2\theta+4}$. Now, to bound the second term $\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}(\hat{\beta}(\boldsymbol{x}_j^{tr})-\beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr})-h(\boldsymbol{x}_j^{tr}))$, we have

$$\begin{aligned}&\frac{1}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|(\hat{\beta}(\boldsymbol{x}_j^{tr})-\beta(\boldsymbol{x}_j^{tr}))(g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr}))| \\ \leq& \frac{B}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})| \\ \leq& \Big|\frac{B}{\lfloor \rho n_{tr} \rfloor}\sum_{j=1}^{\lfloor \rho n_{tr} \rfloor}|g(\boldsymbol{x}_j^{tr})-g_{\gamma}(\boldsymbol{x}_j^{tr})| - B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^1}\Big| + B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^1} \\ \leq& \sqrt{\frac{1}{\delta}}\sqrt{\frac{B^2}{\rho n_{tr}}\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^2}^2} + B\|g-g_{\gamma}\|_{\mathscr{L}_{P_{tr}}^2} \\ \leq& \sqrt{\frac{1}{\delta}}BC\gamma^{\zeta}\sqrt{\frac{1}{\rho n_{tr}}} + C\gamma^{\zeta} = O(\gamma^{\zeta}) = O(\gamma^{\frac{\theta}{2\theta+4}}).\end{aligned} \tag{4.39}$$

where $\mathscr{L}_{P_{tr}}^1$ denotes the 1-norm $\mathbb{E}_{\boldsymbol{x}\sim P_{tr}}|g(\boldsymbol{x})-g_{\gamma}(\boldsymbol{x})|$. Notice the second-to-last line follows from the Chebyshev inequality, the Cauchy-Schwarz inequality, and the last line from (4.26).

Thus, when taking $h = g_{\gamma}$ and $\hat{g} = \hat{g}_{\gamma,\boldsymbol{X}_{NR}^{tr},\boldsymbol{Y}_{NR}^{tr}}$ for some $\gamma > 0$, we can combine (4.32), (4.34),

(4.38) and (4.39) to have

$$\begin{aligned}|V_R(\rho) - \nu| =& O(n_{tr}^{-\frac{1}{2}}) + O(\gamma^{\frac{\theta}{2\theta+4}}) + O(\gamma^{-1} n_{tr}^{-1/2} (n_{tr}^{-1} + n_{te}^{-1})^{\frac{1}{2}}) \\ &+ O((\gamma^{\frac{\theta}{2\theta+4}} + \gamma^{-1/2} n_{tr}^{-1/2} + \gamma^{-1} n_{tr}^{-3/4}) n_{tr}^{-1/2} + n_{te}^{-1/2}) \\ =& O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}} + \gamma^{\frac{\theta}{2\theta+4}} + \gamma^{-\frac{1}{2}} n_{tr}^{-1} + \gamma^{-\frac{1}{2}} n_{tr}^{-\frac{1}{2}} n_{te}^{-\frac{1}{2}}), \end{aligned} \tag{4.40}$$

after simplification. Now, if we take $\gamma = n^{-\frac{\theta+2}{\theta+1}}$ where $n \triangleq \min(n_{tr}, n_{te})$, then (4.40) becomes

$$\begin{aligned}&|V_R(\rho) - \nu| \\ =& O(n^{-\frac{1}{2}} + n^{-\frac{\theta}{2(\theta+1)}} + n^{\frac{\theta+2}{2(\theta+1)}} n^{-1}) = O(n^{-\frac{\theta}{2\theta+2}}) = O(n_{tr}^{-\frac{\theta}{(2\theta+2)}} + n_{te}^{-\frac{\theta}{(2\theta+2)}}), \end{aligned} \tag{4.41}$$

which is the statement of the theorem. However, note that if we choose $\gamma = n^{-1}$, we would achieve the convergence rate of V_{KMM} as $O(n_{tr}^{-\frac{\theta}{(2\theta+4)}} + n_{te}^{-\frac{\theta}{(2\theta+4)}})$. Moreover if $\lim_{n\to\infty} n_{te}^{\frac{6\theta+8}{3\theta+6}} / n_{tr} \to 0$ and we choose $\gamma = n_{tr}^{-1}$, then the rate becomes $O(n_{tr}^{-\frac{\theta}{2\theta+4}} + n_{te}^{-\frac{1}{2}})$. □

Proof of Proposition 1. Fixing $\gamma > 0$, if $g \in \mathcal{H}$ ($i.e., g \in Range(\mathcal{T}_K^{\frac{\theta}{2\theta+4}})$ with $\theta \to \infty$), then by definition of g_γ we would have

$$\|g_\gamma\|_{\mathcal{H}}^2 \leq \frac{\|g_\gamma - g\|_{\mathscr{L}^2_{P_{tr}}}^2 + \gamma \|g_\gamma\|_{\mathcal{H}}^2}{\gamma} \leq \frac{\|g - g\|_{\mathscr{L}^2_{P_{tr}}}^2 + \gamma \|g\|_{\mathcal{H}}^2}{\gamma} = \|g\|_{\mathcal{H}}^2, \tag{4.42}$$

or equivalently $\|g_\gamma\|_{\mathcal{H}} = O(1)$ since the fixed true regression function $\|g\|_{\mathcal{H}} = O(1)$. Thus, a simplified analysis shows

$$\begin{aligned}V_R(\rho) - \nu =& \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) Y_j^{tr} - \nu \\ &+ \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) \end{aligned} \tag{4.43}$$

Note that the first term on the right is nothing but the V_{KMM} estimator with $100 \times \rho$ percent of the training data and we shall denote it as $V_{KMM}(\rho)$ without ambiguity. For the second term, assuming

$\hat{g} = \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}$, is bounded by

$$\begin{aligned}
&\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \hat{g}(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \hat{g}(\boldsymbol{x}_i^{te}) \\
=&\frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{j=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \left\langle \hat{g}, \Phi(\boldsymbol{x}_j^{tr}) \right\rangle_{\mathcal{H}} - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \left\langle \hat{g}, \Phi(\boldsymbol{x}_i^{n_{te}}) \right\rangle_{\mathcal{H}} \\
=&\left\langle \hat{g}, \frac{1}{\lfloor \rho n_{tr} \rfloor} \sum_{i=1}^{\lfloor \rho n_{tr} \rfloor} \hat{\beta}(\boldsymbol{x}_j^{tr}) \Phi(\boldsymbol{x}_j^{tr}) - \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \Phi(\boldsymbol{x}_i^{te}) \right\rangle_{\mathcal{H}} \leq \|\hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} \hat{L}(\hat{\boldsymbol{\beta}}),
\end{aligned} \tag{4.44}$$

Then, by (4.43) and (4.44), we have

$$\begin{aligned}
|V_R(\rho) - \nu| \leq& |V_{KMM}(\rho) - \nu| + \hat{L}(\hat{\boldsymbol{\beta}}) (\|g_\gamma - \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\
=& O(n_{tr}^{-\frac{1}{2}} + n_{te}^{-\frac{1}{2}}),
\end{aligned} \tag{4.45}$$

following (4.42), (4.29) and Theorem 1 of [135]. □

Proof of Proposition 2. If the function g only satisfies the condition $\mathcal{A}_\infty(g, F) \triangleq \inf_{\|f\|_{\mathcal{H}} \leq F} \|g - f\| \leq C(\log F)^{-s}$ for some $C, s > 0$, then we again follow the analysis in the proof of Proposition 1 and arrive at the decomposition in (4.43)

$$\begin{aligned}
|V_R(\rho) - \nu| \leq& |V_{KMM}(\rho) - \nu| + \hat{L}(\hat{\boldsymbol{\beta}}) (\|g_\gamma - \hat{g}_{\gamma, \boldsymbol{X}_{NR}^{tr}, \boldsymbol{Y}_{NR}^{tr}}\|_{\mathcal{H}} + \|g_\gamma\|_{\mathcal{H}}) \\
=& O(\log \frac{n_{tr} n_{te}}{n_{tr} + n_{te}})^{-s},
\end{aligned} \tag{4.46}$$

which is the rate of V_{KMM} by Theorem 3 of [135]. □

Proof of Theorem 2. Define $\epsilon \triangleq \sup_{\theta \in \mathcal{D}} \left| V_R(\theta) - \mathbb{E}[l'(X^{te}, Y^{te}; \theta)] \right|$. We have

$$\mathbb{E}[l'(X_{te}, Y_{te}; \hat{\theta}_R)] - \epsilon \leq V_R(\hat{\theta}_R) \leq V_R(\theta^\star) \leq \mathbb{E}[l'(X_{te}, Y_{te}; \theta^\star)] + \epsilon. \tag{4.47}$$

On the other hand, we know by the triangle inequality that ϵ is bounded by

$$\sup_{\theta\in\mathcal{D}}\Big|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\hat{\beta}(\boldsymbol{x}_j^{tr})l'(\boldsymbol{x}_j^{tr},y_j^{tr};\theta)-\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}l(\boldsymbol{x}_i^{te};\theta)\Big|$$
$$+\sup_{\theta\in\mathcal{D}}\Big|\frac{1}{\lfloor\rho n_{tr}\rfloor}\sum_{j=1}^{\lfloor\rho n_{tr}\rfloor}\hat{\beta}(\boldsymbol{x}_j^{tr})\hat{l}(\boldsymbol{x}_j^{tr};\theta)-\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}\hat{l}(\boldsymbol{x}_i^{te};\theta)\Big|+\sup_{\theta\in\mathcal{D}}\Big|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}l(\boldsymbol{x}_i^{te};\theta)-\mathbb{E}[l(X_{te};\theta)]\Big|,$$

where the first term is bounded by $O(n_{tr}^{-\frac{1}{2}}+n_{te}^{-\frac{1}{2}})$ following Corollary 8.9 in [115]. Moreover, the second term is also $O(n_{tr}^{-\frac{1}{2}}+n_{te}^{-\frac{1}{2}})$ as in (4.44) or Lemma 8.7 in [115]. For the last term, due to the Lipschitz and compact assumption, it follows from Theorem 19.5 of [69] (see also Example 19.7 of [69]) that function class $\mathcal{G}$ is P_{te}-Donsker, which means that

$$\mathbb{G}_n(\theta)\triangleq\sqrt{n_{te}}\left(\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}l(\boldsymbol{x}_i^{te};\theta)-\mathbb{E}_{\boldsymbol{x}\sim P_{te}}[l(\boldsymbol{x};\theta)]\right)$$

converges in distribution to a Gaussian Process $\mathbb{G}_\infty$ with zero mean and covariance function $\mathrm{Cov}(\mathbb{G}_\infty(\theta_1),\mathbb{G}_\infty(\theta$ $\mathbb{E}_{\boldsymbol{x}\sim P_{te}}(l(\boldsymbol{x};\theta_1)l(\boldsymbol{x};\theta_2))-\mathbb{E}_{\boldsymbol{x}\sim P_{te}}l(\boldsymbol{x};\theta_1)\mathbb{E}_{\boldsymbol{x}\sim P_{te}}l(\boldsymbol{x};\theta_2)$. Notice $\mathbb{G}_\infty$ can be viewed as random function in $C(\mathcal{D})$, the space of continuous and bounded function on θ. Since for any $z\in C(\mathcal{D})$, the mapping $z\to\|z\|_\infty\triangleq\sup_{\theta\in\mathcal{D}}z(\theta)$ is continuous with respect to the supremum norm, it follows from the continuous-mapping theorem that $n_{te}^{\frac{1}{2}}\sup_{\theta\in\mathcal{D}}\left|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}l(\boldsymbol{x}_i^{te};\theta)-\mathbb{E}[l(X_{te};\theta)]\right|$ converges in distribution to $\|\mathbb{G}_\infty\|_\infty$ which has finite expectations based on the assumptions on $\mathcal{G}$ (see, e.g., Section 14, Theorem 1 of [149]). Thus, by definition of convergence in distribution, for any $\delta>0$, we can find some constant D' that

$$P(\|\mathbb{G}_n\|_\infty>D')=P(\|\mathbb{G}_\infty\|_\infty>D')+o(1)\leq\delta+o(1),\tag{4.48}$$

which means, we can find some N such that when $n_{te}>N$,

$$P_{te}\Big(\sup_{\theta\in\mathcal{D}}\left|\frac{1}{n_{te}}\sum_{i=1}^{n_{te}}l(\boldsymbol{x}_i^{te};\theta)-\mathbb{E}[l(X_{te};\theta)]\right|>n_{te}^{-\frac{1}{2}}D'\Big)=P_{te}(\|\mathbb{G}_n\|_\infty>D')\leq 2\delta,$$

and consequently, with probability $1 - 2\delta$, we have

$$\sup_{\theta \in \mathcal{D}} \Big| \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te}; \theta) - \mathbb{E}[l(X_{te}; \theta)] \Big| \leq n_{te}^{-\frac{1}{2}} D'.$$

In other words, we also have

$$\sup_{\theta \in \mathcal{D}} \Big| \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} l(\boldsymbol{x}_i^{te}; \theta) - \mathbb{E}[l(X_{te}; \theta)] \Big| = O(n_{te}^{-\frac{1}{2}}),$$

which concludes our proof. □

Chapter 5: Constrained Reinforcement Learning via Policy Splitting

We develop a model-free reinforcement learning approach to solve constrained Markov decision processes, where the objective and budget constraints are in the form of infinite-horizon discounted expectations, and the rewards and costs are learned sequentially from data. We propose a two-stage procedure where we first search over deterministic policies, followed by an aggregation with a mixture parameter search, that generates policies with simultaneous guarantees on near-optimality and feasibility. We also numerically illustrate our approach by applying it to an online advertising problem.

We note the special structure we developed here is currently limited to CMDP with one constraint. A further generalization with multiple constraints might be worth exploring.

5.1 Introduction

Applications of Reinforcement Learning (RL) in online advertising with recommendation systems have been a topic of major research interests ([10, 11, 12]). However, despite their tremendous success, most RL-methods are not designed to learn optimal policies under constraints, yet they appear ubiquitously when facing budget or safety considerations. A standard framework for studying RL under constraints is the Constrained Markov Decision Process (CMDP), where the objective is to maximize the long-run return, with constraints on one or several types of long-run costs. In this Chapter, we consider the case where both the objective and the constraint are in the form of an infinite-horizon cumulative discounted expectation, whereas the returns, costs and transitions are revealed from sequential data. The goal is to design an efficient methodology for the constrained problem by assimilating classical optimality properties of CMDP into RL, in order to efficiently use established RL approaches and obtain policies that enjoy both near-optimality and feasibility.

The CMDP in the form described above is motivated from a range of important applications including online advertising. Sponsored search campaigns, for instance, are designed based on predetermined budgets. Therefore, the marketer has to employ effective strategies to accrue the maximum reward while observing certain monetary constraints throughout the campaign. Similarly, in email campaigns, the marketer can only send out a limited number of emails under different constraints due to user fatigue or limited available discount offers. Thus, it is important to consider information beyond potential revenues, such as the remaining budget or the likely outcomes of different offers. Direct applications of most RL-algorithms do not, in general, consistently produce optimal solutions within these budget constraint. Thus, several lines of work have been devoted to resolve this challenge. In the model-based regime (i.e., parametric-based transition), [150] and [151] consider linear programming, [150] considers state-space extension, and [152] considers policy iterations. However, model-based algorithms suffer when the state or action space gets large as estimating the transition dynamics of the users can be very challenging or even infeasible. In model-free settings, constrained policy optimization (CPO) ([153]) is designed based on trust region policy optimization (TRPO) and its variants ([154, 155]). Through surrogate function approximations, CPO provides safe iterations in each policy update, preventing any constraint violation in the agent's learning process. However, the implementation requires a safe policy to start with and it may be over-conservative to require a safe update in each iteration, especially for areas of advertising where the budget constraint is not as hard a constraint as, say, in auto-driving. Thus, the extra effort and setup in the implementation of CPO might not be as desirable in our setting. Another line of work in tacking constrained MDP uses primal-dual, Lagrangian-based RL methods ([156, 157]), which involves stochastic updates for solving the KKT conditions. In particular, [156] investigates constraints arising from risk criteria such as conditional-value-at-risk or chance constraints while the reward constrained policy optimization (RCPO) in [157] uses an actor-critic updates in the policy space and a stochastic recursion on the Lagrange multiplier updates in the dual space. However, although convergence is guaranteed for primal-dual methods in theory, in practice significant efforts are required to tune the hyper-parameters, especially the learning rates

of the dual variable, as the updates become noisy and unstable around convergence and the training process can easily become too slow or overly greedy.

In this Chapter, we address these issues on the primal-dual formulation and explain the unstable convergence behavior of primal-dual methods around the optimal value. Furthermore, we design a mixing method which aims to alleviate the tuning issues by both exploiting the low-dimensional feature of dual variables (when the number of budget constraints is negligible compared to the cardinality of the state/action space) and investigating a special splitting property of CMDPs ([152]). In particular, for a single budget constraint, the "splitting" property refers to a structure of the optimal randomized policy in CMDP where two possible actions are assigned with a binary distribution to a certain state and the policy stays deterministic elsewhere ([152]). This splitting property contributes to the unstable behaviors of the dual convergence because the RL method is essentially searching for two different optimal policies around the optimal dual value. This splitting property arises from the extreme points of a linear program (LP) formulation of CMDP via the occupation measure ([158]). It reveals the saddle point structure of the Lagrangian and allows us to confine our policy search in a smaller solution space.

Leveraging the splitting property, our approach bypasses the need to search over large spaces of randomized policies and, by solving a sequence of RL problems without restriction under the Lagrangian relaxation, finds candidate deterministic policies with direct application of classical RL-methods (e.g. Q-learning, TD-learning or TRPO). To improve on the undesirable properties of primal-dual methods around convergence, we first propose a discretization scheme which exploits the one-dimensional structure of dual variable and allows for parallel computing. Then we propose a novel feasibility mixing procedure which efficiently mixes the candidate policies and find an optimal randomized policy that would achieve both optimality and feasibility. We provide theoretical justifications on our framework, and also conduct experiments on an online advertisement problem to demonstrate its performance.

The remainder of this Chapter is organized as follows. Section 5.2 presents our problem setting and notations. Section 5.3 describes our Lagrangian formulation and its implications. Section 5.4

presents our main dual Q-learning algorithm that harnesses the splitting property of CMDP in the Lagrangian formulation. Section 5.5 discusses practical implementation, and Section 5.6 illustrates our experimental results.

5.2 Problem Setting

A Constrained Markov Decision Process (CMDP) can be formulated as follows. Let $\mathcal{S}$ be the finite set of states, $\mathcal{A}$ the finite set of actions, and $p(s, a, s')$ the probability measure governing the stochastic transition between states, namely

$$\mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a) = p(s, a, s')$$

with non-negative entries and $\sum_{s'} p(s, a, s') = 1$. Let $r_t = r(s_t, a_t)$ be the corresponding expected reward. Denote Π to be the space of stationary randomized policies π where

$$\mathbb{P}(a_t = a | s_0, a_0, r_1, s_1, a_1, ..., r_t, s_t = s) = \mathbb{P}(a_t = a | s_t = s) = \pi(s, a),$$

and $\sum_a \pi(s, a) = 1, \pi(s, a) \geq 0$ for all a, s. Notice the stationarity comes from the fact that the policy at each state s does not change with t. Moreover, if over any state s, $\pi(s, a)$ is zero for all but one action $a \in \mathcal{A}$, then we say $\pi \in \Pi_0 \subset \Pi$ is a stationary deterministic policy and denote this a by $\pi(s)$. Suppose at each step t, the agent interacting with the environment not only receives random (immediate) reward r_t but also incurs random (immediate) cost denoted by $c_t = c(s_t, a_t)$. Let $s_0 \sim \rho$ be the distribution of the initial state and $\gamma \in [0, 1]$ be the discounted factor. We consider the following CMDP:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & \mathbb{E}_{s_0 \sim \rho, \pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \right] \\ \text{s.t.} \quad & \mathbb{E}_{s_0 \sim \rho, \pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} c_t \right] \leq B, \end{aligned} \tag{5.1}$$

where $\mathbb{E}_{s_0\sim\rho,\pi}$ denotes the expectation under policy π and initial distribution $s_0 \sim \rho$. We confine our policy search in Π because it is well-known (see, e.g., [158]) that the optimal policy $\pi^\star$ for CMDP lies in the space Π. Also, we do not assume the distributions of $r(\cdot,\cdot)$ $c(\cdot,\cdot)$, or $p(\cdot,\cdot,\cdot)$ are known.

5.3 Lagrangian with Reduced Policy Space

A common way to solve CMDP (5.1) is to formulate it as the following LP ([158]):

$$\begin{aligned} \max_{\mathbf{x}\geq 0} \quad & \sum_{s,a} x_{sa} r(s,a) \\ \text{s.t.} \quad & \sum_{s,a} x_{sa} c(s,a) \leq B, \\ & \sum_{a} x_{sa} - \gamma \sum_{s',a} x_{s'a} p(s',a,s) = \rho(s)\ \forall s, \end{aligned} \tag{5.2}$$

where $x_{sa} = \sum_{t=1}^{\infty} \gamma^{t-1}\mathbb{P}(s_t = s, a_t = a|\pi, s_0 \sim \rho)$ is referred to as the *occupation measure* of policy π under initial distribution ρ. It can be interpreted as the total discounted expected number of times state-action pair (s,a) is visited under policy π, so that $\mathbb{E}_\pi[\sum_{t=1}^{\infty}\gamma^{t-1}r_t]$ can be seen to be expressible as $\sum_{s,a} x_{sa} r(s,a)$ and similarly $\mathbb{E}_\pi[\sum_{t=1}^{\infty}\gamma^{t-1}c_t]$ as $\sum_{s,a} x_{sa} c(s,a)$, and the second constraint in (5.2) follows from a first-step Markovian analysis. Moreover, it is shown in [158] that an optimal randomized policy $\pi^\star$ can be computed from an optimal solution $\mathbf{x}^\star$ of (5.2) by letting

$$\pi^\star(s,a) = \frac{x^\star_{sa}}{\sum_a x^\star_{sa}}. \tag{5.3}$$

However, formulating the above optimization problem requires the knowledge of $r(s,a), c(s,a)$ and $p(s,a,s')$ of the MDP which in our setting can only be learned implicitly. Also, the number of state-action pair may get too large to use tabular methods. On the other hand, the more efficient, large-scale approximate RL methods such as TD-learning, Q-learning or TRPO ([159, 160]) cannot directly help us with the search of an optimal randomized policy. To address this issue, we first

consider the dual optimization problem ([161]) of (5.2):

$$\begin{aligned} \min_{\lambda \geq 0, \mathbf{v}} \quad & \sum_{s} v_s \rho(s) + \lambda B \\ \text{s.t.} \quad & v_s \geq r(s,a) - \lambda c(s,a) + \gamma \sum_{s'} p(s,a,s') v_{s'} \ \forall s. \end{aligned} \tag{5.4}$$

For fixed $\lambda \geq 0$, the minimization in (5.4) is exactly the LP formulation for solving the value function of an unconstrained MDP with adjusted reward $r_t^\lambda = r_t - \lambda c_t$ instead of r_t at each step t (plus the constant term λB), and the constraint follows from the Bellman optimality equation ([162]). This allows us to convert (5.1) into the form (5.5) (shown below). Advantageously, for any fixed λ, because of its unconstrained nature, the inner maximization problem in (5.5) now suffices to search for policy π in the deterministic policy space Π_0 instead of the randomized policy space Π. Hence we can apply many suitable approximation algorithms in RL to search for the optimal deterministic policy ([159]). We have the following theorem (Notice the reduction of policy space into Π_0 as a key transition in this dual):

Theorem 5.3.1. *Problem* (5.1) *can be reformulated as*

$$\min_{\lambda \geq 0} \quad \max_{\pi \in \Pi_0} \quad \mathcal{R}(\pi, \rho) - \lambda \big(C(\pi, \rho) - B \big) \tag{5.5}$$

where $\mathcal{R}(\pi, \rho) \triangleq \mathbb{E}_{s_0 \sim \rho, \pi}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$ *and* $C(\pi, \rho) \triangleq \mathbb{E}_{s_0 \sim \rho, \pi}[\sum_{t=1}^{\infty} \gamma^{t-1} c_t]$.

Proof. Based on our discussion and the LP duality, we only have to show that for any fixed $\lambda \geq 0$,

$$\begin{aligned} \min_{\mathbf{v}} \quad & \sum_{s} v_s \rho(s) \\ \text{subject to} \quad & v_s \geq r(s,a) - \lambda c(s,a) + \gamma \sum_{s'} p(s,a,s') v_{s'} \ \forall s \end{aligned} \tag{5.6}$$

is equivalent to

$$\max_{\pi \in \Pi_0} \quad \mathcal{R}(\pi, \rho) - \lambda C(\pi, \rho). \tag{5.7}$$

In particular, for fixed $\lambda \geq 0$, problem (5.6) obtains the optimal expected total discounted reward

$\sum_s v_s \rho(s)$ with adjusted reward $r_t^\lambda = r_t - \lambda c_t$ guaranteed by the Bellman optimality constraint as well as the condition that $\rho(s) > 0, \forall s$ ([162]). On the other hand, given the discounted adjusted reward r_t^λ, we know from classical MDP results that for any unconstrained infinite-horizon discounted MDP there exists a stationary and deterministic optimal policy $\pi^\star \in \Pi_0$ for any initial state distribution satisfying $\rho(s) > 0, \forall s$. Moreover, the optimal expected total discounted reward is $\max_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho) - \lambda C(\pi, \rho)$. □

Theorem 1 suggests that the search for optimal policies can first proceed with a deterministic policy search fixing some set of λ. Then, we optimize with respect to λ in (5.5) to find an optimal $\lambda^\star$ which closes the duality gap between (5.2) and (5.4) with optimal policies that maximize the penalized expected reward $r_t - \lambda^* c_t$ plus the term $\lambda^\star B$.

5.4 Policy Mixing and Dual Q-Learning

The two steps discussed above recover the optimal value of the primal (5.2). However, to recover the optimal, possibly randomized policy, we need to look more closely at the dual problem (5.5). To begin, it is known that if an LP has an optimal solution, then it also has an optimal basic feasible solution ([161]), meaning that we can find optimal solution $\mathbf{x}^\star$ with at most $s+1$ non-zeros entries. This leads to the following proposition.

Proposition 6. *If $\rho(s) > 0 \; \forall s$, then there is an optimal policy $\pi^\star$ for the primal problem* (5.1) *with $\pi^\star(s)$ following a deterministic action for all but possibly one state.*

Proof. Given that we can find optimal solution $\mathbf{x}^\star$ for problem (5.2) with at most $s + 1$ non-zero entries, if we further assume that $\rho(s) > 0$ for all state s, then the second constraint of (5.2) would force any feasible solution $\mathbf{x}$ to satisfy $\sum_a x_{sa} > 0$ for any s. This condition implies that for any s, we can find at least one a such that $x_{sa}^\star > 0$. Since $\mathbf{x}^\star$ has at most $s + 1$ non-zeros entries, we can have at most one positive entry among all entries of $x_{sa}^\star$. It then follows from (5.3) that the optimal policy $\pi^\star$ for (5.1) is deterministic at all states except possibly one, where the optimal policy splits into two possible actions. □

Following Proposition 6, we can characterize an important property regarding the optimal policy for (5.5). In particular, we consider the dual function

$$\mathcal{D}(\lambda) \triangleq \max_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho) - \lambda\big(C(\pi, \rho) - B\big). \tag{5.8}$$

Theorem 5.4.1. *Assume $\rho(s) > 0 \; \forall s$ and the optimal policy $\pi^\star$ for problem* (5.1) *is unique. Then the maximization in* (5.8)*, at the optimal λ^* that solves* (5.5)*, admits either a deterministic optimal policy $\pi^\star$, or a pair of optimal deterministic policies π_1, π_2 with actions different in one state s and $\pi^\star = (1-t)\pi_1 + t\pi_2$ for some $0 < t < 1$.*

Proof. Let $\pi^\star$ be the optimal, possibly randomized policy for the primal (5.1). By the LP duality ([161]), we know the optimal values for (5.1) and (5.5) are equal and we must have, for some $\lambda^\star \in \operatorname*{argmin}_{\lambda \geq 0} \mathcal{D}(\lambda) \geq 0$, that

$$\mathcal{R}(\pi^\star, \rho) = \min_{\lambda \geq 0} \mathcal{D}(\lambda) = \mathcal{D}(\lambda^\star). \tag{5.9}$$

If there exists $\lambda^\star = 0$ where (5.9) holds, then

$$\min_{\lambda \geq 0} \mathcal{D}(\lambda) = \mathcal{D}(0) = \max_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho). \tag{5.10}$$

Combining (5.9) and (5.10), we have $\mathcal{R}(\pi^\star, \rho) = \max_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho)$ and by the uniqueness we have $\pi^\star = \operatorname*{argmax}_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho)$. The primal feasibility of (5.1) guarantees $C(\pi^\star, \rho) \leq B$. In fact, notice in this case, the optimal policy for the unconstrained MDP in (5.1) is actually feasible, and thus CMDP (5.1) reduces to an unconstrained MDP.

On the other hand, if we have $\operatorname*{argmin}_{\lambda \geq 0} \mathcal{D}(\lambda) > 0$, then we observe that $\mathcal{D}(\lambda) = \max_{\pi \in \Pi_0} \mathcal{R}(\pi, \rho) - \lambda\big(C(\pi, \rho) - B\big)$ is the maximum of a finite number (i.e. the number of deterministic policies is finite) of linear functions in λ. Thus, $\mathcal{D}(\lambda)$ is piece-wise linear and convex in λ. Since $\lambda^\star > 0$ is the global minimum of $\mathcal{D}(\lambda)$ and $\mathcal{D}(\lambda)$ is piece-wise linear, we must have $\mathcal{D}^+(\lambda^\star) = \lim_{t \to 0} \frac{\mathcal{D}(\lambda^\star + t) - \mathcal{D}(\lambda^\star)}{t} \geq 0$ and $\mathcal{D}^-(\lambda^\star) = \lim_{t \to 0} \frac{\mathcal{D}(\lambda^\star) - \mathcal{D}(\lambda^\star - t)}{t} \leq 0$.

Now if $\lambda^\star = \underset{\lambda \geq 0}{\text{argmin}} \mathcal{D}(\lambda) > 0$ is not unique, then by convexity we can find an interval of λ with the same optimal $\mathcal{D}(\lambda)$, implying the optimal deterministic policy under this λ is both feasible (zero slope means $C(\pi, \rho) = B$) and optimal. Thus, suppose $\lambda^\star = \underset{\lambda \geq 0}{\text{argmin}} \mathcal{D}(\lambda) > 0$ is unique, then we have $\mathcal{D}^-(\lambda^\star) < 0 < \mathcal{D}^+(\lambda^\star)$, and there exists some $\epsilon > 0$ and policies π_1, π_2 such that

$$\mathcal{D}(\lambda) = \mathcal{D}(\lambda^\star) + \mathcal{D}^+(\lambda^\star)(\lambda - \lambda^\star) = \mathcal{R}(\pi_1, \rho) - \lambda\big(C(\pi_1, \rho) - B\big) \tag{5.11}$$

for $\lambda^\star \leq \lambda \leq \lambda^\star + \epsilon$ and

$$\mathcal{D}(\lambda) = \mathcal{D}(\lambda^\star) + \mathcal{D}^-(\lambda^\star)(\lambda - \lambda^\star) = \mathcal{R}(\pi_2, \rho) - \lambda\big(C(\pi_2, \rho) - B\big) \tag{5.12}$$

for $\lambda^\star - \epsilon \leq \lambda \leq \lambda^\star$. In particular, at $\lambda^\star$, we have

$$\mathcal{R}(\pi_1, \rho) - \lambda^\star\big(C(\pi_1, \rho) - B\big) = \mathcal{R}(\pi_2, \rho) - \lambda^\star\big(C(\pi_2, \rho) - B\big) \tag{5.13}$$

which implies

$$\pi_1 = \pi_2 = \underset{\pi \in \Pi_0}{\text{argmax}} \mathcal{R}(\pi, \rho) - \lambda^\star C(\pi, \rho). \tag{5.14}$$

We know from [163] that for a finite unconstrained MDP problem, there exists a unique optimal value function such that $v^\star(s) \geq v^\pi(s)$ for all state s. Thus, (5.14) and the fact that $\rho(s) > 0 \; \forall s$ implies that we must have

$$v^\star(s) = v^{\pi_1}(s) = v^{\pi_2}(s) \quad \forall s \tag{5.15}$$

where $\mathbf{v}^\star$ is the optimal value function for the MDP with adjusted reward $r_t^{\lambda^\star} = r_t - \lambda^\star c_t$ and $\mathbf{v}^{\pi_i}$ is the value of policy π_i under this adjusted reward. This implies $\mathbf{v}^\star, \mathbf{v}^{\pi_1}$ and $\mathbf{v}^{\pi_2}$ must satisfy all

three forms of the Bellman equations:

$$\begin{aligned} v(s) =& \max_a r^{\lambda^\star}(s,a) + \gamma \sum_{s'} p(s,a,s')v(s'), \\ =& r^{\lambda^\star}(s,\pi_1(s)) + \gamma \sum_{s'} p(s,\pi_1(s),s')v(s') = r^{\lambda^\star}(s,\pi_2(s)) + \gamma \sum_{s'} p(s,\pi_2(s),s')v(s'), \end{aligned} \tag{5.16}$$

for all s. Now, for any $0 \le t \le 1$, let π_t be the randomized policy $\pi_t = (1-t)\pi_1 + t\pi_2$. Then the value of policy π_t uniquely satisfies the following Bellman equation:

$$\begin{aligned} v^{\pi_t}(s) &= (1-t)r^{\lambda^\star}(s,\pi_1(s)) + t \cdot r^{\lambda^\star}(s,\pi_2(s)) \\ &+\gamma \sum_{s'} \Big((1-t)p(s,\pi_1(s),s') + tp(s,\pi_2(s),s') \Big) v^{\pi_t}(s') \end{aligned} \tag{5.17}$$

It follows from (5.16) that $\mathbf{v}^\star$ satisfies (5.17) and is thus the value function (i.e. fixed point) of policy π_t. Thus any policy $\pi_t, 0 \le t \le 1$ is optimal for the MDP with adjusted reward $r_t^{\lambda^\star} = r_t - \lambda^\star c_t$ and achieves primal optimality in the sense that

$$\mathcal{R}(\pi^\star, \rho) = \mathcal{D}(\lambda^\star) = \mathcal{R}(\pi_t, \rho) - \lambda^\star \big(C(\pi_t, \rho) - B \big). \tag{5.18}$$

Now, it follows from (5.11) and (5.12) that $\mathcal{D}^+(\lambda^\star) = B - C(\pi_1, \rho) > 0$ and $\mathcal{D}^-(\lambda^\star) = B - C(\pi_2, \rho) < 0$. Furthermore, $C(\pi_t, \rho)$ can be shown to be a continous function of t. Thus, we must have $C(\pi_t, \rho) = B$ for some $0 < t < 1$. Then such π_t satisfies not only primal feasibility but also primal optimality due to (5.18):

$$\mathcal{R}(\pi^\star, \rho) = \mathcal{R}(\pi_t, \rho) - \lambda^\star \big(C(\pi_t, \rho) - B \big) = \mathcal{R}(\pi_t, \rho). \tag{5.19}$$

The claim that π_1 and π_2 differ by one state now follows from (6) and the uniqueness assumption. The other cases where one or both of $\mathcal{D}^+(\lambda^\star)$ and $\mathcal{D}^-(\lambda^\star)$ are 0 lead to either $t = 0$ or 1, which further lead to deterministic policy. The analysis is similar so we omit it. □

Theorem 5.4.1 postulates that the maximization of the Lagrangian or penalized objective $\mathcal{R}(\pi, \rho) - \lambda^* \big(C(\pi, \rho) - B\big)$ generally leads to multiple (deterministic) optimal solutions, even if the primal problem (5.1) has a unique optimal policy. Note that the maximization of $\mathcal{R}(\pi, \rho) - \lambda^* \big(C(\pi, \rho) - B\big)$ is an unconstrained MDP, which allows us to use any classical RL methods to learn its optimal policy. The key is that in order to retrieve the primal optimal policy, we need to identify *two* optimal policies for this penalized objective, and mix them together with a search for the optimal mixture parameter t.

Before presenting practical algorithms for implementation, we first propose a straightforward theoretical procedure in Algorithm 3 that would demonstrate the asymptotic optimality of our method. For demonstration, we would simply use Q-learning on the penalized problem along with subsequent TD-learning for dual updates. However, we note that Algorithm 3 can be replaced by any type of Actor-Critic updates as in [157]. Notation-wise, we use π^λ to denote the optimal deterministic policy for penalized reward $r_t^\lambda = r_t - \lambda c_t$. Given the simple dual Q-learning method described in Algorithm 3, we have the following Theorem 5.4.2. Notice the N chosen large is fixed and does not grow with iterations.

Algorithm 3 Dual Q-learning on Candidates for Mixture

Input: Dual range $0 \leq \lambda_{min} < \lambda_{max}$, discretization parameter n, maximum episode E_1 and E_2, maximum trajectory M_1 and M_2, learning rate α_e, ϵ_{greedy} for the greedy policy and discretized $\lambda_{min} = \lambda_1 < ... < \lambda_n = \lambda_{max}$.

for $i = 1$ **to** n **do**

Initialize : $e \leftarrow 0$, $\hat{Q}^i_e$, the Q-function array for storage (e.g. to 0), an estimate of $Q^i(s,a) = \mathbb{E}_{\pi^{\lambda_i}}[\sum_{t=0}^{\infty} \gamma^t (r_t - \lambda_i c_t) | s_0 = s, a_0 = a]$ and $\{\hat{v}_{cost}\}^i_e$ cost value function array for storage, an estimate of $\mathbb{E}_{\pi^{\lambda_i}}[\sum_{t=0}^{\infty} \gamma^t c_t | s_0 = s]$.

repeat

$e \leftarrow e + 1$, initialize $t \leftarrow 0$ and sample $s_0 \sim \rho$

while s_t is not terminal **and** $t \leq M_1$ **do**

Take action a_t at s_t derived from $\hat{Q}^i_{e-1}$ using ϵ_{greedy}-greedy policy and observe r_{t+1}, s_{t+1}, then let $\hat{Q}^i_{e-1}(s_t, a_t) \leftarrow \hat{Q}^i_{e-1}(s_t, a_t) + \alpha_e \big(r_{t+1} - \lambda_i c_{t+1} + \gamma \max_{a'} \hat{Q}^i_{e-1}(s_{t+1}, a') - \hat{Q}^i_{e-1}(s_t, a_t)\big)$ and update $t \leftarrow t + 1$

Update $\hat{Q}^i_e \leftarrow \hat{Q}^i_{e-1}$.

until $e \geq E_1$ **or** changes in $\hat{Q}^i$ are small

$e \leftarrow 0$.

repeat

$e \leftarrow e + 1$, initialize $t \leftarrow 0$ and sample $s_0 \sim \rho$

while s_t is not terminal **and** $t \leq M_2$ **do**

$\{\hat{v}_{cost}\}^i_{e-1}(s_t) \leftarrow \{\hat{v}_{cost}\}^i_{e-1}(s_t) + \alpha_e \big(c_{t+1} + \gamma \{\hat{v}_{cost}\}^i_{e-1}(s_{t+1}) - \{\hat{v}_{cost}\}^i_{e-1}(s_t)\big)$

Update $t \leftarrow t + 1$

Update $\{\hat{v}_{cost}\}^i_e \leftarrow \{\hat{v}_{cost}\}^i_{e-1}$.

until $e \geq E_2$ **or** changes in $\hat{V}^i_{cost}$ are small

Compute $\hat{\mathcal{D}}(\lambda_i) = \sum_s (\max_a \hat{Q}^i(s,a))\rho(s) + \lambda_i B$. Find $\pi^{\lambda_i}(s) = \text{argmax}_a \hat{Q}^i(s,a)$

Output: $\pi_1 = \pi^{\lambda_i}$ and $\pi_2 = \pi^{\lambda_{i'}}$ where $\lambda_i = \text{argmin}\{\hat{\mathcal{D}}(\lambda_j) | \sum_s \hat{v}^j_{cost}(s)\rho(s) \leq B\}$ and $\lambda_{i'} = \text{argmin}\{\hat{\mathcal{D}}(\lambda_j) | \sum_s \hat{v}^j_{cost}\rho(s) \geq B, \pi^{\lambda_j} \neq \pi_1\}$.

Theorem 5.4.2. *Assume $\rho(s) > 0\ \forall s$, the optimal policy $\pi^\star$ for problem* (5.1) *is unique and there exists some $\lambda^\star \in \operatorname{argmin} \mathcal{D}(\lambda)$ such that $\lambda_{min} < \lambda^\star < \lambda_{max}$. Fix $n \geq 0$, assume for each Q^i-learning problem and TD-learning problem for $1 \leq i \leq n$, every state and every state-action pair are visited infinitely often. Furthermore, sequence α_e satisfies*

$$\sum_e \alpha_e = \infty \quad \textit{and} \quad \sum_e \alpha_e^2 < \infty. \tag{5.20}$$

Then there exists N large enough and ϵ_g small enough such that if we fix $n = N$ and $\epsilon_{greedy} \leq \epsilon_g$, we will recover a pair of deterministic policies π_1, π_2 such that $\pi^\star = (1-t)\pi_1 + t\pi_2$ for some $0 \leq t \leq 1$ with probability 1 as the number of episode $E_1, E_2 \to \infty$.

Proof. Following Theorem 5.4.1, first consider the case where $\lambda^\star > 0$ is unique and $\mathcal{D}^-(\lambda^\star) < 0 < \mathcal{D}^+(\lambda^\star)$. Then, as discussed in Theorem 5.4.1, (5.11) and (5.12), there exist some $\epsilon > 0$ and policies π_1', π_2' which differ by one state such that $\pi^\star = (1-t)\pi_1' + t\pi_2'$ for some $0 < t < 1$,

$$\mathcal{D}(\lambda) = \mathcal{D}(\lambda^\star) + \mathcal{D}^+(\lambda^\star)(\lambda - \lambda^\star) = \mathcal{R}(\pi_1', \rho) - \lambda\big(C(\pi_1', \rho) - B\big) \tag{5.21}$$

for $\lambda^\star \leq \lambda \leq \lambda^\star + \epsilon$ and some deterministic π_1' while

$$\mathcal{D}(\lambda) = \mathcal{D}(\lambda^\star) + \mathcal{D}^-(\lambda^\star)(\lambda - \lambda^\star) = \mathcal{R}(\pi_2', \rho) - \lambda\big(C(\pi_2', \rho) - B\big) \tag{5.22}$$

for $\lambda^\star - \epsilon \leq \lambda \leq \lambda^\star$ and some deterministic π_2'. It is clear from the definition of $\mathcal{D}(\lambda)$ and our assumption on the uniqueness of $\pi^\star$ that $\pi_1' = \pi^\lambda$ for $\lambda^\star < \lambda < \lambda^\star + \epsilon$ and $\pi_2' = \pi^\lambda$ for $\lambda^\star - \epsilon < \lambda < \lambda^\star$. Then, for $n = N$ large enough, where $(\lambda_{max} - \lambda_{min})/N \leq \epsilon$, we must have some $\lambda^\star - \epsilon \leq \lambda_i \leq \lambda^\star \leq \lambda_{i+1} \leq \lambda^\star + \epsilon$ for some $1 \leq i \leq n$ and due to the strict convexity of $\mathcal{D}(\lambda)$ around $[\lambda^\star - \epsilon, \lambda^\star + \epsilon]$, we must have $\mathcal{D}(\lambda_i) < \mathcal{D}(\lambda_{i-1}) < \ldots < \mathcal{D}(\lambda_1)$ and $\mathcal{D}(\lambda_{i+1}) < \mathcal{D}(\lambda_{i+2}) < \ldots < \mathcal{D}(\lambda_n)$. Now, by the assumption on the Q-learning procedure (infinitely often visit for state-action pair under ϵ-greedy policy, the Robbins-Monro ([164]) type condition (5.20)), it follows that the Q^i-learning for every $1 \leq i \leq n$ converges to the optimal Q^i value (or ϵ_{greedy}-optimal

assuming *optimistic*, large initialization for Q values ([165])) and we can recover the optimal value (λ-adjusted) function $\max_a Q^i(s,a)$ with probability 1 as $E \to \infty$ ([160, 159, 166]). Thus, as $E \to \infty$, we will have $\hat{\mathcal{D}}(\lambda_i) < \hat{\mathcal{D}}(\lambda_{i-1}) < ... < \hat{\mathcal{D}}(\lambda_1)$ and $\hat{\mathcal{D}}(\lambda_{i+1}) < \hat{\mathcal{D}}(\lambda_{i+2}) < ... < \hat{\mathcal{D}}(\lambda_n)$. On the other hand, the assumption also guarantees that the TD learning on $\hat{v}^j_{cost}$ will converge to $v^{\lambda_j}_{cost}$ (or $v^{\pi^{\lambda_j}_{\epsilon_{greedy}}}_{cost}$, where $\pi^{\lambda_j}_{\epsilon_{greedy}}$ is the ϵ_{greedy} greedy policy from the optimal π^{λ_j}). If we pick $\epsilon_{greedy} > 0$ small enough, we can make $\sum_s |v^{\pi^{\lambda_j}}_{cost}(s) - v^{\pi^{\lambda_j}_{\epsilon_{greedy}}}_{cost}(s)|\rho(s)$ arbitrarily small. However, we know from the piece-wise linearity and convexity of $\mathcal{D}(\lambda)$ that, for all $\lambda_j \geq \lambda^\star$, the gradient $B - C(\pi^\lambda, \rho) > 0$ which implies $\sum_s v^{\pi^{\lambda_j}}_{cost}(s)\rho(s) = C(\pi^{\lambda_j}, \rho) < B$, and we can find ϵ_{greedy} small enough such that $\sum_s v^{\pi^{\lambda_j}_{\epsilon_{greedy}}}_{cost}(s)\rho(s) < B$ and thus (in both cases) $\sum_s \hat{v}^j_{cost}(s)\rho(s) < B$ with $\lambda_{i+1} = \text{argmin}\{\hat{\mathcal{D}}(\lambda_j) | \sum_s \hat{v}^j_{cost}(s)\rho(s) \leq B\}$ implying $\pi_1 = \pi'_1$ as $E_1, E_2 \to \infty$. Similarly we can show $\pi_2 = \pi'_2$. For other cases where $\lambda_\star = 0$ and one or both of $\mathcal{D}^+(\lambda^\star)$ and $\mathcal{D}^-(\lambda^\star)$ are 0, it can be shown that the unique deterministic policy $\pi^\star$ can be recovered. □

Theorem 5.4.2 guarantees that with suitable algorithmic parameter choices, Algorithm 3 can retrieve two candidate optimal policies such that their mixture gives rise to the optimal randomized policy for the constrained problem (5.1). Next we will discuss in more detail the implementation issues, including how to search for the mixture parameter.

5.5 Discussion and Implementation

Theorem 5.4.2 not only gives us theoretical guarantees on recovering the candidates for optimal mixtures, but also partially explains why the behavior of a direct primal dual method becomes unstable around convergence. In particular, the splitting of action forces the primal update to search for different optimal polices around the $\lambda^\star$ and makes the convergence especially difficult. To overcome such a difficulty, we use the mixing of policies which is to be explained later in this section. The discretization of dual variable λ is designed for this purpose as well. Notice this special discretization also allows for efficient parallel computing on different λ. On the other hand, the conditions can be restrictive in practice and the implementation for Algorithm 3 becomes inefficient as the accuracy parameters increase. In particular, there are several main issues concerning

the implementation of Algorithm 3:

1. How to find the a reasonable set of $\lambda_{min}, \lambda_{max}$?

2. What if Algorithm 3 cannot converge to the correct pair of policies (e.g. π_1 and π_2 differ by more than one state)?

3. Given two candidate policies π_1, π_2, and the results from Theorem 5.4.1 that $\pi^\star = (1-t)\pi_1 + t\pi_2$ for some $0 \leq t \leq 1$, how do we find t?

The first point is not a major concern. As mentioned, the dual variable λ is one-dimensional and we can use many efficient RL methods such as Q-learning. In fact, we can use RCPO efficiently before we run into convergence issues, at which point we can already observe a good range of dual value λ for which the optimal $\lambda^\star$ is likely to be contained in. To address the second and third issues, we note that in both minimizing $\mathcal{D}(\lambda)$ and mixing $\pi_t = (1-t)\pi_1 + t\pi_2$, it is critical to efficiently estimate $C(\pi, \rho)$ for a given policy π.

Cost Evaluation. Suppose we have found $\pi^\lambda \in \underset{\pi \in \Pi_0}{\text{argmax}}\, \mathcal{R}(\pi, \rho) - \lambda C(\pi, \rho)$. Then an estimate of $C(\pi^\lambda, \rho)$ can help evaluate a sub-gradient ([167]) of the piece-wise linear dual function $\mathcal{D}(\lambda)$, which is given by $B - C(\pi^\lambda, \rho)$. This in turn helps decide a search direction for $\lambda^\star$ based on first-order optimization methods. On the other hand, when mixing the policies $\pi_t = (1-t)\pi_1 + t\pi_2$, we know from duality that

$$\mathcal{R}(\pi^\star) = \mathcal{D}(\lambda^\star) = \mathcal{R}(\pi_t, \rho) - \lambda^\star\big(C(\pi_t, \rho) - B\big). \tag{5.23}$$

Thus, if we can find t such that $C(\pi_t, \rho) = B$, it then follows from (5.23) that policy π_t satisfies primal feasibility and optimality simultaneously and is the solution of (5.1).

There are many ways to estimate $C(\pi, \rho)$, e.g., TD-learning $\sum_s v_s \rho(s)$, or Monte Carlo by [159]. Thus, from now on we assume an efficient oracle $Eval_C(\pi, \rho)$ which takes as input policy π and initial distribution ρ and outputs an estimate of $C(\pi, \rho)$.

Dual Variable Range. Given the oracle $Eval_C(\pi, \rho)$, we can construct algorithms that effec-

Algorithm 4 Dual Variable Range Selection

Input: A threshold $0 < \theta < 1$ (e.g. $\theta = 1/2$), step size λ_{step} and a tolerance for budget constraint τ.
Initialization: λ, λ_{min},λ_{max} (e.g. 0)
Find π^λ by Q-learning
if $B - \tau \leq Eval_C(\pi^\lambda, \rho) \leq B + \tau$, **then**
    Break search and accept π^λ as optimal policy.
if $Eval_C(\pi^\lambda, \rho) < (1 - \theta)B$ **then**
    Set $\lambda_{max} = \lambda$, Break Search and restart algorithm with $\lambda \leftarrow \lambda - \lambda_{step}$. (Also Break if $\lambda_{max} = 0$, suggesting the MDP is unconstrained.)
if $Eval_C(\pi^\lambda, \rho) > (1 + \theta)B$ **then**
    Set $\lambda_{min} = \lambda$. Break Search and restart algorithm with $\lambda \leftarrow \lambda + \lambda_{step}$.

tively select a reasonable pair of λ_{min} and λ_{max}. In particular, given a $\lambda \geq 0$, if we have found π^λ by Q-learning on function $\mathcal{D}(\lambda)$, then by the convexity of $\mathcal{D}(\lambda)$, we know if $C(\pi^\lambda, \rho) > B$, it indicates $\lambda \leq \lambda^\star$ whereas if $C(\pi^\lambda, \rho) < B$, it indicates $\lambda \geq \lambda^\star$. Thus, we can make use of the oracle $Eval_C(\pi, \rho)$ to estimate $C(\pi, \rho)$. However, the estimate would inevitably be corrupted by noise so we want to ensure an empirically over-budget policy π (i.e. $C(\pi, \rho) > B$) is indeed over-budgeted, by setting a "safety margin" θ to account for statistical significance. For example, if $Eval_C(\pi^\lambda, \rho) > (1 + \theta)B$, then with high probability we have $C(\pi^\lambda, \rho) > B$ and we can set $\lambda_{min} = \lambda$. On the other hand, if during the search we have found a policy π^λ that is close to feasibility (i.e. $C(\pi^\lambda, \rho) \approx B$), then we make use of weak duality ([161]):

$$\mathcal{R}(\pi^\lambda, \rho) \approx \mathcal{R}(\pi^\lambda, \rho) - \lambda\big(C(\pi^\lambda, \rho) - B\big) = \mathcal{D}(\lambda) \geq \mathcal{R}(\pi^\star, \rho),$$

and accept π^λ as a near-optimal, near-feasible solution. Of course such cases will not occur in general. Based on these discussion, we propose one possible Algorithm 4.

Feasibility Mixing. As we have discussed in (5.23), we need to build an oracle that given two policies π_1, π_2 with $C(\pi_1, \rho) \leq B$ and $C(\pi_2, \rho) \geq B$, we can find $\pi_t = (1 - t)\pi_1 + t\pi_2$ satisfying $C(\pi_t, \rho) = B$. Here we make use of oracle $Eval_C$ again to present an approximate algorithm that combines linear interpolation and bisection to quickly search for a feasible policy. Specifically, for the interpolation part, we notice that, for $L \leq B \leq U$, $(1 - t)L + tU = B \quad \text{where} \quad t = \frac{B-L}{U-L}$. In

Algorithm 5 Feasibility Mixing

Input: policies π_1, π_2 with $Eval_C(\pi_1, \rho) \leq B$, $Eval_C(\pi_2, \rho) \geq B$, a tolerance for the budget τ
Initialize: $t \leftarrow \frac{B - Eval_C(\pi_1, \rho)}{Eval_C(\pi_2, \rho) - Eval_C(\pi_1, \rho)}$, (or $i \leftarrow 1$, $t_i \leftarrow 1/2$ for direct bisection)
Set policy $\pi_t = (1-t)\pi_1 + t\pi_2$ $B - \tau \leq Eval_C(\pi_t, \rho) \leq B + \tau$, Break search and accept π_t as optimal policy.
if $Eval_C(\pi_t, \rho) < B - \tau$ **then**
    Update $\pi_1 \leftarrow \pi_t$ and $t \leftarrow \frac{B - Eval_C(\pi_1, \rho)}{Eval_C(\pi_2, \rho) - Eval_C(\pi_1, \rho)}$, (or $i \leftarrow i + 1$ $t \leftarrow t + 1/2^i$)
if $Eval_C(\pi_t, \rho) > B + \tau$ **then**
    Update $\pi_2 \leftarrow \pi_t$ and $t \leftarrow \frac{B - Eval_C(\pi_1, \rho)}{Eval_C(\pi_2, \rho) - Eval_C(\pi_1, \rho)}$, (or $i \leftarrow i + 1$ $t \leftarrow t - 1/2^i$)
Output: t (or π_t).

practice, we may use a direct bisection. Feasibility mixing is especially practical because we might only obtain approximately optimal candidate policies π'_1, π'_2 (i.e. they might not be the optimal pair of polices) under two dual variables λ'_1 and λ'_2 (i.e. they might be different from the desired λ_1 and λ_2 in Theorem 5.2) from Algorithm 3 that in turn might only be approximately optimal for λ'_i (meaning that $\mathcal{R}(\pi'_i.\rho) - \lambda'_i(C(\pi'_i, \rho) - B) \leq \mathcal{D}(\lambda_i)$). However, based on the piecewise-linearity and the convexity of $\mathcal{D}(\lambda)$, as long as feasibility mixing is performed, it is straight-forward to show that the reward function of the mixing policy π_t satisfies $\mathcal{D}(\lambda^\star) - \mathcal{R}(\pi_t, \rho) = O(\epsilon_1 \cdot \epsilon_2 \cdot \epsilon_3)$ where $\epsilon_1 = \max_{1 \leq i \leq 2} |\lambda_i - \lambda^\star|$, $\epsilon_2 = \max_{1 \leq i \leq 2} |\mathcal{D}(\lambda_i) - \mathcal{D}(\lambda^\star)|$ and $\epsilon_3 = \max_{1 \leq i \leq 2} |\mathcal{R}(\pi'_i.\rho) - \lambda'_i(C(\pi'_i, \rho) - B)|$.

5.6 Numerical Experiments

5.6.1 Environment Description and Setup

We evaluate the proposed algorithms on a real world dataset collected from [anonymized for review purpose] during a sponsored search campaign portfolio which spans over six months and contains over a million distinct user search trajectories. The dataset provides ad click records of anonymous users before conversion with their corresponding timestamps. The ad click records are associated with a matching of the user's query with a keyword group. This particular dataset has ten different keyword groups each containing hundreds of keywords. Similar to other advertiser-specific data, we do not directly observe the events in which the users did not click on the ad. Similarly, the data does not record the searches for which the ad was not shown to the user for

any reason such as low bid values, budget constraint, etc. On the other hand, a smaller version of the experiment allows a clear validation of our key theorem on policy splitting, because the optimal policy and its two splitting policies in a CMDP is difficult to recover in complicated, large MDPs. However, we note that our algorithm allows for larger experiments in a model-free algorithm setting.

For the experiment setup, we first retrieve the cost information for our sampled dataset with CPC (cost per click) metric averaged at the keyword group level for the similar time period as the collected data. The average cost for the ten keyword groups in our experiment is estimated to be $[0.2, 0.4, 0.25, 0.5, 0.3, 0.6, 0.5, 0.3, 0.3, 0.4]$ in dollars. Additionally, the reward for converting a user is estimated to be worth \$10 for this campaign. Then, we follow the framework in [168] to establish a CMDP. In particular, user state represents the matching of the user's last query with any of the keyword groups that translates to ten states in our experiment. Then, our action space is binary and includes "advertise" and "do not advertise" actions and transition probabilities between states are directly estimated from the data. In order to overcome the issue of estimating transition probabilities for "do not advertise", we follow the remedy suggested by [168]. That is, we assume the transitions between states are independent of the ad presented to the user if the time period between two consecutive searches is longer than one day. Moreover, we bundle all possible advertisement keywords in 10 keyword groups. Finally, we add 4 states, which contain a beginning state, a conversion state, a non-conversion state and eventually the final state to incorporate the situation where users may convert temporarily but eventually become disinterested in the ad push (see Figure 5.1). Consequently, we have 14 states in our environment in total with a transition probability matrix in $\mathbb{R}^{2\times 14\times 14}$. We run Algorithm 3 with hyper-parameters $\lambda_{min} = 0$, $\lambda_{max} = 2$, $M_1 = 10^5$, $E_1 = 3.5 \times 10^5$, $M_2 = 10^4$, $E_2 = 2 \times 10^5$, $\alpha_e = \frac{9}{9+0.2e}$, $\epsilon_{greedy} = 0.2$, $B = 0.45$, $\gamma = 0.6$, $\tau = 10^{-4}$ and early stopping criterion requires $\|\cdot\|_\infty$ norm within 10^{-4}. The metrics here for reward and cost are averaged accumulative rewards and averaged accumulative costs defined in (1), In order to show the advantage of our method, we pick RCPO as a baseline. For the sake of fairness, all experiments are implemented in `Python 3.7` and executed on a standard `1.7 GHz`

Dual-Core Intel Core i7.

Figure 5.1: MDP on advertisement (red node denotes a conversion/non-conversion state).

5.6.2 Algorithm Performances

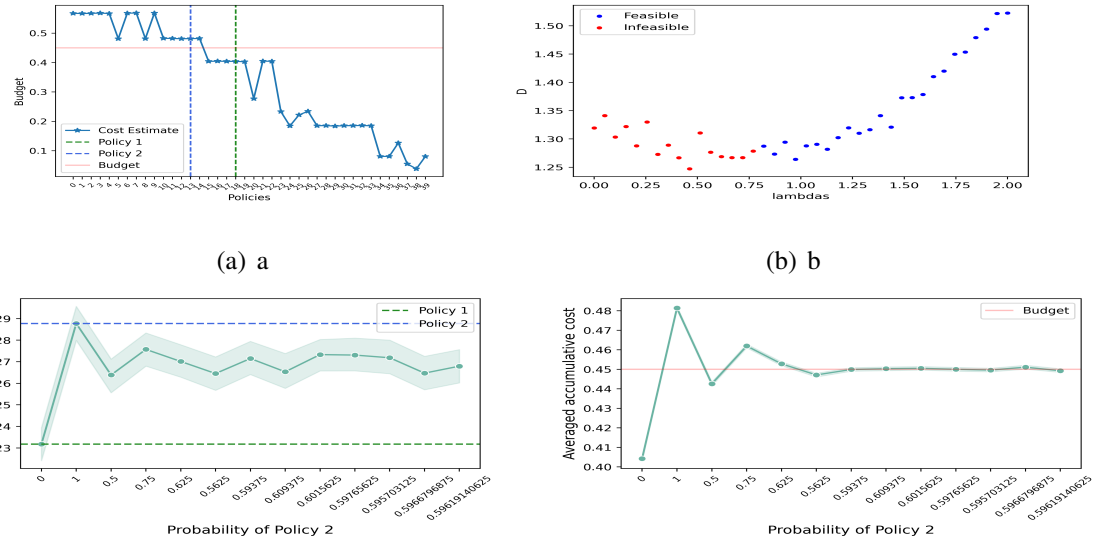
Figure 5.2(a) demonstrates the averaged accumulative costs of the two candidate policies (Policy 1 and Policy 2) selected by Algorithm 3. Moreover, for each λ, $\mathcal{D}(\lambda)$ can be computed efficiently with RL-methods and its convexity is shown in Figure 5.2(b). After identifying two candidate policies from Algorithm 3, we run Algorithm 5 which mixes the policies to satisfy the budget constraint. As shown in Figure 5.2(c), we start with Policies 1 and 2 corresponding to $t = 0$ and 1 and use a simple bisection to search for the target value of t. Figure 5.2(d) shows the searching process stabilizes after a few iterations and the corresponding long-run budget for different mixture policies gradually converges to the target budget value. As we expect, in this case the optimal policy comes from the mixture, one policy going over budget and the other under.

To show the robustness of the procedure, we perform a large number of experiments to see the effectiveness of Algorithm 3 in recovering the correct pair of optimal policies. Figure 5.3 (a)(b) shows that, in this example, the correct pair of policies can be recovered in 78% of the experimental repetitions. More importantly, we plot the distribution of the reward-budget pairs of the resulting mixture policy across all experiments and show that, among the occasions Algorithm 3 does not pick the correct pair, the resulting mixture is still approximately optimal and feasible, within a controllable error margin, showing the stability of the procedure. In addition, we compare the performances between our method and RCPO. As shown in Figure 5.3(c), the learning curve on rewards of RCPO is between the learning curves of two candidate policies. However, as shown in Table 5.1 and 5.3(d), our mixing method can find a randomized policy that has a higher average accumulative reward in lesser time. As discussed, RCPO converges fast initially, yet the convergence slows down and exhibits a zigzag motion when it is quite close to the optimal λ. Advantageously,

our mixing method bypass this problem around convergence.

Methods	Accumulative Rewards	Accumulative Costs	Clock Time (s)
RCPO	1.229	**0.405**	924.961
Policy Mixing (τ = 1e-4)	1.271	**0.449**	839.708
Policy Mixing (τ = 1e-3)	**1.277**	**0.449**	702.927
Policy Mixing (τ = 1e-2)	1.276	**0.449**	**558.763**

Table 5.1: Performance comparison summary (Bold means either the best or valid).



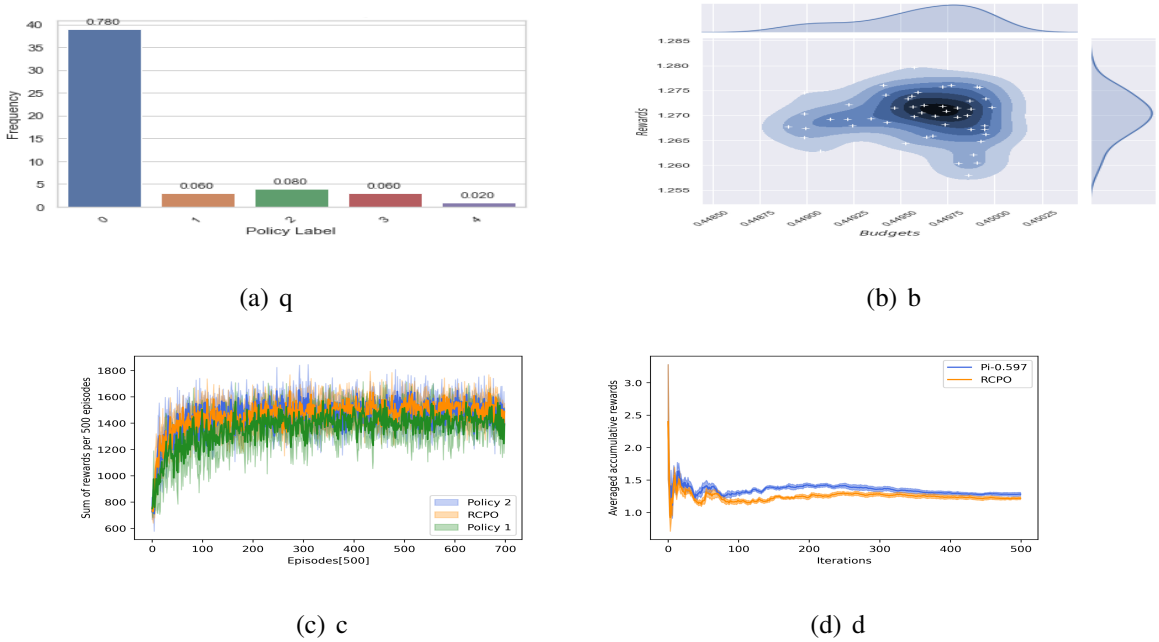
(a) a

(b) b

(c) c

(d) d

Figure 5.2: (a) Budget estimates of policies with different λ; (b) Convexity of $D(\lambda)$; (c) Accumulative adjusted rewards during policy mixing; (d) Accumulative costs during policy mixing.



(a) q (b) b

(c) c (d) d

Figure 5.3: (a) Occurrences of policy pairs (Label 0 denotes valid policy pairs with only one state with different actions); (b) Joint distribution of averaged reward and cost, where each dot represents each experiment and the heat map is estimated from kernel density estimation; (c) Learning curves of policies 1, 2 and RCPO. A tick on x-axis denotes 500 episodes and y-axis denotes the total rewards for every 500 episodes; (d) MC evaluation of averaged accumulative rewards.

5.7 Conclusion

We focus on solving CMDPs which, although arise frequently in practice, are not amenable to efficient solution techniques offered by most established RL-methods on unconstrained problems. Through incorporating the “splitting" property of CMDP in a Lagrangian formulation, our approach investigates the potential issues around convergence for current primal-dual RL-methods and offers a suitable alternative. The approach aims to identify two candidate optimal policies which through mixing would result in an optimal randomized policy of the CMDPs. We illustrate our performances through an online advertising problem with budget calibrated by real-world data.

Chapter 6: Unbiased Sampling of Multidimensional Partial Differential Equations with Random Input

Partial differential equations (PDEs) are important tools for modeling physical or financial systems. However, intrinsic variability of the system or measurement errors bring uncertainty into the model and are commonly represented by random input data. In this chapter, we use multilevel Monte Carlo (MLMC) to construct unbiased estimators for expectations of random parabolic PDE. Building on previous works of Giles (2008) and Li et al.(2018), we obtain estimators with finite variance and finite expected computational cost, but bypassing the curse of dimensionality. For the error analysis in random PDE, we combine rough path theory with numerical stochastic analysis in a novel way.

The rough path part is mostly proof and are left in the Supplementary. Interesting readers can turn to [169]. The use of MLMC in random PDE can be justified by the need for an unbiased estimator and Feynman-Kac formula.

6.1 Introduction

The heat equation is a classic PDE with many applications. In different contexts, the interpretations for the coefficients of PDE vary. In heat conduction, the equation follows from Fourier's law and the solution represents the temperature of the material, while the coefficients characterize the thermal conductivity of the material. In flow dynamics [170], the heat equation follows from Darcy's law for describing the flow of fluids through a porous medium, where the solution represents the fluid pressure and the coefficients characterize the medium permeability, analogous to Fick's second law in diffusion theory. In mathematical finaince, the heat equation governs the risk-neutral pricing of European-style options with given payoff at maturity where coefficients

represent properties of financial markets and underlyiung assets, including risk-free rate, drift rate, volatility, etc [171]. In general, the coefficients in the heat equation reflect properties of medium or underlying systems. In practice, either due to microscopic heterogeneity of the media, intrinsic variability of the system or measurement error from experiments, the coefficients in the PDE are inherently uncertain and are modeled as random fields in probability space. Related literature on the modeling and analysis for the heterogeneous random medium includes [172, 173, 174, 175]. On the other hand, in derivative pricing, while the diffusion coefficient σ can be estimated reasonably accurate due to the characteristics of quadratic variations, the drift coefficient μ is typically difficult to calibrate and modeled as random variable [176, 177].

Specifically, in this chapter we consider a random parabolic partial differential equation (PDE) $u : \mathcal{X} \times \mathbb{R}^+ \to \mathbb{R}$ on a simply connected and compact domain $\mathcal{X} \subseteq \mathbb{R}^d$:

$$\partial_t u(x,t) = \boldsymbol{\mu}^T(x) D_x u(x,t) + \frac{1}{2} trace\left(\sigma(x)\,\sigma^T(x)\,D_{xx}u(x,t)\right), \tag{6.1}$$

with known initial condition $u(\cdot,0) = f(\cdot) : \mathbb{R}^d \to \mathbb{R}$, where $\sigma(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d\times d'}$ is known (sometimes implicitly) but $\boldsymbol{\mu}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is a random field on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Here D_x and D_{xx} denote the first and second order partial derivatives operators, while $trace(\cdot) : \mathbb{R}^{d\times d} \to \mathbb{E}$ denotes the trace operator for matrices. Notice randomness propagates in (6.1) through $\boldsymbol{\mu}(\cdot)$. The solution u is implicitly determined by $\mu(\cdot,\omega)$, the realization of $\boldsymbol{\mu}$ and is hence also random, henceforth denoted as $\boldsymbol{u}$. However, for brevity, we suppress its dependence on Ω and still write $\{u(x,t)\}_{(x,t)\in\mathcal{X}\times\mathbb{R}^+}$ instead of $\{u(x,t,\omega)\}_{(x,t)\in\mathcal{X}\times\mathbb{R}^+}$ for the realization of $\boldsymbol{u}$. Generally, we are interested in estimating statistics or functionals of $\boldsymbol{u}$ (failure probability, moments estimation, e.g.) [173, 178]. As dependence of u on μ is typically implicit and in non-closed form, a popular tool for studying distributional property of $\boldsymbol{u}$ is Monte Carlo method. In particular, we study expectations of the form

$$\nu = \mathbb{E}\left[G(\boldsymbol{u}(x_1,t_1),...,\boldsymbol{u}(x_k,t_k))\right], \tag{6.2}$$

for any $\{(x_i,t_i)\}_{i\in[k]} \subseteq \mathcal{X}\times\mathbb{R}^+$, $k \in \mathbb{Z}^+$ and $G : \mathbb{R}^k \to \mathbb{R}$ satisfying certain regularity conditions.

Notice

$$\{G(\cdot) : G(u) = G(u(x_1, t_1), ..., u(x_k, t_k)) \text{ for } k \in \mathbb{Z}^+ \text{ and } \{(x_i, t_i)\}_{i\in[k]} \subseteq \mathcal{X} \times \mathbb{R}^+\}$$

only constitutes a proper subset of all functionals of u. However, as we shall see, this form of G allows us to bypass the curse of dimensionality and make arbitrarily fine approximation for a wide range of functionals on u, as k gets large.

6.1.1 Background and review of related results

In this chapter, we provide an unbiased estimator for v in (6.2) and could be efficiently implemented by parallel computing architectures. Due to ease of implementation, Monte Carlo method has been widely used for solving PDEs with random input, including quasi-Monte Carlo and multilevel Monte Carlo methods [179, 180, 181]. On the other hand, spectral stochastic methods, with faster convergence rates but suffering from the curse of dimensionality, are also popular for moderate dimensional problems and include stochastic Galerki method and stochastic collocation method [178, 182, 183]. In general, all such methods, including Monte Carlo methods, require approximations to the PDE solution, using determinsitic slovers such as the finite elements method (FEM), the finite difference method (FDM),the finite volume method, etc [184, 185, 186]. In particualr, recent development from [187] combines multilevel Monte Carlo (see [188, 189, 190]), a randomization scheme (see [191]) and FEM to build an unbiased estimator for the solution of elliptic equations with random inputs and Dirichlet boundary conditions. The variance and the expected computational cost of generating such an estimator are shown to be finite. However, as the error analysis based on FEM depends on the underlying dimension d, even though the sampling strategy in [187] achieves a square-root convergence rate, the estimator can achieve both finite variance and finite expected computational cost only when $d \leq 3$. In other words, similar as a substantial amount of recent literature combining the multilevel Monte Carlo technique with the numerical methods for PDE, the procedure in [187] suffers from the curse of dimensionality, as the

rate of convergence (for the numerical solver of PDE and for the estimator) deteriorates with the increase of problem dimensions [184, 179, 186]. On the other hand, Monte Carlo methods with better dependence on problem dimension d are available, but they produce biased estimators [183, 190, 180].

6.1.2 Contribution

The contribution of this chapter is to introduce an unbiased estimator for ν with finite variance, finite expected computational cost to generate and for arbitrary dimension d. Consequently, our method allows for a full Monte Carlo procedure with a traditional square-root convergence rate for any dimension d, therefore preserving the well-known characteristic of the Monte Carlo method in combating the curse of dimensionality. Thus, if the parallel computing cores are relatively cheap and wall-clock time is a relatively hard constraint, one can then independent copies the estimator in parallel servers and combine them to provide confidence intervals with squared-root convergence rate for any d.

The technical contribution of this chapter is potentially of interest in its own right. In order to bypass the curse dimensionality, the construction of the estimator avoids the numerical approximations of PDE (e.g., FDM, FEM) and instead exploits the connection between the parabolic PDEs and stochastic differential equations (SDEs) using the Feynman-Kac formula. Thus, instead of discretizing the mesh size of numerical PDE, we discretize the step for simulating the path of SDE, combining multilevel Monte Carlo [181] with randomization step [191] and an additional randomization from [192] canceling the bias incurred from randomness of $\boldsymbol{\mu}$. The difficulty arises from the this additional randomization step and requires a non-standard technical development. In particular, error analysis in numerical SDE [181, 185] commonly relies on Gronwall's inequality [193]. However, if the same stochastic analysis were applied here, the estimator could not be shown to exhibit both finite variance and finite expected computational cost. To overcome this issue, we turn to the theory of rough paths to obtain "path-by-path" estimates. The rough path theory [194, 195, 196, 197] has received substantial attention in recent literature due to connections to the

theory of regularity structures and nonlinear stochastic PDEs [198]. Even though a considerable amount of literature has been devoted to explore the relations between the theory of rough paths and stochastic numerical analysis in the context of cubature methods [199] or SDEs [169, 200], rough paths estimates have yet to be connected with numerical analysis of random PDEs. In this chapter, we are able to brige this gap which also allows us to overcome our technical difficulty, adding to the literature combining rough paths theory with numerical stochastic analysis.

The rest of the chapter is organized as follows. In Section 2 we lay out notations and assumptions used throughout the chapter. In Section 3 we present preliminary material and roadmap towars the construction of the unbiased Monte Carlo estimator. In Section 4, we provide theoretical analysis and proofs for properties of the estimator. In Section 5, we present simulation studies on numerical experiments. Finally, proofs omitted in the main sections can be found in the Supplementary. We also include a supplementary material for additional technical proofs.

6.2 Preliminaries

6.2.1 Notations and assumptions

We use the following notations and terminology throughout the chapter. The Frobenius norm of vectors and matrices is $\|\cdot\|_F$. The supreme norm is denoted as $\|\cdot\|_\infty$. The d-dimensional Gaussian random vector with mean $\theta \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is denoted $\mathcal{N}(\theta, \Sigma)$. For a natural number $k \in \mathbb{Z}^+$, we denote $[k]$ to be the set $\{1, ..., k\}$. As before, D_x and D_{xx} denote the first and second order (partial) derivatives operators with respect to variables in $\mathcal{X}$. We also use ∂ operator to specify the components of differentiation. For use $X \overset{\mathcal{D}}{=} Y$ to denote two random variables (or stoachastic process) equal in distribution.

Moreover, given $L > 0$, we denote $\mathcal{L}(L)$ to be the space of bounded, Lipschitz continuous and twice continuously differentiable functions defined on $\mathbb{R}^d$ (range not sepcified) such that $\mu \in \mathcal{L}(L)$ if

$$\|\mu\|_\infty \le L, \quad \|D_x\mu\|_\infty \le L \quad \text{and} \quad \|D_{xx}\mu\|_\infty \le L. \tag{6.3}$$

It is worth noting that the analysis somtimes simplifies when we focus on $\mathcal{L}(L)$ for $L > 1$. However, since $\mathcal{L}(L_1) \subseteq \mathcal{L}(L_2)$ for $L_2 \geq L_1$, we always assume $L > 1$ without loss of generality when we say $\mu \in \mathcal{L}(L)$. We also write $\mu \in \mathcal{L}$ when the constant L exists but does not need to be specified. We denote $poly(\cdot)$ (or $poly(\cdot, \cdot)$, $poly(\cdot, \cdot, \cdot)$, etc) to be a (multivariable) polynomial function.

Throughout the chapter, we assume the following regularity conditions. First, we need a Karhunen-Loève type of representation for the random field $\boldsymbol{\mu}$.

Assumption 5. *The random field $\boldsymbol{\mu} : \mathcal{X} \times \Omega \to \mathbb{R}^d$ has the following expansion*

$$\mu(\cdot, \omega) = \sum_{i=1}^{\infty} \frac{\lambda_i}{i^q} \cdot V_i(\omega) \cdot \psi_i(\cdot), \tag{6.4}$$

where $q > 4$ is a fixed constant, $\{\lambda_i\}_{i \geq 1} \subseteq \mathbb{R}$ is uniformly bounded, $\{\boldsymbol{V}_i\}_{i \geq 1}$ is a sequence of i.i.d. $\mathcal{N}(\mathbf{0}, \Sigma_i)$ and $\psi_i(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a sequence of deterministic functions. Moreover, there exists a constant $L > 1$ such that for $i \geq 1$,

$$\max_i \|\Sigma_i\|_F < L, \quad \|\psi_i\|_\infty < L, \quad \|D_x \psi_i\|_\infty < iL, \text{ and } \quad \|D_{xx} \psi_i\|_\infty < i^2 L. \tag{6.5}$$

In fact the proof does not require the assumption on $\{\boldsymbol{V_i}\}_i$ being Gaussian. We only need the tails of $\{\|\boldsymbol{V}_i\|_\infty\}_i$ to decay exponentially fast and uniformly in i (see Supplementary).

Given the representation in (6.4), we provide a technical lemma. Denote $\boldsymbol{S}_n$ as the partial sum process for $\boldsymbol{\mu}$ in (6.4):

$$S_n(\cdot, \omega) = \sum_{i=1}^{n} \frac{\lambda_i}{i^q} \cdot V_i(\omega) \cdot \psi_i(\cdot).$$

Lemma 13. *Under Assumption 5, there exists a random variable $\boldsymbol{L}_1 > 1$ on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}(e^{t\boldsymbol{L}_1}) < \infty$ for $t \in \mathbb{R}^+$ and $\{\boldsymbol{\mu}\} \cup \{\boldsymbol{S}_n\}_{n \geq 1} \subseteq \mathcal{L}(\boldsymbol{L_1})$ almost surely.*

We also need the following smoothness conditions on the deterministic functions σ, f and G.

Assumption 6. *There exists a constant $L > 1$ such that $\sigma(\cdot)$, $f(\cdot)$ and $G(\cdot)$ defined in* (6.1) *and* (6.2) *are in $\mathcal{L}(L)$.*

6.2.2 Definitions

In this section we present definitions from antithetic multilevel Monte Carlo for SDEs [181] related to our estimator. Let $B(\cdot)$ be a d'-dimensional standard Brownian motion and $B_j(\cdot)$ be its j-th component for $j \in [d']$.

Definition 2. *For $n \geq 0$, define $\Delta t_n \triangleq 2^{-n}$ and $t_k^n \triangleq k\Delta t_n$. Define $\Delta B_k^n \triangleq B(t_{k+1}^n) - B(t_k^n)$ and $\Delta B_{j,k}^n \triangleq B_j(t_{k+1}^n) - B_j(t_k^n)$ as the Brownian increments of step size Δt_n at t_k^n and its j-th component for $j \in [d']$.*

Definition 3. *Given $n \geq 0$ and a sequence of Brownian increments $\{\Delta B_k^n\}_{0 \leq k \leq 2^n - 1} \subseteq \mathbb{R}^{[d']}$. For $i, j \in [d']$, $\widetilde{A}_{ij}$ is defined on $\{(t_k^n, t_{k+1}^n)\}_{0 \leq k \leq 2^n - 1}$:*

$$\widetilde{A}_{ij}(t_k^n, t_{k+1}^n) \triangleq \frac{\Delta B_{i,k}^n \cdot \Delta B_{j,k}^n - \boldsymbol{I}_{ij} \cdot (\Delta t_n)}{2}, \tag{6.6}$$

where $\boldsymbol{I}$ is the d'-dimensional identity matrix.

Definition 4. *Given $n \geq 0$ and a sequence of Brownian increments $\{\Delta B_k^n\}_{0 \leq k \leq 2^n - 1}$, define the sequence of antithetic Brownian increments $\{\Delta B_k^{n,a}\}_{0 \leq k \leq 2^n - 1}$ as*

$$\Delta B_{2m}^{n,a} \triangleq \Delta B_{2m+1}^n \quad \text{and} \quad \Delta B_{2m+1}^{n,a} \triangleq \Delta B_{2m}^n \qquad \text{for } 0 \leq m \leq 2^{n-1} - 1\,. \tag{6.7}$$

The $\widetilde{A}_{ij}$ for $i, j \in [d']$ in (6.6) can be equivalently defined for $\{\Delta B_k^{n,a}\}_{0 \leq k \leq 2^n - 1}$. We denote it as $\widetilde{A}_{ij}^a$. It also follows from Definition 4, given Brownian motion $B(\cdot)$,

$$\Delta B_{2k}^n + \Delta B_{2k+1}^n = \Delta B_k^{n-1} = \Delta B_{2k+1}^{n,a} + \Delta B_{2k}^{n,a} \tag{6.8}$$

for $n \geq 1$, $0 \leq k \leq 2^{n-1} - 1$.

Moreover, it is easy to check that $\{\Delta B_k^n\}_{0 \leq k \leq 2^n - 1} \overset{\mathcal{D}}{=} \{\Delta B_k^{n,a}\}_{0 \leq k \leq 2^n - 1}$ for $n \geq 0$.

Definition 5. *Given the representation of* $\boldsymbol{\mu}$ *in* (6.4) *and* $\gamma > 0$*, define* $\boldsymbol{\mu}^{(n)}$ *as*

$$\mu^{(n)}(\cdot, \omega) \triangleq \sum_{i=1}^{\lfloor 2^{n\gamma} \rfloor} \frac{\lambda_i}{i^q} \cdot V_i(\omega) \cdot \psi_i(\cdot),$$

and the i*-th component of* $\boldsymbol{\mu}^{(n)}$ *is denoted* $\boldsymbol{\mu}_i^{(n)}$ *for* $i \in [d]$.

Note in Definition 5 we supress the dependence on γ as we treat it as one of the hyperparameters which is considered fixed throughout the chapter. The details will be provided in the sequel. Finally, it follows directly that $\{\boldsymbol{\mu}^{(n)}\}_{n\geq 0} \subseteq \mathcal{L}(\boldsymbol{L}_1)$ for the same $\boldsymbol{L}_1$ in Lemma 13.

6.3 Construction of the unbiased estimator

We denote W as our unbiased estimator for v in (6.2). In this section we present the construction of W in several steps. For ease of presentation, we illustrate the case for $k = 1$ and $t = 1$ in (6.2). The case for general $(k, t) \in \mathbb{Z}^+ \times \mathbb{R}^+$ follows in a straightforward manner.

6.3.1 Probabilistic representation of $u(x, t)$

For $\mu \in \mathcal{L}$, the solution $u(x, t)$ in (6.1) is connected to a d-dimensional diffusion process by the Feynman-Kac formula. For a brief introduction on SDE and Feynman-Kac formula, see, e.g., [201, 202].

Proposition 7. *Suppose* $(x, t) \in \mathcal{X} \times \mathbb{R}^+$ *and* $\mu(\cdot) \in \mathcal{L}$ *in* (6.1)*. Then under Assumption 6, solution of the PDE in* (6.1) *satisfies*

$$u(x, t) = \mathbb{E} f(X_t), \tag{6.9}$$

where the expectation is taken w.r.t. to the d*-dimensional diffusion process* $\{X_s\}_{0\leq s\leq t}$ *with* $X_0 = x$ *and governed by the SDE (i.e., the unique strong solution):*

$$dX_s = \mu(X_s)dt + \sigma(X_s)dB_s \tag{6.10}$$

for $0 \leq s \leq t$*. Here* B_s *is a* d'*-dimensional Brownian motion.*

Proof. For $\mu \in \mathcal{L}$, the existence and uniqueness of strong solution $\{X_s, 0 \le s \le t\}$ follow from the Lipschitz condition on $\sigma(\cdot)$. The rest follows from the Feynman-Kac formula (see Section 4.4 in [201]). □

We motivate the construction of W in two steps. First, given $\mu(\cdot, \omega)$, we construct an estimator $Z(\mu)$ such that $\mathbb{E}Z(\mu) = u(x, 1)$ (we are letting $t = 1$ w.l.o.g). We write $Z(\mu)$ to stress that Z is constructed while keeping the random field realization μ fixed. Here the expectation is not taken w.r.t the randomness in $\boldsymbol{\mu}$ but the randomness in the estimator Z itself. Next, we construct estimator $W(\mu)$ such that $\mathbb{E}W(\mu) = G(\mathbb{E}Z(\mu)) = G(u(x, 1))$, again with μ fixed. After these two steps, we can sample $\boldsymbol{\mu}$ and construct $W = W(\boldsymbol{\mu})$ as above. Then, the unbiasedness of W follows:

$$\mathbb{E}W = \mathbb{E}[\mathbb{E}[W(\mu)|\boldsymbol{\mu} = \mu]] = \mathbb{E}[\mathbb{E}[G(u(x, 1))|\boldsymbol{\mu} = \mu]] = \mathbb{E}[G(\boldsymbol{u}(x, 1))] = \nu. \tag{6.11}$$

Notice this construction does not guarantee the finite variance or finite expected computational cost of W for arbitary dimension d. For now we focus on the construction of Z and W.

6.3.2 Multilevel Monte Carlo

Section 6.3.1 allows for estimators based on discretization schemes for SDEs (e.g., Euler scheme, Milstein scheme, see [185]) rather than the ones for PDEs (e.g., FEM, FDM), which do not suffer from curse of dimensionality in the context of linear parabolic PDEs. However, estimators directly from numerical schemes are biased. The multilevel Monte Carlo method (MLMC) combines different "levels" of numerical estimators [189, 190]. In particular,

$$Z_{\text{MLMC}} = \sum_{n=0}^{N} \frac{1}{M_n} \sum_{i=1}^{M_n} \Delta_n^{(i)} + \frac{1}{N_0} \sum_{i=1}^{N_0} f(X_0^{(i)}(1)). \tag{6.12}$$

where $\{\Delta_n^{(i)}\}_{i\in[M_n]}$ and $\{X_0^{(i)}(1)\}_{i\in[M_0]}$ are generally I.I.D. copies. Here, Δ_n is any estimator satisfying

$$\mathbb{E}\Delta_n = \mathbb{E}f(X_{n+1}(1)) - \mathbb{E}f(X_n(1)), \tag{6.13}$$

where $X_m(1)$ for $m \geq 0$ corresponds to any discretization scheme for SDE solution X_t from (6.10) at $t = 1$ and m is a generic index indicating the level of discretization. N is a truncating integer (typically large) for the telescope sum

$$\mathbb{E}Z_{\text{MLMC}} = \sum_{n=0}^{N} \mathbb{E}f(X_{n+1}(1)) - \mathbb{E}f(X_n(1)) + \mathbb{E}f(X_0(1)) = \mathbb{E}f(X_{N+1}(1)),$$

to control bias. However, the advantage of MLMC is the variance reduction from the efficient coupling in Δ_n [188, 189]. In fact, the finite variance of our estimator hinges on Δ_n proposed by so called antithetic MLMC for multidimensional SDEs [181].

However, in this chapter, we do not assume we can explicitly sample $\boldsymbol{\mu}$ in (6.4). Thus, we can not directly apply the antithetic MLMC in [181] as we need to approximate the random field by $\boldsymbol{\mu}^{(n)}$ in Definition 5. We summarize our discretization scheme into the following Algorithm: Num_Sol$(\cdot,\cdot,\cdot)$. Notations in Algorithm Num_Sol$(\cdot,\cdot,\cdot)$ are defined in Section 6.2.2, and $\gamma > 0$ is considered fixed, to be sepcified later.

Algorithm 6 Num_Sol: discretization scheme for SDE

procedure NUM_ SOL $(x, n, \{\Delta B_k^n\}_{0\leq k\leq 2^n-1})$

input: starting point $x \in \mathbb{R}^d$, discretization level $n \geq 0$ and Brownian increments $\{\Delta B_k^n\}_{0\leq k\leq 2^n-1}$.

$X_n(0) \leftarrow x, \mu^{(n)}(\cdot) \leftarrow \sum_{i=1}^{\lfloor 2^{n\gamma}\rfloor} \frac{\lambda_i}{i^q} V_i \phi_i(\cdot)$

compute $\{\widetilde{A}_{ij}(t_k^n, t_{k+1}^n)\}_{i,j\in[d'],0\leq k\leq 2^n-1}$ from $\{\Delta B_k^n\}_{0\leq k\leq 2^n-1}$

for $0 \leq k \leq 2^n - 1$ and $i \in [d]$ **do**

$$X_{i,n}(t_{k+1}^n) \leftarrow X_{i,n}(t_k^n) + \mu_i^{(n)}(X_n(t_k^n))\Delta t_n + \sum_{j=1}^{d'} \sigma_{ij}(X_n(t_k^n))\Delta B_{j,k}^n$$
$$+ \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \frac{\partial \sigma_{ij}}{\partial x_l}(X_n(t_k^n))\sigma_{lm}(X_n(t_k^n))\widetilde{A}_{mj}(t_k^n, t_{k+1}^n)$$

output: $\{X_{i,n}(t_k^n)\}_{i\in[d],0\leq k\leq 2^n}$ (or $\{X_n(t_k^n)\}_{0\leq k\leq 2^n} \subseteq \mathbb{R}^d$)

From (6.8) in Remark 6.2.2, given $\{\Delta B_k^{n+1}\}_{0\leq k\leq 2^{n+1}-1}$, we create $\{\Delta B_k^n\}_{0\leq k\leq 2^n-1}$ and $\{\Delta B_k^{n+1}\}_{0\leq k\leq 2^{n+1}-1}$. The Δ_n in the modified antithetic scheme uses coupling of three discretizations from the three

Brownian increments above:

$$\Delta_n \triangleq \frac{1}{2}\big(f(X^f_{n+1}(1)) + f(X^a_{n+1}(1))\big) - f(X_n(1)). \tag{6.14}$$

where

$$\begin{aligned} X^f_{n+1}(\cdot) &\leftarrow \text{Num\_Sol}(x, n+1, \{\Delta B^{n+1}_k\}_{1\le k\le 2^{n+1}-1}) \\ X^a_{n+1}(\cdot) &\leftarrow \text{Num\_Sol}(x, n+1, \{\Delta B^{n+1,a}_k\}_{1\le k\le 2^{n+1}-1}) \\ X_n(\cdot) &\leftarrow \text{Num\_Sol}(x, n, \{\Delta B^{n}_k\}_{1\le k\le 2^{n}-1}) \end{aligned} \tag{6.15}$$

The notations X^f_{n+1} and X^a_{n+1} comes from [181]. They represent the “fine" and “antithetic" solutions on level $n+1$ versus the“coarse" solution X_n on level n. Moreover, note (6.13) is satisfied for Δ_n in (6.14). In particular, $X^f(\cdot) \overset{\mathcal{D}}{=} X^a(\cdot)$ since $\{\Delta B^n_k\}_{0\le k\le 2^n-1} \overset{\mathcal{D}}{=} \{\Delta B^{n,a}_k\}_{0\le k\le 2^n-1}$.

6.3.3 Bias removal via additional randomization

After the construction of MLMC (6.12), we note the bias exists as long as N is finite. In this section we present a bias removal technique via additonal randomness, originally proposed by [192, 191], for the construction of both Z and W in Section 6.3.1.

Definition 6 (Construction of $Z(\mu)$)**.** *Given $\theta > 0$, a fixed hyperparameter, let $N \sim Geom(1-2^{-\theta})$ be a geometric R.V. with $p_n \triangleq \mathbb{P}(N = n) = (1-2^{-\theta})(2^{-\theta n})$, $n \ge 0$. Let $n_0 \ge 0$ be the base discretization level for estimator $X_{n_0} \leftarrow$ Num_Sol$(x, n_0,$ $\{\Delta B^{n_0}_k\}_{1\le k\le 2^{n_0}-1})$ and Δ_n as defined in* (6.14)*. Then*

$$Z(\mu) \triangleq f(X_{n_0}(1)) + \frac{\Delta_{N+n_0}}{p_N}. \tag{6.16}$$

In practice, a larger value of n_0 gives lower variance of Z at the cost of a higher computational cost. We can use the same Brownian path to for $X_{n_0}(1)$ and Δ_{N+n_0} in (6.16). We summarize the procedure for obtaining $Z(\mu)$ into an Algorithm: Unbiased_Z$(\cdot,\cdot)$.

Algorithm 7 Generate $Z(\mu)$ (hyperparameters θ and γ fixed)

procedure UNBIASED_Z(x, n_0) with input $x \in \mathbb{R}^d$ and $n_0 \geq 0$.

Generate $N \leftarrow Geom(1-2^{-\theta})$, and $\boldsymbol{V}_i \leftarrow \mathcal{N}(0, \Sigma_i)$ for $1 \leq i \leq \lfloor 2^{(N+n_0+1)\gamma} \rfloor$.

$\mu^{(N+n_0+1)} \leftarrow \sum_{i=1}^{\lfloor 2^{(N+n_0+1)\gamma} \rfloor} \frac{\lambda_i}{i^q} V_i \phi_i(\cdot)$ and same for $\mu^{(N+n_0)}, \mu^{(n_0)}$

sample a Brownian path at times $\{t_k^{N+n_0+1}\}_{0 \leq k \leq 2^{N+n_0+1}}$

store Brownian increments $\{\Delta B_k^{N+n_0+1}\}_{0 \leq k \leq 2^{N+n_0+1}-1}$, $\{\Delta B_k^{n_0}\}_{0 \leq k \leq 2^{n_0}-1}$

store $\{\Delta B_k^{N+n_0+1,a}\}_{0 \leq k \leq 2^{N+n_0+1}-1}$ and $\{\Delta B_k^{N+n_0}\}_{0 \leq k \leq 2^{N+n_0}-1}$

$X_{n_0}(\cdot) \leftarrow Num\_Sol(x, n_0, \{\Delta B_k^{n_0}\}_{1 \leq k \leq 2^{n_0}-1})$ and $p_N \leftarrow (1-2^{-\theta})(2^{-\theta N})$

Compute Δ_{N+n_0} from (6.15) and (6.14)

Output $Z(\mu) \leftarrow \frac{\Delta_{N+n_0}}{p_N} + f(X_{n_0}(1))$

The additional randomness via geometric R.V. is also used for debiasing $W(\mu)$. We summarize the Algorithm: Unbiased_W for generateing $W(\mu)$, for a general $k \in \mathbb{Z}^+$ and $G(\cdot) : \mathbb{R}^k \to \mathbb{R}$.

Definition 7. *Given* $M \in \mathbb{Z}^+$ *and* $\{Z_{ij}\}_{i \in [k], j \in [M]}$. *For* $a, b \in \mathbb{Z}^+$ *and* $a \leq b \leq M$,

$$S(a, b; \{Z_{ij}\}) \triangleq G\left(\frac{1}{b-a+1}\sum_{j=a}^{b} Z_{1j}, ..., \frac{1}{b-a+1}\sum_{j=a}^{b} Z_{kj}\right) \tag{6.17}$$

Definition 8 (Construction of $W(\mu)$)**.** *Let* $\{Z_i(\mu)\}_i$ *be I.I.D. copies of random variables* $Z(\mu)$ *in* (6.16). *Define*

$$\begin{aligned}\widetilde{\Delta}_n \triangleq & S(1, 2^{\widetilde{N}+n_1+1}; \{Z_{ij}\}) \\ & - \frac{1}{2}\Big(S(1, 2^{\widetilde{N}+n_1}; \{Z_{ij}\}) + S(2^{\widetilde{N}+n_1}+1, 2^{\widetilde{N}+n_1+1}; \{Z_{ij}\})\Big).\end{aligned} \tag{6.18}$$

Let $n_1 \geq 1$ *be the base level and* $\widetilde{N} \sim Geom(1-2^{-1.5})$ *with* $\widetilde{p}_n \triangleq \mathbb{P}(\widetilde{N} = n) = 2^{-1.5n}(1-2^{-1.5})$ *for* $n \geq 0$. *Then,*

$$W(\mu) = \frac{\widetilde{\Delta}_{\widetilde{N}+n_1}}{\widetilde{p}_{\widetilde{N}}} + S(1, 2^{n_1}; \{Z_{ij}\}). \tag{6.19}$$

Algorithm 8 Generate $W(\mu)$.

procedure UNBIASED_W($\{x_i\}_{i\in[k]}, n_0, n_1$)

input: starting points $\{x_i\}_{i\in[k]} \subseteq \mathbb{R}^d$, base level $n_0 \geq 0$ and $n_1 \geq 1$.

Generate $\widetilde{N} \leftarrow Geom(1 - 2^{-1.5})$

for $1 \leq i \leq k$ **do**

  for $1 \leq j \leq 2^{\widetilde{N}+n_1+1}$ **do**

    Generate $Z_{ij} \leftarrow$ UNBIASED_Z(x_i, n_0)

compute $\widetilde{\Delta}_{\widetilde{N}+n_1}$ in (6.18) and $S(1, 2^{n_1}; \{Z_{ij}\})$

$p_{\widetilde{N}} \leftarrow 2^{-1.5\widetilde{N}}(1 - 2^{-1.5})$

Output $W(\mu) \leftarrow \frac{\widetilde{\Delta}_{\widetilde{N}+n_1}}{p_{\widetilde{N}}} + S(1, 2^{n_1}; \{Z_{ij}\})$

Notice that if we denote N_{ij} to be the geometric random variable generated for $Z_{ij} \rightarrow$ Unbiased_Z(x_i, n_0). Then, let

$$m = \max_{i\in[k], j\in[2^{\widetilde{N}+n_1+1}]} N_{ij} \quad \text{and} \quad M = \lfloor 2^{(m+n_0+1)\gamma} \rfloor.$$

We only need to generate $\boldsymbol{V}_1, ..., \boldsymbol{V}_M$ for approximating random field $\boldsymbol{\mu}$ and use them for generating all $\{Z_{ij}\}$.

6.4 Main results

In this section, we present the analysis on the moments and complexity for estimator W and Z. We show our estimator for ν is unbiased, has a finite variance and can be generated with finite computational cost.

6.4.1 Unbiasedness

To show the unbiasedness of Z, we need a couple of a technical lemma on the approximation error from the discretization scheme $X_n(\cdot) \leftarrow$ Num_ Sol.

Lemma 14. *Given* $\mu \cup \{\mu^{(n)}\}_{n\geq 1} \subseteq \mathcal{L}(L_1)$ *for* $L_1 > 1$, *let* $\{X_t\}_{t\in[0,1]}$ *be the solution of the SDE in* (6.10) *and* $\{X_n(t_k^n)\}_{0\leq k\leq 2^n}$ *be the numerical solution from Num_ Sol. Then, for appropriate choice*

of γ and θ, there exists $\epsilon > 0$ and $C > 1$ such that,

$$\mathbb{E}\|X_n(t) - X_t\|_\infty^4 \leq e^{CL_1}\Delta t_n^{2-\epsilon}, \tag{6.20}$$

for all $t \in \{t_k^n\}_{0\leq k\leq 2^n}$.

Given $\mu \in \mathcal{L}(L_1)$, typical results from stochastic analysis (e.g.,[181]) show that $\mathbb{E}\|X_n(t) - X_t\|_\infty^4 = O(\Delta t_n^2)$. Such is a standard error bound for numerical SDEs obtained from Gronwall's inequality [193, 185] which has the form

$$\mathbb{E}\|X_n(t) - X_t\|_\infty^4 \leq e^{CL_1^4}\Delta t_n^2, \tag{6.21}$$

However, when $\boldsymbol{\mu}$ is random and $e^{\boldsymbol{L}_1^p}$ may not have finite expectation for p greater than 1, which becomes a technical challenge for showing finite variance. Instead of Gronwall's inequality, we use rough path techniques in [200] to develop an path-wise bound and trade the term $e^{CL_1^p}$ for e^{CL_1} by giving up ϵ order from Δt_n^2 in (6.21).

Corollary 6.4.0.1. *Under the same setting of Lemma 14 plus Assumption 6, we have*

$$\lim_{n\to\infty} \mathbb{E}f(X_n(1)) = \mathbb{E}f(X_1). \tag{6.22}$$

Proof. From Assumption 6 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}(f(X_n(1)) - f(X_1))^2 \leq L^2\mathbb{E}\|X_n(1) - X_1\|_\infty^2 \leq L^2\sqrt{\mathbb{E}\|X_n(1) - X_1\|_\infty^4}. \tag{6.23}$$

which converge to 0 by Lemma 14. □

Lemma 15. *Under the same setting of Lemma 14 plus Assumption 6, we have*

$$\mathbb{E}Z = u(x, 1).$$

Proof. Note

$$\mathbb{E}[\frac{\Delta_{N+n_0}}{p_N}] = \mathbb{E}[\mathbb{E}[\frac{\Delta_{N+n_0}}{p_N}|N]] = \sum_{n=0}^{\infty} \frac{\mathbb{E}\Delta_{n+n_0}}{p_n} \cdot p_n = \sum_{n=0}^{\infty} \mathbb{E}\Delta_{n+n_0}$$

It then follows from (6.13) and Lemma 14 that

$$\mathbb{E}Z = \mathbb{E}fX_{n_0}(1) + \sum_{n=0}^{\infty} \mathbb{E}fX_{n+n_0+1}(1) - \mathbb{E}fX_{n+n_0}(1) = \mathbb{E}fX_1.$$

The conclusion now follows from Proposition 7. □

Lemma 15 proves the unbiasedness of Z. We also need technical lemmas on the fourth moment of Δ_n and Z. Note finite fourth moment implies finite variance.

Lemma 16. *Under the same setting of Lemma 14, there exist $\delta > 0$ and $C > 1$ that*

$$\mathbb{E}\Delta_n^4 \leq e^{CL_1}\Delta t_n^{4-\delta}, \tag{6.24}$$

$$\mathbb{E}f(X_{n_0}(1))^4 \leq poly(L_1), \tag{6.25}$$

for some polynomial function $poly(\cdot)$ satisfying $poly(x) > 1$ when $x > 1$.

Lemma 17. *Under the setting of Lemma 16, there exists appropriate choice of γ and θ with $3\theta < 4 - \delta$ such that*

$$\mathbb{E}Z^4 \leq e^{CL_1}, \tag{6.26}$$

for some $C > 1$.

Proof. It follows from

$$|\sum_{n=1}^{N} a_n|^p \leq N^{p-1} \sum_{n=1}^{N} |a_n|^p, \tag{6.27}$$

that $\mathbb{E}Z^4$ is bounded by

$$8\sum_{n=0}^{\infty} \frac{\mathbb{E}\Delta_{N+n_0}^4}{p_n^3} + 8\mathbb{E}|f(X_{n_0}(1))|^4 \leq 8\Big(\frac{e^{CL_1}}{(1-2^{-\theta})^3} \sum_{n=0}^{\infty} \frac{\Delta t_n^{4-\delta}}{\Delta t_n^{3\theta}} + poly(L_1)\Big) \tag{6.28}$$

according to Lemma 16. Since $4-\delta > 3\theta$, the conclusion follows if we can find some $C' > 1$ such that (6.28) can be bounded by $e^{C'L_1}$. This can be done since for any $poly(\cdot)$, we can find $c > 0$ such that $poly(x) < e^{cx}$ when $x > 1$. □

We can now show the unbiasedness of W.

Lemma 18. *Under the same setting of Lemma 16 plus Assumptions 6, we have*

$$\mathbb{E}W = G(\mathbb{E}Z). \tag{6.29}$$

Proof. It follows from 17 that

$$\lim_{n\to\infty} \mathbb{E}\left|\frac{\sum_{j=1}^{n} Z_j}{n} - \mathbb{E}Z\right| = 0. \tag{6.30}$$

It then follows from the bound of $\|D_x G\|_\infty$ in Assumption 6 that

$$\lim_{n\to\infty} \mathbb{E}G\Big(\frac{\sum_{j=1}^{2^n} Z_j}{2^n}\Big) = G(\mathbb{E}Z) \tag{6.31}$$

as $n \to \infty$. Now, since $\mathbb{E}\widetilde{\Delta}_n = \mathbb{E}G(\frac{\sum_{j=1}^{2^{n+1}} Z_j(\mu)}{2^{n+1}}) - \mathbb{E}G(\frac{\sum_{j=1}^{2^n} Z_j(\mu)}{2^n})$, the rest of the proof follows as in Lemma 15. □

6.4.2 Variance and computational cost

After unbiasedness, we now show W has finite variance and finite computational cost. We start with several technical lemmas and then proceed to the main theorem.

Lemma 19. *Under the setting of Lemma 14 plus Assumptions 6, $\widetilde{\Delta}_n$ satisfies*

$$\mathbb{E}\widetilde{\Delta}_n^2 \le e^{CL_1}\Delta t_n^2 \tag{6.32}$$

for some $C > 1$.

Lemma 20. *Under the setting of Lemma 14 plus Assumptions 6, we have*

$$\mathbb{E}W^2 \leq e^{CL_1} \tag{6.33}$$

for some $C > 1$.

To discuss the computational cost for generating Z, denoted as $cost_Z$, we denote the cost for generating then $X_n(\cdot) \leftarrow$ Num_Sol by $cost_n$. Then, notice

$$cost_Z = cost_{n_0} + cost_{N+n_0} + 2cost_{N+n_0+1}, \tag{6.34}$$

due to the computation of $X_{n_0}(\cdot)$, $X_{N+n_0}(\cdot)$, $X^f_{N+n_0+1}(\cdot)$ and $X^a_{N+n_0+1}(\cdot)$.

Lemma 21. *There exists appropriate choice of* γ *and* θ *with* $\theta > 1 + \gamma$ *and the computational cost for generating* Z *has finite expectation:*

$$\mathbb{E}(cost_Z) < \infty. \tag{6.35}$$

Proof. Consider the $cost_n$. For fixed n, one needs to generate 2^n Brownian increments and $O(2^{\gamma n})$ of $\boldsymbol{V}_i$ for $\boldsymbol{\mu}^{(n)}$. Then, to compute $X_n(1)$, one needs 2^n recursions in Num_Sol and each iteration requires $O(2^{\gamma n})$ computation to evaluate

$$\phi_1(X_n(t_k^n)), ..., \phi_{2^{\lfloor \gamma n \rfloor}}(X_n(t_k^n))$$

in $\boldsymbol{\mu}^{(n)}(X_n(t_k^n))$. Thus,

$$cost_n \sim O(2^{(1+\gamma)n}) \tag{6.36}$$

Therefore, from (6.34) and $p_n \sim 2^{-\theta n}$, we have

$$\begin{aligned}\mathbb{E}(cost_Z) \leq & \mathbb{E}(cost_{n_0}) + \mathbb{E}(cost_{N+n_0}) + 2\mathbb{E}(cost_{N+n_0+1}) \\ = & \mathbb{E}(cost_{n_0}) + \sum_{n=0}^{\infty} \mathbb{E}(cost_{n+n_0}) p_n + 2\sum_{n=0}^{\infty} \mathbb{E}(cost_{n+n_0+1}) p_n \\ \sim & O\left(2^{(1+\gamma)n_0}(1 + \sum_{n=0}^{\infty} 2^{(1+\gamma-\theta)n} + 2^{1+\gamma}\sum_{n=0}^{\infty} 2^{(1+\gamma-\theta)n})\right) < \infty \end{aligned} \tag{6.37}$$

since $\theta > 1 + \gamma$. □

Now we discuss the computational cost for generating W, denoted by $cost_W$. The construction W consists of $\widetilde{\Delta}_{\widetilde{N}+n_1}$ and $G(\sum_{j=1}^{2^{n_1}} Z_j/2^{n_1})$. The computation cost of both lies in generating samples of Z:

$$cost_W = \sum_{j=1}^{2^{\widetilde{N}+n_1+1}} cost_{Z_j}. \tag{6.38}$$

Lemma 22. *The total expected computational cost of W satisfies*

$$\mathbb{E}(cost_W) < \infty. \tag{6.39}$$

Proof. Using Wald's identity and Lemma 21, we have

$$\begin{aligned}\mathbb{E}(cost_W) = & \mathbb{E}(2^{\tilde{N}+n_1+1})\mathbb{E}(cost_Z) \\ = & 2^{n_1+1}\left(\sum_{n=0}^{\infty} 2^{-0.5n}(1 - 2^{-1.5})\right)\mathbb{E}(cost_Z) < \infty. \end{aligned} \tag{6.40}$$

□

6.4.3 Main theorem

Theorem 6.4.1. *Under Assumptions 5-6 and appropriate choice of γ and θ, W is an unbiased estimator for ν. Moreover, W has a finite variance and the computational cost for generating W has finite expectation.*

Proof. The appropriate choice for γ and θ that satisfies all Lemma conditions are referred to Definition 9. The unbiasedness follows from Lemma 18. The finite expected computational cost follows from Lemma 22.Since $\boldsymbol{\mu}(\cdot) \in \mathcal{L}(\boldsymbol{L}_1)$ almost surely for the $\boldsymbol{L}_1$ in Lemma 13, for any realization, we have $\mu(\cdot, \omega) \in \mathcal{L}(L_1(\omega))$ with probability 1. To show the finite variance property of W, note it follows from Lemma 13 and Lemma 20 that,

$$\mathbb{E}W^2 = \mathbb{E}[\mathbb{E}[W^2(\mu)|\boldsymbol{\mu} = \mu]] \leq \mathbb{E}[\mathbb{E}[e^{CL_1}|\boldsymbol{\mu} = \mu]] \leq \mathbb{E}[e^{C\boldsymbol{L}_1}] < \infty.$$

□

6.5 Simulation

Example 1 Consider the one-dimensional SDE known as the Ornstein-Uhlenbeck Process [185]:

$$\begin{cases} dX_t = & -\boldsymbol{\alpha} X_t dt + dB_t \qquad \text{for } t \geq 0 \\ X_0 = & 0 \end{cases}, \tag{6.41}$$

where $\boldsymbol{\alpha} \in \mathbb{R}$ is random. Given the realization $\alpha(\omega)$, the solution are known exactly,

$$X_1 = e^{-\alpha t} \int_0^1 e^{\alpha s}\, dB_s. \tag{6.42}$$

Consequently, given the realization $\alpha(\omega)$, using Itô's isometry, it can be shown that X_1 is Gaussian with mean 0 and variance $(2\alpha(\omega))^{-1}(1 - e^{-2\alpha(\omega)})$. For simulation, we set $\boldsymbol{\alpha}$ to be Gaussian with mean 1 and variance 0.05^2 along with $f(x) = x^2$, $G(x) = e^{-x^2}$. Then, it follows from direct calculation that $\mathbb{E}[f(X_1)|\alpha] = \frac{1-e^{-2\alpha}}{2\alpha}$, and

$$\mathbb{E}[G(\mathbb{E}[f(X_1)|\alpha])] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \cdot 0.05^2}} \cdot\ e^{-\frac{(x-1)^2}{2\cdot 0.05^2}} \cdot\ e^{-(\frac{(1-e^{-2x})}{2x})^2} dx \approx 0.8291.$$

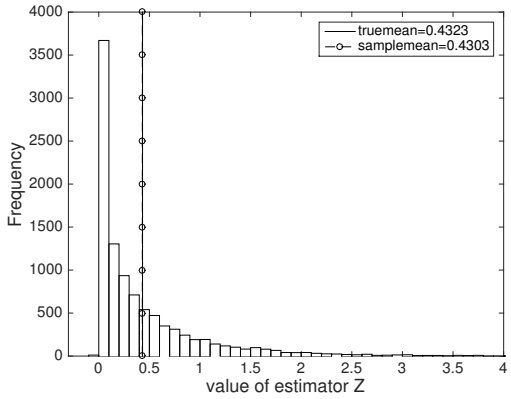
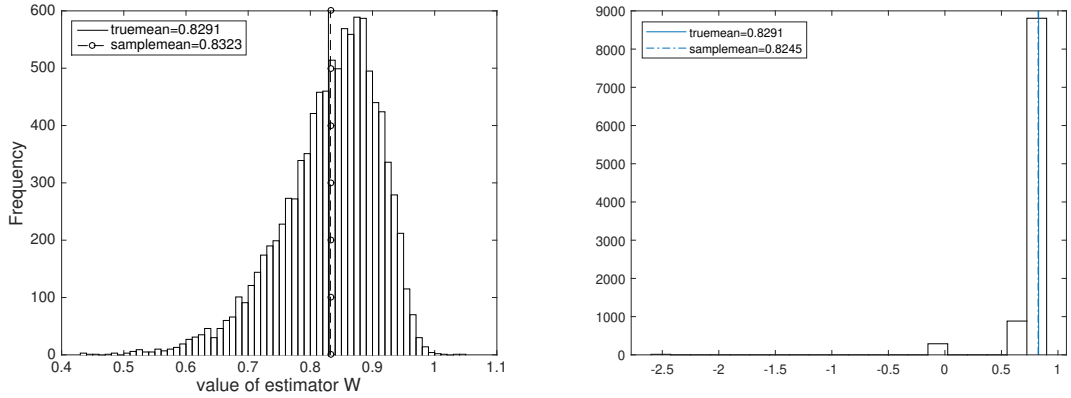


Figure 6.1: Histogram of Estimator Z when $\alpha = 1$



(a) Histogram of Estimator W

(b) Estimators based on Numerical PDE

Figure 6.2: Comparsion of Multilevel Estimators based on Antithetic Numerical SDE or Numerical PDE

To check the unbiasedness property of Z, we first fix $\alpha = 1$ in simulation so that $\mathbb{E}[f(X_1)|\alpha = 1] \approx 0.4323$. Picking $n_0 = 5$ as the base level, we generate 10,000 copies of Z with $\alpha = 1$. A sample mean of 0.4303 is obtained to compare with its true mean 0.4323, as in Figure 1. Then, we pick $n_1 = 5$ and generate 10,000 copies of W to obtain a sample mean of 0.8323 while the true mean is 0.8291, as in Figure 2a. Furthermore, in Figure 2b, we generate 10000 copies of unbiased estimators of $G(u(x, 1))$ using the multilevel Monte Carlo estimator based on a finite difference numerical PDE solver similarly as the methods proposed in [187]. In both cases, the sample size is 10,000 and the difference between sample mean and true mean is within a 95% confidence interval. Overall, the findings are consistent with our theoretical results on the unbiasedness.

Example 2 In this example, we consider the more complicated SDE:

$$\begin{cases} dX_t &= -\boldsymbol{\mu}(X_t)dt + \cos(X_t)dB_t \qquad \text{for } t \geq 0 \\ X_0 &= 0, \end{cases} \tag{6.43}$$

where $\boldsymbol{\mu}(x) = \sum_{i=1}^{\infty} i^{-4} \sin(ix)\boldsymbol{V}_i$ and we compare the proposed method with the standard Monte Carlo method with bias. We take $\gamma = \frac{1}{3}$ and $\theta = \frac{4}{3}$ for simplicity. Similar to the previous example, we take $n_0 = n_1 = 5$. We generate $10,000$ copies of our estimator and compare it with $10,000$ copies of a standard Monte Carlo estimator where we remove the debiasing part $\frac{\Delta_N}{p_N}$ in *both* estimator Z and W. As a result, using the CLT, we compute a 95% confidence interval $[0.4610, 0.4656]$ for our estimator while we obtain an interval $[0.5189, 0.5255]$ for the standard Monte Carlo estimator. As we can see, these two intervals are not overlapping, suggesting that the standard Monte Carlo estimator has a non-negligible bias.

6.6 Supplementary: Proofs

6.6.1 Proof of Lemma 19

Proof. Denote $S(a,b) \triangleq \frac{\sum_{j=a}^{b} Z_j}{b-a+1}$ as before but also $S^k(a,b) = (S(a,b) - \mathbb{E}Z)^k$. Then, as in [192], a second order Taylor expansion of $G(\cdot)$ around $\mathbb{E}Z(\mu)$ gives

$$\begin{aligned} \widetilde{\Delta}_n =& G(S(1,2^{n+1})) - \frac{1}{2}\Big(G(S(1,2^n)) + G(S(2^n+1,2^{n+1}))\Big) \\ =& G^{'}(\mathbb{E}Z(\mu))(S(1,2^{n+1}) - \frac{1}{2}\Big(S(1,2^n) + S(2^n+1,2^{n+1})\Big) \\ &+ \frac{G^{''}(\xi_1)}{2}S^2(1,2^{n+1}) - \frac{G^{''}(\xi_2)}{4}S^2(1,2^n) - \frac{G^{''}(\xi_3)}{4}S^2(2^n+1,2^{n+1}), \end{aligned} \tag{6.44}$$

where ξ_1 is between $\mathbb{E}Z(\mu)$ and $S(1, 2^{n+1})$, similarly ξ_2 between $\mathbb{E}Z(\mu), S(1, 2^n)$ and ξ_3 between $\mathbb{E}Z(\mu), S(2^n + 1, 2^{n+1})$. Thus, it follows from (6.27) and Assumption 6 that

$$|\widetilde{\Delta}_n|^2 \leq \frac{3L^2}{4}\big(S^4(1, 2^{n+1}) + \frac{1}{4}S^4(1, 2^n) + \frac{1}{4}S^4(2^n + 1, 2^{n+1})\big). \tag{6.45}$$

However, $(Z_j(\mu) - \mathbb{E}Z(\mu))$ are I.I.D. with mean 0. In particular, when we write out the expansion in (6.45) and take expectation, the terms with odd power will vanish

$$\begin{aligned}
\mathbb{E}[(Z_i(\mu) - \mathbb{E}Z(\mu))^2(Z_j(\mu) - \mathbb{E}Z(\mu))(Z_k(\mu) - \mathbb{E}Z(\mu))] &= 0, \\
\mathbb{E}[(Z_i(\mu) - \mathbb{E}Z(\mu))^3(Z_j(\mu) - \mathbb{E}Z(\mu))] &= 0, \\
\mathbb{E}[(Z_i(\mu) - \mathbb{E}Z(\mu))(Z_j(\mu) - \mathbb{E}Z(\mu))(Z_k(\mu) - \mathbb{E}Z(\mu))(Z_l(\mu) - \mathbb{E}Z(\mu))] &= 0.
\end{aligned} \tag{6.46}$$

Thus, taking expectation in (6.45) gives $\mathbb{E}\widetilde{\Delta}_n^2$ is bounded by

$$C\binom{2^{n+1}}{2}2^{-4n} \cdot \mathbb{E}\big(Z(\mu) - \mathbb{E}Z(\mu)\big)^4 \tag{6.47}$$

for some $C > 1$ since $\mathbb{E}(Z_j(\mu) - \mathbb{E}Z(\mu))^2(Z_i(\mu) - \mathbb{E}Z(\mu))^2 \leq \mathbb{E}(Z_j(\mu) - \mathbb{E}Z(\mu))^4$. Since $\binom{n}{2} = O(n^2)$, we have

$$\binom{2^{n+1}}{2}2^{-4n} \leq C\Delta t_n^2, \tag{6.48}$$

for some (different) $C > 1$. Finally, we bound $\mathbb{E}\big(Z(\mu) - \mathbb{E}Z(\mu)\big)^4$ by Lemma 17:

$$\mathbb{E}\big(Z(\mu) - \mathbb{E}Z(\mu)\big)^4 \leq e^{CL_1} \tag{6.49}$$

for some $C > 1$. Thus, we conclude there exists some $C > 1$ that $\mathbb{E}\widetilde{\Delta}_n^2 \leq e^{CL_1}\Delta t_n^2$. □

6.6.2 Proof of Lemma 20

Proof. Denote $S(a,b) \triangleq \frac{\sum_{j=a}^{b} Z_j}{b-a+1}$ as before but also $S^k(a,b) = (S(a,b) - \mathbb{E}Z)^k$. By Lemma 17, Assumption 6 on $G(\cdot)$ and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}\Big|G(S(1,2^{n_1}))\Big|^2 \leq& \mathbb{E}(|G(0)| + L|S(1,2^{n_1})|)^2 \\
\leq& |G(0)|^2 + 2|G(0)|L|S(1,2^{n_1})| + L^2 S^2(1,2^{n_1}) \\
\leq& C + Ce^{CL_1}
\end{aligned} \tag{6.50}$$

for some $C > 1$. Now, using (6.27), (6.50) and Lemma 19, we have

$$\begin{aligned}
\mathbb{E}W^2 \leq& 2\mathbb{E}\Big(\frac{\widetilde{\Delta}^2_{N+n_1}}{\widetilde{p}^2_N} + \Big|G(\frac{\sum_{j=1}^{2^{n_1}} Z_j}{2^{n_1}})\Big|^2\Big) \\
=& 2\sum_{n=0}^{\infty} \frac{\mathbb{E}\widetilde{\Delta}^2_{n_1+n}}{\widetilde{p}_n} + 2\mathbb{E}\Big|G(\frac{\sum_{j=1}^{2^{n_1}} Z_j}{2^{n_1}})\Big|^2 \\
\leq& \frac{2e^{CL_1}}{(1-2^{-1.5})} \sum_{n=0}^{\infty} \frac{2^{-2n}}{2^{-1.5n}} + 2C + 2Ce^{CL_1} \leq e^{C'L_1}
\end{aligned} \tag{6.51}$$

for some $C' > 1$. The last inequality follows since for any a, b and c, there exists d such that $a + ce^{bx} < e^{dx}$ when $x > 1$.

□

6.6.3 Definitions and supporting lemmas

The following definition discusses the appropriate hyperparameters as well as choice of ϵ, δ in Lemma 14 and Lemma 16. Finally, α and β are used for rough path estimates in the sequel.

Definition 9. *Let $\epsilon > 0$ to be small enough so*

$$\epsilon < \frac{1}{144} \quad \textit{and} \quad \epsilon < \frac{q-4}{36}\frac{1}{2+q}, \tag{6.52}$$

where $q > 4$ is from Assumption 5. Define

$$\alpha \triangleq \frac{1}{2} - \epsilon, \;\; \beta \triangleq \frac{1}{2} + 2\epsilon, \;\; \gamma \triangleq \frac{1}{3} - 12\epsilon, \quad \theta \triangleq \frac{4}{3} - \frac{23}{2}\epsilon \;\; \textit{and} \;\; \delta \triangleq 33\epsilon \tag{6.53}$$

It is easy to check:

$$\gamma \geq \frac{1}{4}, \; (3 + \frac{q-4}{2})\gamma > 1, \; 8(2\alpha - \beta) > 4 - \delta > 3\theta > 0 \;\; \textit{and} \;\; \theta > 1 + \gamma > 0 \tag{6.54}$$

The next definition is used for rough path estimates as well. Notice we have extend the definition of $\widetilde{A}$ in Definition 3 to include generat $(s,t) \in [0,1] \times [0,1]$.

Definition 10. *Let $\{B(t)\}_{t\in[0,1]}$ be a Brownian motion on $[0,1]$ and α, β be defined as in Definition 9. Then, define*

$$\|B\|_\alpha \triangleq \sup_{0\leq s\leq t\leq 1} \frac{\|B(t) - B(s)\|_\infty}{|t-s|^\alpha} \textit{ and } \|A\|_{2\alpha} \triangleq \sup_{0\leq s\leq t\leq 1} \max_{1\leq i,j\leq d'} \frac{|A_{ij}(s,t)|}{|t-s|^{2\alpha}}$$

$$\|\tilde{A}\|_{2\alpha} \triangleq \sup_{0\leq s\leq t\leq 1} \max_{1\leq i,j\leq d'} \frac{|\widetilde{A}_{ij}(s,t)|}{|t-s|^{2\alpha}} \textit{ and } \Gamma_{\widetilde{R}} \triangleq \sup_{\substack{0\leq s\leq t\leq 1 \\ s,t\in D_n, n\geq 1}} \max_{1\leq i,j\leq d'} \frac{|\widetilde{R}^n_{i,j}(s,t)|}{|t-s|^\beta \Delta t_n^{2\alpha-\beta}},$$

where D_n is the dyadic rationals(i.e. multiples of $\frac{1}{2^n}$) in $[0,1]$ and for $i, j \in [d'], i \neq j$,

$$\begin{aligned}
A_{ij}(s,t) &\triangleq \int_s^t (B_i(u) - B_i(s)) dB_j(u) \\
\widetilde{A}_{i,j}(s,t) &\triangleq \frac{(B_i(t) - B_i(s))(B_j(t) - B_j(s))}{2} \\
\widetilde{A}_{i,i}(s,t) = A_{i,i}(s,t) &\triangleq \frac{(B_i(t) - B_i(s))^2 - (t-s)}{2}, \\
\widetilde{R}^n_{i,j}(t^n_l, t^n_m) &\triangleq \sum_{k=l+1}^{m} \{A_{i,j}(t^n_{k-1}, t^n_k) - \widetilde{A}_{i,j}(t^n_{k-1}, t^n_k)\}.
\end{aligned}$$

The proofs for the following lemmas are left in the supplementary material.

Lemma 23. *Fixing $\epsilon > 0$, let $\{\mathbf{Z}_n\}_{n\geq 1}$ be a sequence I.I.D. standard d-dimensional Gaussian*

random vectors (i.e., $\Sigma_n = I_d$ for all $n \geq 1$). Then, the R.V.

$$M_\epsilon \triangleq \sup_{n\geq 1} \frac{\|\boldsymbol{Z}_n\|_\infty}{n^\epsilon},$$

has finite moment-generating function (i.e., $\mathbb{E}[e^{tM_\epsilon}] < \infty$) for all $t \geq 0$.

Lemma 24. *The quantities $\|B\|_\alpha$, $\|A\|_{2\alpha}$, $\|\tilde{A}\|_{2\alpha}$ and $\Gamma_{\widetilde{R}}$ defined in Definition 10 have moments of arbitrary order.*

Lemma 25. *Let $X_n(\cdot)$ be the discretization in Num_Sol generated under $\mu^{(n)}(\cdot) \in \mathcal{L}(L_1), L_1 > 1$ and Brownian motion $B(\cdot)$. Then, there exists $poly(\cdot) : \mathbb{R}^3 \to \mathbb{R}$ such that*

$$\|X_n(t) - X_n(r)\|_\infty \leq poly(L_1, \|B\|_\alpha, \|\widetilde{A}\|_{2\alpha})|t-r|^\alpha$$

for $r, t \in \{t_k^n\}_{0\leq k\leq 2^n}$ and $n \geq 0$. Moreover, $poly(x, \cdot, \cdot) > 1$ for $x > 1$.

Lemma 26. *Let $X_n^\mu(\cdot)$ be the discretization from Num_Sol but generated under $\mu(\cdot) \in \mathcal{L}(L_1), L_1 > 1$ (instead of $\mu^n(\cdot)$, same as in [181]) and Brownian motion $B(\cdot)$. Also, let $\{X_t\}_{t\in[0,1]}$ be the solution of SDE in* (6.10)*. Then, there exists $poly(\cdot) : \mathbb{R}^4 \to \mathbb{R}$ that,*

$$\|X_n^\mu(t) - X_t\|_\infty \leq poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha}, \Gamma_{\widetilde{R}})\Delta t_n^{2\alpha-\beta}. \tag{6.55}$$

for $n \geq 0$ and $t \in \{t_k^n\}_{0\leq k\leq 2^n}$. Moreover, $poly(x, \cdot, \cdot, \cdot) > 1$ for $x > 1$.

6.6.4 Proof of Lemma 13

Proof of Lemma 13. Let $\{\boldsymbol{V}_n\}_{n\geq 1}$ be I.I.D. $\mathcal{N}(\boldsymbol{0}, \Sigma_n)$ with the covariance matrix $\|\Sigma_n\|_F < L$ for all $n \geq 1$ as in Assumption 5. Since $\{\Sigma_n\}_n$ are positive semi-definite, the square root matrices $\Sigma_n^{\frac{1}{2}}$

satisfy,

$$\begin{aligned}\|\Sigma_n^{\frac{1}{2}}\|_F^2 =& trace((\Sigma_n^{\frac{1}{2}})^T(\Sigma_n^{\frac{1}{2}})) = trace(\Sigma_n)\\ =& \sum_i \lambda_i \leq \sum_i (\lambda_i^2 + 1) \leq trace(\Sigma_n^T\Sigma_n) + d\\ \leq& \|\Sigma_n\|_F^2 + d \leq L^2 + d,\end{aligned}$$

where λ_i are the eigenvalues of Σ_n. Thus, if we set $L' = \sqrt{L^2 + d} > 1$, we have $\|\Sigma_n^{\frac{1}{2}}\|_F < L'$ for all $n \geq 1$. Finally, by the equivalence of matrix norms, there exists $L'' > 1$ such that $\|\Sigma_n^{\frac{1}{2}}\|_\infty < L''$, for all $n \geq 1$.

Consequently, if we let $\{\boldsymbol{Z}_n\}_{n\geq 1}$ be a sequence of I.I.D. d-dimensional standard Gaussian, then $\{\Sigma_n^{\frac{1}{2}} \cdot \boldsymbol{Z}_n\}_{n\geq 1}$ are distributed as $\{\boldsymbol{V}_n\}_{n\geq 1}$, and we define

$$M_{\frac{q-4}{2}} \triangleq \sup_{n\geq 1} \frac{\|\Sigma_n^{\frac{1}{2}} \cdot \boldsymbol{Z}_n\|_\infty}{n^{\frac{q-4}{2}}} \leq L'' \sup_{n\geq 1} \frac{\|\boldsymbol{Z}_n\|_\infty}{n^{\frac{q-4}{2}}}. \tag{6.56}$$

It then follows from Lemma 23 that, the random variable $M_{\frac{q-4}{2}}$ and thus

$$N_{\frac{q-4}{2}} \triangleq \sup_{n\geq 1} \frac{\|\boldsymbol{V}_n\|_\infty}{n^{\frac{q-4}{2}}}$$

has finite moment-generating function for all $t \geq 0$. Finally, to bound $\|D_x\boldsymbol{\mu}\|_\infty$,

$$\begin{aligned}\|D_x\mu\|_\infty \leq& \sum_{n=1}^\infty \frac{|\lambda_n|\|V_n\|_\infty}{n^{4+\frac{q-4}{2}} n^{\frac{q-4}{2}}} \|D_x\psi_n\|_\infty\\ \leq& \sum_{n=1}^\infty \frac{|\lambda_n| nL}{n^{4+\frac{q-4}{2}}} N_{\frac{q-4}{2}} \leq C N_{\frac{q-4}{2}},\end{aligned}$$

for some $C > 1$, by Assumptions 5-6. Similarly, we can bound $\|\boldsymbol{\mu}\|_\infty$ and $\|D_{xx}\boldsymbol{\mu}\|_\infty$ by R.V. with finite moment-generating function. The same bound applies for $\boldsymbol{S}_n$, $\boldsymbol{\mu}^{(n)}$ and $\bar{\boldsymbol{\mu}}^{(n)}$ and we can this uniform (random) bound by $\boldsymbol{L}_1$ with condition $\boldsymbol{L}_1 > 1$. □

6.6.5 Proof of Lemma 14

Proof of Lemma 14. Let $\{X_t\}_{t\in[0,1]}$, $X_n(\cdot) \leftarrow$ Num_Sol and $X_n^\mu(\cdot)$ be as defined in Lemma 25 and 26. Then, for $t \in [0,1]$,

$$\|X_n(t) - X_t\|_\infty \le \|X_n(t) - X_n^\mu(t)\|_\infty + \|X_n^\mu(t) - X_t\|_\infty. \tag{6.57}$$

To bound $\|X_n^\mu(t) - X_t\|_\infty$, Lemma 26 provides some $poly(x,\cdot,\cdot,\cdot) > 1$ for $x > 1$ that

$$\|X_n^\mu(t) - X_t\|_\infty^4 \le poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha}, \Gamma_{\widetilde{R}})\Delta t_n^{4(2\alpha-\beta)}.$$

However, Lemma 24 states $\|B\|_\alpha$,$\|A\|_{2\alpha}$ and $\Gamma_{\widetilde{R}}$ have moments of arbitrary order. Thus, for $\mu(\cdot) \in \mathcal{L}(L_1)$, $L_1 > 1$, we can find some $poly'(x) > 1$ for $x > 1$ such that

$$\begin{aligned}\mathbb{E}\|X_n^\mu(t) - X_t\|_\infty^4 \le& \mathbb{E}[poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha}, \Gamma_{\widetilde{R}})]\Delta t_n^{4(2\alpha-\beta)} \\ \le& poly'(L_1)\Delta t_n^{4(2\alpha-\beta)} \le e^{CL_1}\Delta t_n^{4(2\alpha-\beta)},\end{aligned} \tag{6.58}$$

for some appropriately chosen $C > 1$ (note this is possible since $L_1 > 1$). Combining this with (6.57), we have

$$\mathbb{E}\|X_n(t) - X_t\|_\infty^4 \le 8\mathbb{E}\|X_n(t) - X_n^\mu(t)\|_\infty^4 + 8e^{CL_1}\Delta t_n^{4(2\alpha-\beta)}. \tag{6.59}$$

On the other hand, if we can show

$$\mathbb{E}\|X_n(t) - X_n^\mu(t)\|_\infty^4 \le e^{CL_1}\Delta t_n^{4\alpha}, \tag{6.60}$$

for some $C > 1$, we can show

$$\mathbb{E}\|X_n(t) - X_n^\mu(t)\|_\infty^4 \le e^{CL_1}\Delta t_n^{4(2\alpha-\beta)}, \tag{6.61}$$

since $4\alpha = 2 - 4\epsilon > 2 - 16\epsilon = 4(2\alpha - \beta)$ by Lemma 9 and $\Delta t_n < 1$. Finally, we can conclude the proof using (6.59),(6.61) and by adjusting the constant C and ϵ (Lemma 14 states $2 - 2\epsilon$ and we have $2\alpha - \beta = 2 - 16\epsilon$ here). To prove (6.60), we define $\bar{\mu}^{(n)}(\cdot) \triangleq \mu - \mu^{(n)} = \sum_{i=\lfloor 2^{n\gamma} \rfloor + 1}^{\infty} \frac{\lambda_i}{i^q} V_i \psi_i(\cdot)$. As shown in Section 6.6.4, the proof of Lemma 13, that

$$\|\bar{\mu}^{(n)}(\cdot)\|_\infty \leq L_1 \Delta t_n^{3 + \frac{q-4}{2}}. \tag{6.62}$$

Then, for $X_n^\mu(\cdot) - X_n(\cdot)$, we notice the recursion:

$$X_{i,n}^\mu(t_{k+1}^n) - X_{i,n}(t_{k+1}^n) = X_{i,n}^\mu(t_k^n) - X_{i,n}(t_k^n) + \eta_{i,n,k}, \tag{6.63}$$

for $i \in [d]$ and $0 \leq k \leq 2^n - 1$, obtained by modifying Num_Sol with $\eta_{i,n,k}$:

$$\begin{aligned}
\eta_{i,n,k} \triangleq\ & \left(\mu_i^{(n)}(X_n^\mu(t_k^n)) - \mu_i^{(n)}(X_n(t_k^n))\right)\Delta t_n + \bar{\mu}^{(n)}(X_n^\mu(t_k^n))\Delta t_n \\
& + \sum_{j=1}^{d} \Big(\sigma_{ij}(X_n^\mu(t_k^n)) - \sigma_{ij}(X_n(t_k^n))\Big)\Delta B_{j,k}^n \\
& + \sum_{j=1}^{d'} \sum_{l=1}^{d} \sum_{m=1}^{d'} \Big(\frac{\partial \sigma_{ij}}{\partial x_l}(X_n^\mu(t_k^n))\sigma_{lm}(X_n^\mu(t_k^n)) \\
& - \frac{\partial \sigma_{ij}}{\partial x_l}(X_n(t_k^n))\sigma_{lm}(X_n(t_k^n))\Big)\widetilde{A}_{mj}(t_k^n, t_{k+1}^n).
\end{aligned} \tag{6.64}$$

Furthermore, for convenience, define for $i \in [d]$ and $0 \leq k \leq 2^n$, $\xi_{n,k} \triangleq X_n^\mu(t_k^n) - X_n(t_k^n)$ and $\xi_{i,n,k} \triangleq X_{i,n}^\mu(t_k^n) - X_{i,n}(t_k^n)$ so that (6.63) becomes, $\xi_{i,n,k+1} = \xi_{i,n,k} + \eta_{i,n,k}$. Given $\mu \cup \{\mu^{(n)}\}_n \subseteq \mathcal{L}(L_1)$, $L_1 > 1$, taking expectation in (6.64) after raising it to the fourth power,

$$\begin{aligned}
& \mathbb{E}(\xi_{i,n,k+1}^4) \\
=& \mathbb{E}(\xi_{i,n,k}^4) + \mathbb{E}(\eta_{i,n,k}^4) + 3\mathbb{E}(\xi_{i,n,k}^3 \eta_{i,n,k}) + 3\mathbb{E}(\xi_{i,n,k}\eta_{i,n,k}^3) + 6\mathbb{E}(\xi_{i,n,k}^2 \eta_{i,n,k}^2).
\end{aligned} \tag{6.65}$$

Recalling (6.60), it sufficies to show $\mathbb{E}\|X_n(t) - X_n^\mu(t)\|_\infty^4 = \mathbb{E}\|\xi_{n,2^n}\|_\infty^4 \leq e^{CL_1}\Delta t_n^{4\alpha}$ for some $C > 1$. Thus, it further sufficies to prove there exists $poly(\cdot)$ that:

- (I) When n is large that $2^n > poly(L_1)$, $\mathbb{E}|\xi_{i,n,k}|^4 \leq e^{CL_1 \cdot t_k^n} \Delta t_n^{4\alpha}$ for all i, k.
- (II) When $2^n \leq poly(L_1)$, there exists $poly'(\cdot)$ with $poly'(x) > 1$ for $x > 1$ such that for $\mathbb{E}|\xi_{i,n,k}|^4 \leq poly'(L_1)\Delta t_n^{4\alpha}$ for all i, k.

The proof of above statement would conclude (after further adjustments of constants) the proof. So we focus on proving statement (I) and (II).

Proof of statement (I) Given $\mu \cup \{\mu^{(n)}\}_n \subseteq \mathcal{L}(L_1), L_1 > 1$, we do inductions on $0 \leq k \leq 2^n$. First of all, when $k = 0$, for $i \in [d]$, the claim holds since $\xi_{i,n,0} = X_{i,n}^{\mu}(0) - X_{i,n}(0) = x - x = 0$. Next, for $0 \leq k \leq 2^n - 1$, assume that the induction hypothesis holds so that for all $0 \leq j \leq k$,

$$\mathbb{E}|\xi_{i,n,j}^4| \leq e^{CL_1 t_j^n} \cdot \Delta t_n^{4\alpha}, \tag{6.66}$$

for $i \in [d]$ and some $C > 1$. We need to show

$$\mathbb{E}|\xi_{i,n,k+1}^4| \leq e^{CL_1 t_{k+1}^n} \cdot \Delta t_n^{4\alpha}, \tag{6.67}$$

for all $i \in [d]$. To do so, we bound every term on the right hand side of (6.65). For $\eta_{i,n,k}^4$, define $\bar{d} \triangleq \max\{d, d'\}$, by Definition 10 $|\eta_{i,n,k}|$ is bounded by

$$\begin{aligned}
|\eta_{i,n,k}| \leq & \|\partial \mu_i^{(n)}\|_\infty \|\xi_{n,k}\|_\infty \Delta t_n + \|\bar{\mu}^{(n)}\|_\infty \Delta t_n \\
& + \bar{d} L \|\xi_{n,k}\|_\infty \|B\|_\alpha \Delta t_n^\alpha + \bar{d}^3 L \|\xi_{n,k}\|_\infty \|A\|_{2\alpha} \Delta t_n^{2\alpha} \\
\leq & L_1 \|\xi_{n,k}\|_\infty \Delta t_n + L_1 \Delta t_n^{4 + \frac{q-4}{2}} + \bar{d} L \|\xi_{n,k}\|_\infty \|B\|_\alpha \Delta t_n^\alpha \\
& + \bar{d}^3 L \|\xi_{n,k}\|_\infty \|A\|_{2\alpha} \Delta t_n^{2\alpha},
\end{aligned}$$

where the last inequality follows from (6.62). Since $\xi_{n,k}$ is independent of the shifted Brownian motion $\{B(t) - B(t_k^n)\}_{t_k^n \leq t \leq t_{k+1}^n}$, quantities $\|B\|_\alpha$ and $\|A\|_{2\alpha}$ associated to $\{B(t) - B(t_k^n)\}_{t_k^n \leq t \leq t_{k+1}^n}$ are thus independent of $\xi_{n,k}$. Consequently, it then follows from Lemma 24 that we can $C' > 1$ such

that

$$\begin{aligned}\mathbb{E}\eta_{i,n,k}^4 \leq & C'\big(L_1^4\mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n^4 + L_1^4\Delta t_n^{16+2(q-4)} \\ & + \mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n^{4\alpha} + \mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n^{8\alpha}\big) \\ \leq & C'' L_1^4 \bar{d}^4 e^{CL_1\cdot t_k^n}\cdot \Delta t_n^{8\alpha}, \end{aligned} \tag{6.68}$$

for some $C'' > 1$ where the last line follows from the induction hypothesis and $8\alpha < 16 + 2(q-4)$ in Definition 9. To bound $\mathbb{E}(\xi_{i,n,k}^3\eta_{i,n,k})$ in (6.65), we observe the terms in (6.64). We use (6.62) along with the martingale property (i.e., the independence of ΔB_k^n and $X_n(t_k^n)$) to obtain

$$\begin{aligned}&\mathbb{E}(\xi_{i,n,k}^3\eta_{i,n,k}) \\ =&\mathbb{E}\bigg[\big(X_{i,n}^{\mu}(t_k^n) - X_{i,n}^{\mu^{(n)}}(t_k^n)\big)^3\Big(\big(\mu_i^{(n)}(X_n^{\mu}(t_k^n)) - \mu_i^{(n)}(X_n^{\mu^{(n)}}(t_k^n))\big)\Delta t_n \\ &+ \bar{\mu}^{(n)}(X_n^{\mu}(t_k^n))\Delta t_n\Big)\bigg] \\ \leq&\mathbb{E}[L_1\|\xi_{n,k}\|_\infty^4\Delta t_n] + \mathbb{E}[L_1\|\xi_{n,k}\|_\infty^3\Delta t_n^{4+\frac{q-4}{2}}] \leq 2L_1\bar{d}^4 e^{CL_1\cdot t_k^n}\Delta t_n^{4\alpha+1}.\end{aligned}$$

The inequality follows from induction hypothesis, Hölder's inequality and the fact that $\alpha < 4 + \frac{q-4}{2}$ as in Definition 9. For the bound on $\mathbb{E}(\xi_{i,n,k}^2\eta_{i,n,k}^2)$, notice the bound on $|\eta_{i,n,k}|$ and the fact $\mathbb{E}\big(B(t)-B(s)\big)^2 = O(|t-s|)$ and $\mathbb{E}\big(\widetilde{A}_{ij}(s,t)\big)^2 = O((t-s)^2)$ (see, for example, [201]), we can find some $C' > 1$ that $\mathbb{E}(\xi_{i,n,k}^2\eta_{i,n,k}^2)$ is bounded by

$$\begin{aligned}\mathbb{E}(\xi_{i,n,k}^2\eta_{i,n,k}^2) \leq & C'\Big(\mathbb{E}(\|\xi_{n,k}\|_\infty^4)L_1^2\Delta t_n^2 + \mathbb{E}(\|\xi_{n,k}\|_\infty^2)L_1^2\Delta t_n^{4+q} \\ & + \mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n + \mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n^2\Big) \\ \leq & C'\mathbb{E}(\|\xi_{n,k}\|_\infty^4)\Delta t_n\big(L_1^2\Delta t_n + 2\big) + C'\mathbb{E}(\|\xi_{n,k}\|_\infty^4)^{\frac{1}{2}}\Delta t_n^{2\alpha+1}\big(L_1^2(\omega)\Delta t_n\big) \\ \leq & 2C''\bar{d}^4 e^{CL_1\cdot t_k^n}(L_1^2\Delta t_n + 1)\Delta t_n^{4\alpha+1}, \end{aligned} \tag{6.69}$$

for some $C'' > 1$. The last line follows from induction hypothesis. The second to last line follows from Hölder's inequality and the fact that $2\alpha + 2 < 4 + q$ as in Definition 9. Finally, to bound $\mathbb{E}(\xi_{i,n,k}\eta_{i,n,k}^3)$ in (6.65), we again use inequality (6.68), induction hypothesis and Hölder's inequality to obtain

$$\mathbb{E}(\xi_{i,n,k}\eta_{i,n,k}^3) \leq (\mathbb{E}(\xi_{i,n,k}^4))^{\frac{1}{4}}(\mathbb{E}(\eta_{i,n,k}^4))^{\frac{3}{4}} \leq C'' L_1^3 \bar{d}^3 e^{CL_1 \cdot t_k^n} \Delta t_n^{7\alpha}.$$

Now we are ready to prove the induction hypothesis. Let

$$C = 12C''\bar{d}^4 + 6\bar{d}^4 + 1 \quad \text{and} \quad poly(x) = \left(C''(x^4\bar{d}^4 + 3x^3\bar{d}^3 + 12x^2\bar{d}^4)\right)^3. \tag{6.70}$$

It is easy to check that $C > 1$ and the polynomial $poly(x) > 1$ for $x > 1$. Then, it follows from Definition 9 and standard calculation that if n is large enough that $2^n > poly(L_1)$ (i.e., $\Delta t_n < (poly(L_1))^{-1}$), then

$$\begin{aligned}
&C'' L_1^4 \bar{d}^4 \Delta t_n^{4\alpha-1} + 3C'' L_1^3 \bar{d}^3 \Delta t_n^{3\alpha-1} + 12C'' L_1^2 \bar{d}^4 \Delta t_n \\
\leq &(C'' L_1^4 \bar{d}^4 + 3C'' L_1^3 \bar{d}^3 + 12C'' L_1^2 \bar{d}^4)\Delta t_n^{3\alpha-1} \\
= &(poly(L_1))^{\frac{1}{3}} \Delta t_n^{3\alpha-1} < (\mathcal{P}(L_1))^{\frac{4}{3}-3\alpha} < 1,
\end{aligned} \tag{6.71}$$

where the last inequality follows $\Delta t_n = 2^{-n} \leq (poly(L_1))^{-1}$,$\frac{4}{3} - 3\alpha < 0$, $L_1 > 1$ and $poly(x) > 1$ for $x > 1$. Thus, for n such that $2^n > poly(L_1)$, we use (6.65), Hölder's inequality and the bounds above to obtain

$$\begin{aligned}
&e^{CL_1 \cdot t_k^n} \Delta t_n^{4\alpha} \Big(1 + C'' L_1^4 \bar{d}^4 \Delta t_n^{4\alpha} \\
&+ 6L_1 \bar{d}^4 \Delta t_n + 3C'' L_1^3 \bar{d}^3 \Delta t_n^{3\alpha} + 12C'' \bar{d}^4 (L_1^2 \Delta t_n + 1)\Delta t_n \Big) \\
\leq &e^{CL_1 t_k^n} \Delta t_n^{4\alpha} \Big(1 + (6\bar{d}^4 + 12C''\bar{d}^4 + 1)L_1 \Delta t_n \Big) \\
= &e^{CL_1 t_k^n} \Delta t_n^{4\alpha} (1 + CL_1 \Delta t_n) \leq e^{CL_1 t_{k+1}^n} \Delta t_n^{4\alpha},
\end{aligned}$$

where the last line follows from convexity of exponential function: $e^y \geq e^x + e^x \cdot (y-x)$ for $y \geq x$. The second to last inequality follows from (6.70), (6.71) and the fact that $L_1 > 1$. This concludes the induction. However, since $t_n^k \leq 1$ for all $0 \leq k \leq 2^n$, we have proven that when $2^n > poly(L_1)$ (i.e., $\Delta t_n < (poly(L_1))^{-1}$),

$$\mathbb{E}\|X_{i,n}^{\mu}(t) - X_{i,n}(t)\|_{\infty}^4 \leq e^{CL_1} \cdot \Delta t_n^{4\alpha}, \tag{6.72}$$

for all $i \in [d]$ and $t \in [0,1]$.

Proof of statement (II) We extend the result to the case when $2^n \leq poly(L_1)$. By observing (6.63), we can find some $poly'(\cdot)$ with $poly'(x) > 1$ for $x > 1$ so that:

$$\left|(X_{i,n}^{\mu}(t_{k+1}^n) - X_{i,n}(t_{k+1}^n)) - (X_{i,n}^{\mu}(t_k^n) - X_{i,n}(t_k^n))\right| \leq poly'(L_1, \|B\|_{\alpha}, \|\widetilde{A}\|_{2\alpha})\Delta t_n^{\alpha}.$$

Since the number of iterations in the discretization is at most $2^n \leq poly(L_1)$, thus $\|X_{i,n}^{\mu}(\cdot) - X_{i,n}(\cdot)\|_{\infty} \leq poly(L_1)poly'(L_1, \|B\|_{\alpha}, \|\widetilde{A}\|_{2\alpha})\Delta t_n^{\alpha}$, and from Lemma 24:

$$\mathbb{E}\|X_{i,n}^{\mu}(\cdot) - X_{i,n}(\cdot)\|_{\infty}^4 \leq poly''(L_1)\Delta t_n^{4\alpha}, \tag{6.73}$$

for some $poly''(\cdot)$ with $poly''(x) > 1$ for $x > 1$. This concludes the proof of Lemma 14. □

The proof for Lemma 16 is similar and left in the supplementary materials along with other technical lemmas.

6.7 Supplementary Material

We use (#)(SP) to quote equations in this supplementary material, to distinguish from equations in the main text.

Lemma 27. *Let $X_{n+1}^f(1)$ and $X_{n+1}^a(1)$ from (3.7) with $\mu^{(n+1)}(\cdot) \in \mathcal{L}(L_1)$. Then, there exists*

$poly(\cdot) : \mathbb{R} \to \mathbb{R}$ *such that*

$$\mathbb{E}\|X^f_{n+1}(1) - X^a_{n+1}(1)\|^8_\infty \leq poly(L_1)\Delta t_n^{8(2\alpha-\beta)}. \tag{6.74}$$

Moreover, $poly(x) > 1$ *for* $x > 1$.

6.8 Proof of Lemma 4.3

Proof of (4.6) in Lemma 4.3. Assume w.l.o.g. $x = 0$. Given $\mu \cup \{\mu^{(n)}\}_n \subseteq \mathcal{L}(L_1), L_1 > 1$, similar as in [181], we can use Burkholder-Davis-Gundy inequality [203] to find $C > 1$ that

$$\mathbb{E} \sup_{1\leq k\leq 2^n} |X_{i,n}(t^n_k)|^4 \leq C(L_1^4 + L^4 2^n(\sum_{k=0}^{2^n} \Delta t_n^2 + \sum_{k=0}^{2^n} \Delta t_n^4)) < \mathcal{P}(L_1),$$

for some polynomial function $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$. Finally, the claim on $\mathbb{E}|f(X_{n_0}(1))|^4$ follows from the bound on $\|D_x f\|_\infty$ in Assumption 2. □

Proof of (4.5) in Lemma 4.3. It follows from Equation (4.8) in [181], $|\Delta_n|^p$ is bounded by

$$\begin{aligned}|\Delta_n|^p \leq & 2^{p-1}L^p\mathbb{E}\|\frac{1}{2}(X_{n+1}(1) + X^a_{n+1}(1)) - X_n(1)\|^p_\infty \\ & + 2^{-p-1}L^p\mathbb{E}\|X_{n+1}(1) - X^a_{n+1}(1)\|^{2p}_\infty.\end{aligned}$$

when we have $p \geq 2$. Thus, to prove Lemma 4.3 and $\mathbb{E}(\Delta_n^4) \leq e^{CL_1}\Delta t_n^{4-\delta}$, it is sufficient to bound $\mathbb{E}\|\frac{1}{2}(X_{n+1}(1) + X^a_{n+1}(1)) - X_n(1)\|^4_\infty$ and $\mathbb{E}\|X_{n+1}(1) - X^a_{n+1}(1)\|^8_\infty$. Note that bound on $\mathbb{E}\|X_{n+1}(1) - X^a_{n+1}(1)\|^8_\infty$ is provided by Lemma 27 since $4 - \delta < 8(2\alpha - \beta)$ as in Definition A.1 and $poly(L_1) < e^{CL_1}$ for appropriately chosen $C > 1$.

It remains for us to bound $\mathbb{E}\|\frac{1}{2}(X_{n+1}(1) + X^a_{n+1}(1)) - X_n(1)\|^4_\infty$. First we write the recursion for

$X_{n+1}(\cdot)$ over the coarse step Δt_n instead of Δt_{n+1}:

$$
\begin{aligned}
X_{i,n+1}(t^n_{k+1}) &= X_{i,n+1}(t^n_k) + \mu_i^{(n+1)}(X^f_{n+1}(t^n_k)) + \sum_{j=1}^{d'} \sigma_{ij}(X_{n+1}(t^n_k))\Delta B^n_{j,k} \\
&+ \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \frac{\partial \sigma_{ij}}{\partial x_l}(X_{n+1}(t^n_k))\sigma_{lm}(X_{n+1}(t^n_k))\widetilde{A}_{mj}(t^n_k, t^n_{k+1}) + N^f_{i,n,k} + M^{f,(1)}_{i,n,k} + M^{f,(2)}_{i,n,k} + M^{f,(3)}_{i,n,k} \\
&- \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \frac{\partial \sigma_{ij}}{\partial x_l}(X_{n+1}(t^n_k))\sigma_{lm}(X_{n+1}(t^n_k))(\Delta B^{n+1}_{j,2k}\Delta B^{n+1}_{m,2k+1} - \Delta B^{n+1}_{m,2k}\Delta B^{n+1}_{j,2k+1}),
\end{aligned}
$$

where we define

$$
\begin{aligned}
M^{f,(2)}_{i,n,k} &\triangleq \Big(\sum_{j=1}^{d'} (\sigma_{ij}(X_{n+1}(t^{n+1}_{2k+1})) - \sigma_{ij}(X_{n+1}(t^n_k))) \\
&- \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \big(\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm}\big)(X_{n+1}(t^n_k))\Delta B^{n+1}_{m,2k}\Big)\Delta B^{n+1}_{j,2k+1},
\end{aligned}
$$

$$
\begin{aligned}
M^{f,(3)}_{i,n,k} &\triangleq \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'} \Big(\big(\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm}\big)(X_{n+1}(t^{n+1}_{2k+1})) \\
&\qquad - \big(\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm}\big)(X_{n+1}(t^n_k))\Big)\widetilde{A}_{mj}(t^{n+1}_{2k+1}, t^n_{k+1}), \\
M^{f,(1)}_{i,n,k} &\triangleq \Big(\sum_{j=1}^{d} \frac{\partial \mu_i^{(n+1)}}{\partial x_j}(X_{n+1}(t^n_k)) \sum_{m=1}^{d'} \sigma_{jm}(X_{n+1}(t^n_k))\Delta B^{n+1}_{m,2k}\Big)\frac{\Delta t_n}{2},
\end{aligned}
$$

$$
\begin{aligned}
N^f_{i,n,k} &\triangleq \big(\mu_i^{(n+1)}(X_{n+1}(t^{n+1}_{2k+1})) - \mu_i^{(n+1)}(X_{n+1}(t^n_k))\big)\frac{\Delta t_n}{2} - M^{f,(1)}_{i,n,k} \\
=&\Big(\sum_{j=1}^{d}\frac{\partial \mu_i^{(n+1)}}{\partial x_j}(X_{n+1}(t^n_k))(X_{j,n+1}(t^{n+1}_{2k+1}) - X_{j,n+1}(t^n_k)) \\
&+\frac{1}{2}\sum_{j=1}^{d}\sum_{m=1}^{d}\frac{\partial^2 \mu_i^{(n+1)}}{\partial x_j \partial x_m}(\eta)(X_{j,n+1}(t^{n+1}_{2k+1}) - X_{j,n+1}(t^n_k))(X_{m,n+1}(t^{n+1}_{2k+1}) - X_{m,n+1}(t^n_k))\Big)\frac{\Delta t_n}{2} - M^{f,(1)}_{i,n,k} \\
=&\Big(\sum_{j=1}^{d}\frac{\partial \mu_i^{(n+1)}}{\partial x_j}(X_{n+1}(t^n_k))\Big(\mu_j^{(n+1)}(X_{n+1}(t^n_k))\frac{\Delta t_n}{2} + \sum_{m,l,\widetilde{m}}\big(\frac{\partial \sigma_{jm}}{\partial x_l}\cdot\sigma_{l\widetilde{m}}\big)(X_{n+1}(t^n_k))\widetilde{A}_{m\widetilde{m}}(t^n_k, t^{n+1}_{2k+1})\Big) \\
&+\frac{1}{2}\sum_{j,m=1}^{d}\frac{\partial^2 \mu_i^{(n+1)}}{\partial x_j \partial x_m}(\rho)(X_{j,n+1}(t^{n+1}_{2k+1}) - X_{j,n+1}(t^n_k))(X_{m,n+1}(t^{n+1}_{2k+1}) - X_{m,n+1}(t^n_k))\Big)\frac{\Delta t_n}{2},
\end{aligned}
$$

for some ρ that lies between $X_{n+1}(t^n_k)$ and $X_{n+1}(t^{n+1}_{2k+1})$. Furthermore, we similarly define $N^a_{i,n,k}$, $M^{a,(\cdot)}_{i,n,k}$ associated with $X^a_{n+1}(\cdot)$, $B^{n+1,a}(t)$ and $\widetilde{A}^a(t^{n+1}_{k+1}, t^{n+1}_k)$ so that:

$$
\begin{aligned}
X^a_{i,n+1}(t^n_{k+1}) &= X^a_{i,n+1}(t^n_k) + \mu_i^{(n+1)}(X^a_{n+1}(t^n_k)) + \sum_{j=1}^{d'}\sigma_{ij}(X^a_{n+1}(t^n_k))\Delta B^n_{j,k} \\
&+\sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\frac{\partial \sigma_{ij}}{\partial x_l}(X^a_{n+1}(t^n_k))\sigma_{lm}(X^a_{n+1}(t^n_k))\widetilde{A}^a_{mj}(t^n_k, t^n_{k+1}) + N^a_{i,n,k} + M^{a,(1)}_{i,n,k} + M^{a,(2)}_{i,n,k} + M^{a,(3)}_{i,n,k}. \\
&-\sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\frac{\partial \sigma_{ij}}{\partial x_l}(X^a_{n+1}(t^n_k))\sigma_{lm}(X^a_{n+1}(t^n_k))(\Delta B^{n+1,a}_{j,2k}\Delta B^{n+1,a}_{m,2k+1} - \Delta B^{n+1,a}_{m,2k}\Delta B^{n+1,a}_{j,2k+1}).
\end{aligned}
$$

Now, we write the recursion for $\bar{X}_{n+1}(\cdot) \triangleq \frac{1}{2}\big(X_{n+1}(\cdot) + X^a_{n+1}(\cdot)\big)$ over Δt_n:

$$\begin{aligned}
\bar{X}_{i,n+1}(t^n_{k+1}) =& \bar{X}_{i,n+1}(t^n_k) + \mu_i^{(n+1)}(\bar{X}_{n+1}(t^n_k))\Delta t_n + \sum_{j=1}^{d'} \sigma_{ij}(\bar{X}_{n+1}(t^n_k))\Delta B^n_{j,k} \\
&+ \sum_{j,m=1}^{d'} \sum_{l=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_l}(\bar{X}_{n+1}(t^n_k))\sigma_{lm}(\bar{X}_{n+1}(t^n_k))\widetilde{A}_{mj}(t^n_k, t^n_{k+1}) + R_{i,n,k}, \\
R_{i,n,k} \triangleq& N^{(1)}_{i,n,k} + M^{(1)}_{i,n,k} + M^{(2)}_{i,n,k} + M^{(3)}_{i,n,k} \\
&+ \frac{1}{2}(N^f_{i,n,k} + M^{f,(1)}_{i,n,k} + M^{f,(2)}_{i,n,k} + M^{f,(3)}_{i,n,k} + N^a_{i,n,k} + M^{a,(1)}_{i,n,k} + M^{a,(2)}_{i,n,k} + M^{a,(3)}_{i,n,k}), \\
N^{(1)}_{i,n,k} \triangleq& \frac{1}{2}\Big(\mu_i^{(n+1)}(X_{n+1}(t^n_k)) + \mu_i^{(n+1)}(X^a_{n+1}(t^n_k))\Big) - \mu_i^{(n+1)}(\bar{X}_{n+1}(t^n_k)),
\end{aligned}$$

$$\begin{aligned}
M^{(1)}_{i,n,k} \triangleq& \sum_{j=1}^{d'} \Big(\frac{1}{2}\big(\sigma_{ij}(X_{n+1}(t^n_k)) + \sigma_{ij}(X^a_{n+1}(t^n_k))\big) - \sigma_{ij}(\bar{X}_{n+1}(t^n_k))\Big)\Delta B^n_{j,k}, \\
M^{(2)}_{i,n,k} \triangleq& \sum_{j,m=1}^{d'} \sum_{l=1}^{d} \Big(\frac{1}{2}\big((\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X_{n+1}(t^n_k)) + (\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X^a_{n+1}(t^n_k))\big) \\
&- (\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(\bar{X}_{n+1}(t^n_k))\Big)\widetilde{A}_{mj}(t^n_k, t^n_{k+1}), \\
M^{(3)}_{i,n,k} \triangleq& \sum_{j,m=1}^{d'} \sum_{l=1}^{d} \frac{1}{2}\big((\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X_{n+1}(t^n_k)) - (\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X^a_{n+1}(t^n_k))\big) \\
&\cdot (\Delta B^{n+1}_{j,2k}\Delta B^{n+1}_{m,2k+1} - \Delta B^{n+1}_{m,2k}\Delta B^{n+1}_{j,2k+1}).
\end{aligned}$$

Finally, subtract the recursion in Num_Sol for $X_n(\cdot)$ from $\bar{X}_n(\cdot)$ to obtain

$$\begin{aligned}
&\bar{X}_{i,n+1}(t^n_{k+1}) - X_{i,n}(t^n_{k+1}) \\
=&\bar{X}_{i,n+1}(t^n_k) - X_{i,n}(t^n_k) + (\mu^{(n)}_i(\bar{X}_{n+1}(t^n_k)) - \mu^{(n)}_i(X_n(t^n_k)))\Delta t_n \\
&+ (\mu^{(n+1)}_i - \mu^{(n)}_i)(\bar{X}_{i,n+1}(t^n_k))\Delta t_n \\
&+ \sum_{j=1}^{d'} \big(\sigma_{ij}(\bar{X}_{i,n+1}(t^n_k)) - \sigma_{ij}(X_{i,n}(t^n_k))\big)\Delta B^n_{j,k} \\
&+ \sum_{j,m=1}^{d'} \sum_{l=1}^{d} \big((\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(\bar{X}_{i,n+1}(t^n_k)) - (\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X_{i,n}(t^n_k))\big)\widetilde{A}_{mj}(t^n_k, t^n_{k+1}) \\
&+ R_{i,n,k}.
\end{aligned}$$

Now, similarly as in the proof of Lemma 4.1, we simplify the notation by defining

$$\xi_{i,n,k} \triangleq \bar{X}_{i,n+1}(t^n_k) - X_{i,n}(t^n_k) \qquad \text{and} \qquad \xi_{n,k} \triangleq \bar{X}_{n+1}(t^n_k) - X_n(t^n_k),$$

and also

$$\begin{aligned}
&\eta_{i,n,k} \\
\triangleq&(\mu^{(n)}_i(\bar{X}_{n+1}(t^n_k)) - \mu^{(n)}_i(X_n(t^n_k)))\Delta t_n + \sum_{j=1}^{d'} \big(\sigma_{ij}(\bar{X}_{i,n+1}(t^n_k)) - \sigma_{ij}(X_{i,n}(t^n_k))\big)\Delta B^n_{j,k} \\
&+ \sum_{j,m=1}^{d'} \sum_{l=1}^{d} \big((\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(\bar{X}_{i,n+1}(t^n_k)) - (\frac{\partial \sigma_{ij}}{\partial x_l} \cdot \sigma_{lm})(X_{i,n}(t^n_k))\big)\widetilde{A}_{mj}(t^n_k, t^n_{k+1}) \\
&+ R_{i,n,k} + (\mu^{(n+1)}_i - \mu^{(n)}_i)(\bar{X}_{i,n+1}(t^n_k))\Delta t_n,
\end{aligned}$$

for $0 \le k \le 2^n - 1$, so that $\xi_{i,n,k+1} = \xi_{i,n,k} + \eta_{i,n,k}$. Given $\mu \cup \{\mu^{(n)}\}_n \subseteq \mathcal{L}(L_1), L_1 > 1$, we want to find $C > 1$ and $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$, such that if $2^n > poly(L_1)$, then

$$\mathbb{E}(\xi_{i,n,k})^4 \le e^{CL_1 \cdot t^n_k} \Delta t_n^{4-\delta}, \tag{6.75}$$

for all $i \in [d]$ and $0 \le k \le 2^n$. Similarly, we prove by induction on $0 \le k \le 2^n$ and by bounding the terms of $R_{i,n,k}$. We start by bounding $N^{(1)}_{i,n,k}$ using Taylor expansion,

$$
\begin{aligned}
N^{(1)}_{i,n,k} &\triangleq \frac{1}{2}\Big(\mu_i^{(n+1)}(X_{n+1}(t_k^n)) - \mu_i^{(n+1)}(X^a_{n+1}(t_k^n))\Big) - \mu_i^{(n+1)}(\bar{X}_{n+1}(t_k^n)) \\
&= \sum_{j,m=1}^{d} \Big(\frac{\partial^2 \mu_i(\rho_1)}{\partial x_j \partial x_m} + \frac{\partial^2 \mu_i(\rho_1^{'})}{\partial x_j \partial x_m}\Big)(X_{j,n+1}(t_k^n) - X^a_{j,n+1}(t_k^n))(X_{m,n+1}(t_k^n) - X^a_{m,n+1}(t_k^n))\frac{\Delta t_n}{16},
\end{aligned}
$$

where ρ_1 and ρ_1' lie somewhere between $X^a_{n+1}(t_k^n)$ and $X_{n+1}(t_k^n)$. Now we use Lemma 6.74 on $(X_{j,n+1}(t_k^n) - X^a_{j,n+1}(t_k^n))(X_{m,n+1}(t_k^n) - X^a_{m,n+1}(t_k^n))$ and Hölder's inequality to get

$$
\mathbb{E}(N^{(1)}_{i,n,k})^4 < poly(L_1)\Delta t_n^{8(2\alpha-\beta)+4}
$$

$poly(\cdot)$ with $poly(x) > 1$ when $x > 1$. In fact, based on the similar analysis on

$$
N^f_{i,n,k}, M^{(1)}_{i,n,k}, M^{(2)}_{i,n,k}, M^{(3)}_{i,n,k}, M^{f,(1)}_{i,n,k}, M^{f,(2)}_{i,n,k}, M^{f,(3)}_{i,n,k}
$$

as above, we can find some $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$ that

$$
\mathbb{E}R^4_{i,n,k} \le poly(L_1)\Delta t_n^{8(2\alpha-\beta)+2}. \tag{6.76}
$$

Then, as in proof of Lemma 4.1, we prove the hypothesis in (6.75) by induction. First of all, when $k = 0$, for $i \in [d]$, the claim holds since $\xi_{i,n,0} \triangleq X^{\mu}_{i,n}(0) - X_{i,n}(0) = x - x = 0$. Now, fixing $0 \le k \le 2^n - 1$ and $i \in [d]$, suppose the induction hypothesis holds so that we can find $C > 1$ where $\mathbb{E}|\xi^4_{i,n,j}| \le e^{CL_1 \cdot t_j^n} \cdot \Delta t_n^{4-\delta}$ for all $0 \le j \le k$. We want to show, $\mathbb{E}|\xi^4_{i,n,k+1}| \le e^{CL_1 \cdot t_{k+1}^n} \cdot \Delta t_n^{4-\delta}$

for all $i \in [d]$. Use similar analysis as in proof of Lemma 4.1, we obtain

$$\begin{aligned}
\mathbb{E}\eta_{i,n,k}^4 \leq & e^{C_1 L_1 t_k^n} \Delta t_n^{5-\delta} (poly(L_1) \Delta t_n^{1+8(2\alpha-\beta)-(4-\delta)} + 2L_1^4 \Delta t_n^3 + 2L^4 \Delta t_n), \\
\mathbb{E}(\xi_{i,n,k}^3 \eta_{i,n,k}) \leq & e^{CL_1 t_k^n} \Delta t_n^{5-\delta} (2L_1 + 2L_1 \Delta t_n^{\frac{\delta}{4}+2(2\alpha-\beta)-1}), \\
\mathbb{E}\xi_{i,n,k}(\eta_{i,n,k})^3 \leq & e^{CL_1 t_k^n} \Delta t_n^{5-\delta} (poly(L_1) + 2L_1^4 + 2L^4) \Delta t_n^{\frac{1}{2}}, \\
\mathbb{E}(\xi_{i,n,k})^2 (\eta_{i,n,k})^2 \leq & e^{CL_1 t_k^n} \Delta t_n^{5-\delta} \left(polyP(L_1) \Delta t_n^{8(2\alpha-\beta)-(4-\delta)} + 2L_1^4 \Delta t_n^2 + 2L^4 \right)^{\frac{1}{2}}.
\end{aligned} \tag{6.77}$$

for some $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$. Thus, we can find some $poly'(\cdot)$ with $poly'(x) > 1$ when $x > 1$, such that $2^n > poly'(L_1)$, we can find Let $C = 5 + 2L > 1$ that

$$\begin{aligned}
(2L_1 + 2L_1 \Delta t_n^{\frac{\delta}{4}+2(2\alpha-\beta)-1}) \leq & 3L_1, \\
(poly(L_1) \Delta t_n^{1+8(2\alpha-\beta)-(4-\delta)} + 2L_1^4 \Delta t_n^3 + 2L^4 \Delta t_n) \leq & 1, \\
(poly(L_1) + 2L_1^4 + 2L^4) \Delta t_n^{\frac{1}{2}} \leq & 1, \\
\left(poly(L_1) \Delta t_n^{8(2\alpha-\beta)-(4-\delta)} + 2L_1^4 \Delta t_n^2 + 2L^4 \right)^{\frac{1}{2}} \leq & 2L^2.
\end{aligned} \tag{6.78}$$

Consequently, when $2^n > poly'(L_1)$, we use the bound in Equations (6.78) (SP) to obtain

$$\begin{aligned}
\mathbb{E}(\xi_{i,n,k+1})^4 \leq & e^{CL_1 t_k^n} \Delta t_n^{4-\delta} + e^{CL_1 t_k^n} \Delta t_n^{5-\delta} (3L_1 + 2 + 2L^2) \\
\leq & e^{CL_1 t_k^n} \Delta t_n^{4-\delta} \cdot (1 + CL_1 \Delta t_n) \leq e^{CL_1 t_{k+1}^n} \Delta t_n^{4-\delta},
\end{aligned} \tag{6.79}$$

where the last line follows from convexity of exponential function $e^y \geq e^x + e^x \cdot (y - x)$ for $y \geq x$. Now we use the method as in the proof of Lemma 4.1 to extend the induction hypothesis to the case where $\Delta t_n \leq poly'(L_1)$. This concludes the proof. □

6.9 Proof of Supporting Lemmas

First, we use the Levy-Ciesielski construction of the Brownian motion (see [202]).

Lemma 28. *Let* $\{U_j^m : 1 \leq j \leq 2^{m-1}, m \geq 1\}$ *along with* U_0^0 *be a sequence of I.I.D standard*

normal random variables, and we define

$$H(t) \triangleq \mathbf{I}(0 \le t < 1/2) - \mathbf{I}(1/2 \le t \le 1), \tag{6.80}$$

along with its family of functions $\{H_j^m(t) = 2^{m/2}H(2^{m-1}t - j + 1) : 1 \le j \le 2^{m-1}, m \ge 1\}$ *and constant function* $H_0^0(\cdot) = 1$. *Now, if we define* $B(t)$ *for* $t \in [0,1]$ *by*

$$B(t) \triangleq U_0^0 \int_0^t H_0^0(s)ds + \sum_{m\ge 1}\sum_{j=1}^{2^{m-1}} \left(U_j^m \int_0^t H_j^m(s)ds\right), \tag{6.81}$$

then it can be shown that the right-hand side converges uniformly on [0,1] almost surely and the process $\{B(t)\}_{t\in[0,1]}$ *is a standard Brownian motion on [0,1].*

Proof. See Section 2.3 of [201]. □

Changing the sign of a standard Gaussian does not change its distribution. Thus the above theoretical construction a way to define $B^{(n+1),a}(t)$ related to Definition 4.

Corollary 6.9.0.1. *Fixing* $n \ge 0$ *and the sequence of I.I.D. standard Gaussian* $\{U_j^m : 1 \le j \le 2^{m-1}, m \ge 1\}$ *along with* U_0^0, *we can define*

$$B^{n+1,a}(t) \triangleq U_0^0 \int_0^t H_0^0(s)ds + \sum_{j=1}^{2^n}\left(-U_j^{n+1}\int_0^t H_j^{n+1}(s)ds\right) + \sum_{\substack{m\ge 1\\ m\ne n+1}}\sum_{j=1}^{2^{m-1}}\left(U_j^m\int_0^t H_j^m(s)ds\right), \tag{6.82}$$

which is a again Brownian motion on [0,1].

Lemma 29. *Given a sequence of I.I.D. standard Gaussian* $\{U_j^m : 1 \le j \le 2^{m-1}, m \ge 1\}$. *For* $0 \le t \le 1$ *and* $n \ge 0$, *define* $B(t)$ *as in* (6.81) *and* $B^{n+1,a}(t)$ *as in* (6.82). *Let*

$$\begin{aligned}\Delta B_k^{n+1} &= B(t_{k+1}^{n+1}) - B(t_k^{n+1}),\\ \Delta B_k^{n+1,a} &= B^{(n+1),a}(t_{k+1}^{n+1}) - B^{(n+1),a}(t_k^{n+1}).\end{aligned} \tag{6.83}$$

for $1 \le k \le 2^{n+1}-1$. *Then* ΔB_k^{n+1} *and* $\Delta B_k^{n+1,a}$ *satisfy equations (2.5) and (2.6) in Definition 2.3. Thus, we may regard* $X_{n+1}^a(\cdot)$ *to be* $X_{n+1}(\cdot)$ *generated under Brownian motion* $B^{n+1,a}(\cdot)$ *instead of* $B(\cdot)$.

Proof of Lemma 29. By Definition 28, for $n \ge 1$ and $0 \le k \le 2^{n-1}-1$,

$$\begin{cases} \int_{t_{2k}^n}^{t_{2k+1}^n} H_j^m(t)dt = \int_{t_{2k+1}^n}^{t_{2k+2}^n} H_j^m(t)dt & \text{for all} \quad m \ne n \quad \text{and} \quad 1 \le j \le 2^{m-1} \\ \int_{t_{2k}^n}^{t_{2k+1}^n} H_j^m(t)dt = -\int_{t_{2k+1}^n}^{t_{2k+2}^n} H_j^m(t)dt & \text{for all} \quad m = n \quad \text{and} \quad 1 \le j \le 2^{m-1}. \end{cases} \tag{6.84}$$

Thus, we have that, for $0 \le k \le 2^n - 1$,

$$\begin{aligned} B^{n+1,a}(t_{2k+1}^{n+1}) - B^{n+1,a}(t_{2k}^{n+1}) =& B(t_{2k+2}^{n+1}) - B(t_{2k+1}^{n+1}) = \Delta B_{2k}^{n+1,a}, \\ B^{n+1,a}(t_{2k+2}^{n+1}) - B^{n+1,a}(t_{2k+1}^{n+1}) =& B(t_{2k+1}^{n+1}) - B(t_{2k}^{n+1}) = \Delta B_{2k+1}^{n+1,a}, \end{aligned}$$

by simply taking the difference in (6.82) and checking (6.84). □

Proof of Lemma A.2. Following Definition A.2, define $R_{i,j}^n(t_l^n, t_m^n) = \sum_{k=l+1}^m A_{i,j}(t_{k-1}^n, t_k^n)$ for $0 \le l < m \le 2^n$, $i, j \in [d']$ and $i \ne j$. Then, we can define

$$\Gamma_R \triangleq \sup_{n\ge 1} \sup_{\substack{0\le s\le t\le 1 \\ s,t\in D_n}} \max_{1\le i,j\le d', i\ne j} \frac{|R_{i,j}^n(s,t)|}{|t-s|^\beta \Delta t_n^{2\alpha-\beta}},$$

$$\Gamma_{R-\tilde{R}} \triangleq \sup_{n\ge 1} \sup_{\substack{0\le s\le t\le 1 \\ s,t\in D_n}} \max_{1\le i,j\le d', i\ne j} \frac{|R_{i,j}^n(s,t) - \tilde{R}_{i,j}^n(s,t)|}{|t-s|^\beta \Delta t_n^{2\alpha-\beta}},$$

Observing the definition for both the case $i = j$ and $i \ne j$, we have

$$\|\tilde{A}\|_{2\alpha} \le \|A\|_{2\alpha} + \|B\|_\alpha^2 \qquad \text{and} \qquad \Gamma_{\tilde{R}} \le \Gamma_R + \Gamma_{R-\tilde{R}}. \tag{6.85}$$

Now, following Lemma 3.1 in [200], we define a family of random variables $(L_{i,j}^n(k) : k =$

$0, 1, ..., 2^{n-1}, i, j \in [d'], i \neq j, n \geq 1$) satisfying $L^n_{i,j}(0) = 0$ and

$$L^n_{i,j}(k) = L^n_{i,j}(k-1) + \left(B_i\left(t^n_{2k-1}\right) - B_i\left(t^n_{2k-2}\right)\right)\left(B_j\left(t^n_{2k}\right) - B_j\left(t^n_{2k-1}\right)\right).$$

Then, following Lemma 3.4 and its proof in [200], we define, for $i, j \in [d']$ and $i \neq j$,

$$N_{i,j,2} = \max\{n : |L^n_{i,j}(m) - L^n_{i,j}(l)| > (m-l)^\beta \Delta t_n^{2\alpha} \quad \text{for some} \quad 0 \leq l < m \leq 2^{n-1}\},$$

and define $N_2 = \max\{N_{i,j,2} : i, j \in [d'], i \neq j\}$ along with

$$\Gamma_L \triangleq \max\{1, \max_{1\leq i,j\leq d', i\neq j} \max_{n<N_2} \max_{0\leq l<m<2^{n-1}} \frac{|L^n_{i,j}(m) - L^n_{i,j}(l)|}{(m-l)^\beta \Delta t_n^{2\alpha}}\}.$$

Finally, we use Definition A.2 and apply the result of Lemma 3.5 in [200] to write:

$$\Gamma_R \leq \frac{2^{-(2\alpha-\beta)}}{1-2^{-(2\alpha-\beta)}} \cdot \Gamma_L, \qquad \text{and} \qquad \|A\|_{2\alpha} \leq \Gamma_R \cdot \frac{2}{1-2^{-2\alpha}} + \|B\|^2_\alpha \cdot \frac{2^{1-\alpha}}{1-2^{-\alpha}}.$$

Thus, it suffices to show that $\|B\|_\alpha, \Gamma_L$ and $\Gamma_{R-\tilde{R}}$ has finite moments of every order. For $\|B\|_\alpha$, it follows from Borell's inequality for continuous Gaussian random fields (see Section 2.3 of [204]). To show the result for Γ_L, we follow the proof of Lemma 3.4 in [200] to show that $\mathbb{P}(N_{i,j,2} \geq n)$ is bounded by

$$\begin{aligned}
&\sum_{h=n}^{\infty} \mathbb{P}(|L^h_{i,j}(m) - L^h_{i,j}(l)| > (m-l)^\beta \Delta t_n^{2\alpha} \quad \text{for some} \quad 0 \leq l < m \leq 2^{n-1}) \\
&\leq \sum_{h=n}^{\infty} 2^{2h} \exp(-\theta' 2^{h(1-2\alpha)}) \leq \exp(-\frac{\theta'}{2} \cdot 2^{n(1-2\alpha)}) \sum_{h=0}^{\infty} 2^{2h} \exp(-\frac{\theta'}{2} \cdot 2^{h(1-2\alpha)}) \\
&\leq C \exp(-\frac{\theta'}{2} \cdot 2^{n(1-2\alpha)}),
\end{aligned} \tag{6.86}$$

for some $C > 1$ and $\theta' > 0$. It follows that,

$$\mathbb{P}(N_2 \geq n) \leq C(d')^2 \exp(-\frac{\theta'}{2} \cdot 2^{n(1-2\alpha)}), \tag{6.87}$$

$$\mathbb{E}\left(\exp\left(\eta N_2\right)\right) \leq \sum_{n=1}^{\infty} C(d')^2 \exp\left(\eta n\right) \exp(-\frac{\theta'}{2} \cdot 2^{n(1-2\alpha)}) < \infty, \tag{6.88}$$

for every $\eta > 0$. On the other hand, since for $m > l, n \leq N_2$, we have

$$(m-l)^{-\beta} \Delta t_n^{-2\alpha} = (m-l)^{-\beta} 2^{2\alpha n} \leq 2^{2\alpha N_2},$$

$$\Gamma_L \leq 1 + 2^{2\alpha N_2} \cdot \big(\max_{1 \leq i,j \leq d', i \neq j} \max_{n < N_2} \max_{0 \leq l < m < 2^{n-1}} |L_{i,j}^n(m) - L_{i,j}^n(l)| \big).$$

Since N_2 has a finite moment-generating function on the real line according to (6.88), in order to establish that Γ_L has finite moments of every order, it suffices to show that

$$\mathbb{E}[(\sum_{n=1}^{N_2} \sum_{1<l<m}^{2^{n-1}} \sum_{1 \leq i \neq j}^{d'} |L_{i,j}^n(m) - L_{i,j}^n(l)|)^k] < \infty.$$

for every $k \geq 1$. Letting $\bar{n}$ be the number of total elements being summed up inside the previous expectation, it follows that $\bar{n} \leq N_2 \cdot 2^{2N_2} (d')^2$ and therefore, by (6.27)

$$\begin{aligned}
&\mathbb{E}(\sum_{n=1}^{N_2} \sum_{1<l<m}^{2^{n-1}} \sum_{1 \leq i \neq j}^{d'} |L_{i,j}^n(m) - L_{i,j}^n(l)|)^k \leq \mathbb{E}\bar{n}^{k-1} \sum_{n=1}^{N_2} \sum_{1<l<m}^{2^{n-1}} \sum_{1 \leq i \neq j}^{d'} |L_{i,j}^n(m) - L_{i,j}^n(l)|^k \\
&\leq \sum_{n=1}^{N_2} \sum_{1<l<m}^{2^{n-1}} \sum_{1 \leq i \neq j}^{d'} \mathbb{E}\left[(N_2 \cdot 2^{2N_2} (d')^2)^{k-1} |L_{i,j}^n(m) - L_{i,j}^n(l)|^k I(N_2 \geq n) \right].
\end{aligned}$$

To bound these terms, we first show that, fixing any $h \geq 1$, $\mathbb{E}|L_{i,j}^n(m) - L_{i,j}^n(l)|^h$ is uniformly bounded for any $n \geq 1, 1 \leq l < m \leq 2^{n-1}, 1 \leq i, j \leq n$ and $i \neq j$. Let $\{Y_{i'}\}_{i' \geq 1}$ be I.I.D. with $Y \overset{\mathcal{D}}{=} Z_1 \cdot Z_2$ where Z_1, Z_2 are independent standard Gaussian. It follows from Hölder's inequality and Jensen's inequality that we can find $C_h > 0$ such that $\mathbb{E}|\frac{\sum_{i=1}^n Y_{i'}}{n}|^h < C_h$ for all $n \geq 1$. Then $\mathbb{E}|L_{i,j}^n(m) - L_{i,j}^n(l)|^h < C_h$ follows from $|L_{i,j}^n(m) - L_{i,j}^n(l)| \overset{d}{=} |\Delta t_n \sum_{i'=1}^{m-l} Y_{i'}| \leq |\frac{\sum_{i'=1}^{m-l} Y_{i'}}{m-l}|$. Specifically $\mathbb{E}|L_{i,j}^n(m) - L_{i,j}^n(l)|^{4k} < C_{4k}$ for all $n \geq 1$. Now we can use Hölder's inequality multiple times and

the fact that N_2 has moment-generating function to conclude:

$$E\left[(d')^{2(k-1)}2^{3N_2(k-1)}|L^n_{i,j}(m)-L^n_{i,j}(l)|^k I(N_2\geq n)\right]\leq C' f(N_2\geq n)^{1/2},$$

for $C'>1$ and it follows from (6.87) that Γ_L has moments of every order. Finally, for $\Gamma_{R-\tilde{R}}$, define $\{\tilde{L}^n_{i,j}(k): k=0,1,...,2^n, i,j\in[d'], i\neq j, n\geq 1\}$ with $\tilde{L}^n_{i,j}(0)$ and

$$\tilde{L}^n_{i,j}(k)=\tilde{L}^n_{i,j}(k-1)+\left(B_i\left(t^n_k\right)-B_i\left(t^n_{k-1}\right)\right)\left(B_j\left(t^n_k\right)-B_j\left(t^n_{k-1}\right)\right),$$

$$\tilde{N}_2=\max\{n: |\tilde{L}^n_{i,j}(m)-\tilde{L}^n_{i,j}(l)|>(m-l)^\beta\Delta t_n^{2\alpha} \text{ for some } 0\leq l<m\leq 2^n, i,j\in[d'], i\neq j\},$$

$$\Gamma_{\tilde{L}}\triangleq\max\{1,\max_{1\leq i,j\leq d',i\neq j}\max_{n<\tilde{N}_2}\max_{0\leq l<m<2^{n-1}}\frac{|\tilde{L}^n_{i,j}(m)-\tilde{L}^n_{i,j}(l)|}{(m-l)^\beta\Delta t_n^{2\alpha}}\}.$$

Then, for $1\leq i,j\leq d', i\neq j, n\geq 1$ and $0\leq s<t\leq 1, s,t\in D_n$, we have

$$R^n_{i,j}(s,t)-\tilde{R}^n_{i,j}(s,t)=\sum_{k=s2^n+1}^{t2^n}\widetilde{A}_{i,j}(t^n_{k-1},t^n_k)=\tilde{L}^n_{i,j}(t2^n)-\tilde{L}^n_{i,j}(s2^n), \tag{6.89}$$

which implies $\Gamma_{R-\tilde{R}}\leq\Gamma_{\tilde{L}}$. We can now proceed to show $\Gamma_{\tilde{L}}$ has finite moments of every order in the similar fashion as we did for Γ_L. This completes the proof. □

Proof of Lemma A.3. Let $X^{\mathcal{M}}_n(\cdot)$ be the Milstein discretization for Δt_n :

$$X^{\mathcal{M}}_{i,n}(t^n_{k+1})=X^{\mathcal{M}}_{i,n}(t^n_k)+\mu_i(X^{\mathcal{M}}_n(t^n_k))\Delta t_n+\sum_{j=1}^{d'}\sigma_{ij}(X^{\mathcal{M}}_n(t^n_k))\Delta B^n_{j,k}$$
$$+\sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\frac{\partial\sigma_{ij}}{\partial x_l}(X^{\mathcal{M}}_n(t^n_k))\sigma_{lm}(X^{\mathcal{M}}_n(t^n_k))A_{mj}(t^n_k,t^n_{k+1}),$$

where we use $A_{ij}(s,t)$ instead of $\widetilde{A}_{ij}(s,t)$ defined in (24) (This distinguishes $X^{\mathcal{M}}_n(\cdot)$ from $X_n(\cdot)$, our antithetic scheme). Then, $\mu\cup\{\mu^{(n)}\}_n\subseteq\mathcal{L}(L_1), L_1>1$, we compute constant C_1 explicitly in terms of $L_1, \|B\|_\alpha$ and $\|A\|_{2\alpha}$ (denoted as $M, \|Z\|_\alpha$ and $\|A\|_{2\alpha}$ in [200]) such that for n large enough and $r,t\in D_n$, $\|X^{\mathcal{M}}_n(t)-X^{\mathcal{M}}_n(r)\|_\infty\leq C_1|t-r|^\alpha$. See page 305 of [200, Lemma 6.1].

To get the result for $X_n(\cdot)$ instead of $X_n^{\mathcal{M}}(\cdot)$, we follow page 283 of [200, Lemma 2.1], replacing $\|A\|_{2\alpha}$ by $\|\tilde{A}\|_{2\alpha}$ in notation, we define

$$\begin{cases} C_1(\delta) &= \bar{d}L_1\|B\|_\alpha + 1/2, \\ C_2(\delta) &= \bar{d}^3L_1^2\|A\|_{2\alpha} + 1/2, \\ C_3(\delta) &= \frac{2}{1-2^{1-3\alpha}}(\bar{d}L_1C_1(\delta)^2\|B\|_\alpha + \bar{d}^2L_1C_2(\delta)\|B\|_\alpha + \bar{d}^2L_1^2\|B\|_\alpha + 2\bar{d}^3L_1^2C_1(\delta)\|A\|_{2\alpha}), \end{cases}$$

and find some $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$ so that if

$$\delta = (\mathcal{P}(L_1, \|B\|_\alpha, \|A\|_{2\alpha}))^{-1},$$

then

$$C_3(\delta)\delta^{2\alpha} + L_1\delta^{1-\alpha} + \bar{d}^3L_1^2\|A\|_{2\alpha}\delta^\alpha < 1/2 \quad \text{and} \quad C_3(\delta)\delta^\alpha < 1/2, \tag{6.90}$$

so that Equation (6.4) in page 308 of [200, Lemma 6.1] is satisfied:

$$\begin{cases} C_1(\delta) &\geq \bar{d}L_1\|B\|_\alpha + L_1\delta^{1-\alpha} + \bar{d}L_1\|B\|_\alpha + \bar{d}^3L_1^2\|\widetilde{A}\|_{2\alpha}\delta^\alpha, \\ C_2(\delta) &\geq \bar{d}^3L_1^2\|A\|_{2\alpha} + \bar{d}^3L_1^2\|\widetilde{A}\|_{2\alpha}, \\ C_3(\delta) &\geq \frac{2}{1-2^{1-3\alpha}}(\bar{d}L_1C_1(\delta)^2\|B\|_\alpha + \bar{d}^2L_1C_2(\delta)\|B\|_\alpha + \bar{d}^2L_1^2\|B\|_\alpha + 2\bar{d}^3L_1^2C_1(\delta)\|A\|_{2\alpha}), \end{cases}$$

which gives, according to line $12-17$ of page 308 of [200, Lemma 6.1], that

$$\|X_n(t) - X_n(r)\|_\infty \leq \frac{2}{\delta}C_1(\delta)|t-r|^\alpha, \tag{6.91}$$

for all n large enough where $\Delta t_n \leq \frac{1}{2}\delta$. Notice we changed the result to address $X_n(\cdot)$ instead of $X_n^{\mathcal{M}}(\cdot)$, and so far it follows from an modification of [200, Lemma 6.1].

To extend the result for n where $\Delta t_n > \frac{\delta}{2}$, notice the recursion in Num_Sol is carried out at most 2^n number of times and $2^n = (\Delta t_n)^{-1} < 2(\delta)^{-1} = 2poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha})$. By analyzing Num_Sol,

we have

$$\begin{aligned}\|X_n(t_{k+1}^n) - X_n(t_k^n)\|_\infty \leq & \bar{d}(CL_1\Delta t_n + \bar{d}L\|B\|_\alpha \Delta t_n^\alpha + \bar{d}^3L^2\|A\|_{2\alpha}\Delta t_n^{2\alpha}) \\ \leq & \bar{d}(CL_1 + \bar{d}L\|B\|_\alpha + \bar{d}^3L^2\|A\|_{2\alpha})\Delta t_n^\alpha,\end{aligned}$$

for some $C > 1$. Since $\Delta t_n < 1$, thus, for $\Delta t_n > \frac{\delta}{2}$, $\|X_n(t) - X_n(r)\|_\infty$ is bounded by

$$\begin{aligned}&\frac{|t-r|\Delta t_n^\alpha}{\Delta t_n}\bar{d}(CL_1 + \bar{d}L\|B\|_\alpha + \bar{d}^3L^2\|A\|_{2\alpha}) \\ \leq & \bar{d}(CL_1 + \bar{d}L\|B\|_\alpha + \bar{d}^3L^2\|A\|_{2\alpha})\frac{|t-r|2^{1-\alpha}}{\delta^{1-\alpha}} \\ \leq & 2\bar{d}(CL_1 + \bar{d}L\|B\|_\alpha + \bar{d}^3L^2\|A\|_{2\alpha}) \cdot poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha}) \cdot |t-r|^\alpha, \end{aligned} \tag{6.92}$$

where the last line follows from $poly(x) > 1$ when $x > 1$ and $|t - r| < 1$. The second to last line follows from $\Delta t_n > \frac{\delta}{2}$. We now combine (6.91) and (6.92) and let

$$\begin{aligned}&poly'(L_1, \|B\|_\alpha, \|A\|_{2\alpha}) \\ \triangleq & 2\bar{d}(CL_1 + \bar{d}L\|B\|_\alpha + \bar{d}^3L^2\|A\|_{2\alpha}) \cdot poly(L_1, \|B\|_\alpha, \|A\|_{2\alpha}) \cdot \frac{2}{\delta}C_1(\delta),\end{aligned}$$

be the polynomial $\|X_n(t) - X_n(r)\|_\infty \leq poly'(L_1, \|B\|_\alpha, \|\widetilde{A}\|_{2\alpha})|t-r|^\alpha$ for all n. □

Proof of Lemma A.4. The discretization $\hat{X}^n(\cdot)$ from Equation (2.4) on page 280 of [200] is defiend as:

$$\begin{aligned}\hat{X}_i^n(t_{k+1}^n) = & \hat{X}_i^n(t_k^n) + \mu_i(\hat{X}^n(t_k^n))\Delta t_n + \sum_{j=1}^{d'}\sigma_{ij}(\hat{X}^n(t_k^n))\Delta B_{j,k}^n \\ & + \sum_{j=1}^{d'}\sum_{l=1}^{d}\sum_{m=1}^{d'}\frac{\partial\sigma_{ij}}{\partial x_l}(\hat{X}^n(t_k^n))\sigma_{lm}(\hat{X}^n(t_k^n))\hat{A}_{mj}(t_k^n, t_{k+1}^n),\end{aligned}$$

where $\hat{A}_{i,j}(s,t) = 0$ for $i \neq j$ and $\hat{A}_{i,i}(s,t) = A_{i,i}(s,t)$ for $i \in [d]$ as in Definition A.2. Moreover,

it is defined on page 280 of [200], as in Definition A.2, that

$$R^n_{i,j}(t^n_l, t^n_m) \triangleq \sum_{k=l+1}^{m} \{A_{i,j}(t^n_{k-1}, t^n_k) - \hat{A}_{i,j}(t^n_{k-1}, t^n_k)\}$$

$$\Gamma_R \triangleq \sup_{\substack{n\geq 0 \\ 0\leq s\leq t\leq 1 \\ s,t\in D_n}} \max_{1\leq i,j\leq d'} \frac{|R^n_{i,j}(s,t)|}{|t-s|^\beta \Delta t_n^{2\alpha-\beta}}.$$

With a slight change in notation, we replace M with $L_1(\omega)$, $\|Z\|_\alpha$ with $\|B\|_\alpha$, then according to [200, Theorem 2.1], we can find constant G (for notation consistency with [200]) explicitly in terms of $L_1, K_\alpha, K_{2\alpha}$ and K_R such that $\|\hat{X}^n(t) - X_t\|_\infty \leq G\Delta t_n^{2\alpha-\beta}$ where we may take $K_\alpha = \|B\|_\alpha, K_{2\alpha} = \|A\|_{2\alpha}$ and $K_R = \Gamma_R + 1$. To prove a similar result for $\|X^\mu_n(t) - X_t\|_\infty$ instead of $\|\hat{X}^n(t) - X_t\|_\infty$, we replace Γ_R with our $\Gamma_{\widetilde{R}}$ defined in Definition A.2, the proof will follow exactly as in the proof of Theorem 2.1 in [200][Proposition 6.1 and 6.2]. Particularly, we are able to compute constant G in terms of $L_1(\omega), \|B\|_\alpha, \|A\|_{2\alpha}$ and $\Gamma_{\widetilde{R}}$ such that $\|X^\mu_n(t) - X_t\|_\infty \leq G\Delta t_n^{2\alpha-\beta}$, for n large enough. However, we can extend the result to hold for all n using the the method in the proof of Lemma A.3. Moreover, following Section 2.2 on pages 282−283 of [200] (part of which is shown in Lemma A.3), the construction of the constant G only involves multiplication and addition among the variables L_1,$\|B\|_\alpha$,$\|A\|_{2\alpha}$, $\Gamma_{\widetilde{R}}$ and constants. Thus there exists $poly''(\cdot) : \mathbb{R}^4 \to \mathbb{R}$ with $poly''(x,\cdot,\cdot,\cdot) > 1$ when $x > 1$

$$\|X^\mu_n(t) - X_t\|_\infty \leq poly''(L_1, \|B\|_\alpha, \|A\|_{2\alpha}, \Gamma_{\widetilde{R}})\Delta t_n^{2\alpha-\beta}.$$

□

Proof of Lemma 27. Given $\mu \cup \{\mu^{(n)}\}_n \subseteq \mathcal{L}(L_1), L_1 > 1$. Denote $\{X(t;\mu,B)\}_{t\in[0,1]}$ to be the solution of SDE under $\mu(\cdot)$ and Brownian motion B. Let $X_n(t;\mu^{(n+1)},B) \leftarrow$ Num_Sol but under $\mu^{(n+1)} \in \mathcal{L}_1$ instead of $\mu^{(n)}$. Since ΔB^n_k are the same for $B(\cdot)$ and $B^{(n+1),a}(\cdot)$ by Eqautions (2.6), we have $X_n(1;\mu^{(n+1)},B) = X_n(1;\mu^{(n+1)},B^{n+1,a})$.

Thus $\|X_{n+1}(1) - X^a_{n+1}(1)\|_\infty$ is bounded by

$$\begin{aligned}
&\|X_{n+1}(1) - X_n(1;\mu^{(n+1)}, B)\|_\infty + \|X^a_{n+1}(1) - X_n(1;\mu^{(n+1)}, B^{n+1,a})\|_\infty \\
\leq &\|X_{n+1}(1) - X(1;\mu^{(n+1)}, B)\|_\infty + \|X_n(1;\mu^{(n+1)}, B) - X(1;\mu^{(n+1)}, B)\|_\infty \\
&+ \|X^a_{n+1}(1) - X(1;\mu^{(n+1)}, B^{n+1,a})\|_\infty + \|X_n(1;\mu^{(n+1)}, B^{n+1,a}) - X(1;\mu^{(n+1)}, B^{n+1,a})\|_\infty \\
\leq &2\Big(poly''(L_1, \|B\|_\alpha, \|A\|_{2\alpha}, \Gamma_{\widetilde{R}}) + poly''(L_1, \|B^{n+1,a}\|_\alpha, \|A^{n+1,a}\|_{2\alpha}, \Gamma_{\widetilde{R}^{n+1,a}})\Big)\Delta t_n^{2\alpha-\beta}.
\end{aligned}$$

The last line follows from Lemma A.4 where quantity $\|B^{n+1,a}\|_\alpha, \|A^{n+1,a}\|_{2\alpha}, \Gamma_{\widetilde{R}^{n+1,a}}$ is defined for $B^{n+1,a}(\cdot)$ as for $B(\cdot)$ in Definition A.1. Now, raising above inequality to the eighth power and using A.2, there exists some $poly(\cdot)$ with $poly(x) > 1$ when $x > 1$

$$\mathbb{E}\|X_{n+1}(1) - X^a_{n+1}(1)\|^8_\infty \leq poly(L_1)\Delta t_n^{8(2\alpha-\beta)},$$

for all $n \geq 0$. □

Proof of Lemma A.1. By the Gaussian tail bound $\int_\xi^\infty e^{-\frac{t^2}{2}} dt \leq \frac{1}{\xi} e^{-\frac{\xi^2}{2}}$ for all $\xi > 0$,

$$\mathbb{P}(M_\epsilon > b) = 1 - \prod_{n=1}^{\infty} \mathbb{P}(|Z_n| \leq bn^\epsilon) \leq 1 - \prod_{n=1}^{\infty}(1 - \frac{2}{\sqrt{2\pi}\cdot bn^\epsilon} e^{-\frac{b^2 n^{2\epsilon}}{2}}).$$

Thus, we have $\mathbb{E}[e^{tM_\epsilon}] = \int_0^\infty \mathbb{P}(e^{tM_\epsilon} > b)db$ which is bounded by

$$3 + \int_3^\infty \mathbb{P}(M_\epsilon > \frac{\log(b)}{t})db \leq 3 + \int_3^\infty (1 - \prod_{n=1}^{\infty}(1 - \frac{2t}{\sqrt{2\pi}n^\epsilon \cdot \log(b)} e^{-\frac{\log(b)^2 n^{2\epsilon}}{2t^2}}))db < \infty,$$

according to calculation. □

Chapter 7: Unbiased Gradient Simulation for Stochastic Composition Optimization

We introduce unbiased gradient simulation algorithms for solving stochastic composition optimization (SCO) problems. We show that the unbiased gradients generated by our algorithms have finite variance and finite expected computational cost. Therefore, the unbiased gradients can be directly used to solve SCO problems by applying the Stochastic Gradient Descent method (SGD) and have an iteration complexity of $O(\epsilon^{-1})$ for strongly convex SCOs. We also show how to combine unbiased gradient simulation with variance reduction techniques such as stochastic variance reduced gradient (SVRG) or stochastically controlled stochastic gradient (SCSG) to achieve state-of-the-art theoretical convergence rates as well as practical performances. Finally, we illustrate the effectiveness of our algorithms through experiments on datasets arising from statistics and machine learning, specifically, Cox's partial likelihood model and conditional random field models.

7.1 Introduction

In statistics and machine learning, we often encounter the generic stochastic optimization problem

$$\min_{x\in\mathcal{D}} F(x) \triangleq \mathbb{E}_v f_v(x), \tag{7.1}$$

where f_v is a convex function indexed by random variable v, $\mathbb{E}_v$ denotes expectation with respect to v, and $\mathcal{D} \subset \mathbb{R}^d$ is a compact convex set. A special case of (7.1) is the empirical risk minimization

(ERM) problem when v is from the uniform random variable on $\{1, 2, \ldots, n\}$, that is,

$$\min_{x \in \mathcal{D}} F_n(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{7.2}$$

When obtaining the full gradient is computationally intensive, a popular method for solving these problems is the (projected) stochastic gradient descent (SGD) algorithm, which can be described by the following update rule for $t = 1, 2, ...$

$$x_t = \Pi_{\mathcal{D}}\{x_{t-1} - \lambda_t \nabla f_{v_t}(x_{t-1})\}, \tag{7.3}$$

where v_t is sampled from the distribution of v for generic optimization problems and from the uniform distribution on $\{1, 2, ..., n\}$ for ERM problems, λ_t is the step size, and $\Pi_{\mathcal{D}}$ is the projection operator on to $\mathcal{D}$. It is well known that convergence of SGD requires a diminishing step size λ_t and thus results in a worse convergence rate than gradient descent algorithms. [205] observed that the inferior rate of SGD is caused by the fact that stochastic gradients do not converge to 0 as the iterates converge to the optimal solution. Base on this observation, they improved the SGD by applying a control variate variance reduction technique to the stochastic gradient generation which is known as the SVRG algorithm. SVRG has been shown to converge linearly to the optimal solution for strongly convex ERM problems and performs well in practice. These algorithms implicitly assume that the gradient of each member function $f_v(\cdot)$ is easy to compute. But this assumption does not hold in the so-called stochastic composition optimization (SCO) problem [206]:

$$\min_{x \in \mathcal{D}} F(x) \triangleq \mathbb{E}_v f_v(\mathbb{E}_w g_w(x)),$$

where v and w are random variables with certain known joint distributions nor its finite sample version:

$$\min_{x\in\mathcal{D}} F_n(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i\{\frac{1}{m_i}\sum_{j=1}^{m_i} g_{ij}(x)\}. \tag{7.4}$$

Problems of this form arise in many areas such as reinforcement learning and risk-averse learning to graphical models, econometrics and survival analysis. As far as we know, all current algorithms that are used to solve SCO problems are based on *biased* stochastic gradient oracles. The convergence rates for these algorithms are unsatisfactory compared to the algorithms for solving generic stochastic optimization problems, except for the Comp-SVRG algorithms in [207]. Their algorithms are also based on *biased* stochastic gradients, but the modified variance reduced gradients vanish as the iterates converge to the optimal solution. Therefore, linear convergence can be proved for the finite sum version of SCO when strong convexity is present. However, the number of samples that are needed to construct a variance reduced gradient depends on the condition number of the objective function. All these drawbacks are the result of biased stochastic gradients. If unbiased stochastic gradients can be generated for SCO problems, we can treat SCO problems in the same way that we treat generic stochastic optimization problems and apply SGD and its variants to solve it.

7.1.1 Contributions

The contributions of this chapter can be summarized as follows.

- We introduce unbiased gradient simulation algorithms that are based on a multilevel Monte Carlo technique for solving smooth SCO problems. We also show that the output of these algorithms has finite variance and its expected computational cost is finite.

- Based on our unbiased gradient simulation algorithms, a stochastic composition optimization problem can be considered as a generic stochastic optimization problem. This is because we can simply apply SGD to solve SCO problems and achieve the same iteration complexity as

using SGD to solve generic stochastic optimization problems.

- We also show that our unbiased gradient simulation algorithm can be combined with variance reduction techniques including SVRG [205] and SCSG [208], yielding variance reduced optimization algorithms that converge linearly to the optimum of a SCO problem.

7.1.2 Related work

In the current SCO literature, as far as we know, all the algorithms used to solve SCO problems are based on *biased* stochastic gradients. [206] first proposed a generic algorithm for solving (7.5) with an iteration complexity of $O(\epsilon^{-3/2})$ for strongly convex objectives and $O(\epsilon^{-4})$ for general convex objectives. This result is further improved to $O(\epsilon^{-5/4})$ for strongly convex objectives and $O(\epsilon^{-7/2})$ for general convex objectives in [209]. For strongly convex objectives with finite sum structure, ([207]) modified the SVRG algorithm and achieved a sample complexity $O((m+n)\log(1/\epsilon))$. Stochastic algorithms using biased gradient methods also appeared in [210] for non-convex SCOs.

We propose unbiased gradient simulation methods that are based on a multilevel Monte Carlo technique for solving smooth SCO problems. Unbiased simulation methods for functions of expectations using multilevel Monte Carlo techniques were developed in [191] and [192]. Such techniques have been heavily used in simulation algorithms to solve problems that require high accuracy estimates such as stochastic differential equation [211, 187, 212], stochastic partial differential equations [213], and Markov Chains [214]. They also have been used to reduce computational cost through variance reduction techniques [188, 189, 179, 184].

We also consider variance reduced stochastic gradient algorithms that are based on unbiased gradient simulation. A number of variance reduction techniques have been proposed for strongly convex ERM problems in the literature including control variate see SVRG in ([205]) and SDCA in ([215]), incremental gradients in [216] and SAGA in [217], and importance sampling in [218]. The analysis of these methods and their variants can be find in [219, 220, 221, 222, 223, 224].

A summary of the *iteration complexity* for current algorithms on smooth SCO is presented in

Table 1. In particular, SimGD, SimVRG and SCSimG are proposed in this chapter. We report iteration complexity instead of sample complexity due to the special randomization component in the gradient estimator construction. This component is critical for our estimator to be unbiased, but the trade-off is the difficulty during the analysis of sample complexity. We will discuss the related issue into detail in later sections.

Table 7.1: Iteration complexity of different algorithms for solving smooth SCO problems.

	Convex	Strongly Convex
Basic SCGD [206]	$O(1/\epsilon^4)$	$O(1/\epsilon^{3/2})$
Accelerating SCGD [209]	$O(1/\epsilon^{7/2})$	$O(1/\epsilon^{5/4})$
Compositional SVRG-1 [207]	N.A.	$O(\log(1/\epsilon))$
Compositional SVRG-2 [207]	N.A.	$O(\log(1/\epsilon))$
SimGD (our variant of SGD)	$O(1/\epsilon^2)$	$O(1/\epsilon)$
SimVRG (our variant of SVRG)	N.A.	$O(\log(1/\epsilon))$
SCSimG (our variant of SCSG)	N.A.	$O(\log(1/\epsilon))$

The basic SCGD and accelerating SCGD makes 2 sampling queries in every iteration, Compositional SVRG-1 and Compositional SVRG-2 make $\sum_{i=1}^{n} m_i$ and additional constant number of sampling queries in every iteration. SimGD makes a random number of sampling queries in every iteration and the expectation of this random number is finite. SimSVRG makes $\sum_{i=1}^{n} m_i$ and additional random number of sampling queries in every iteration and the the expectation of this random number is finite. SCSimG makes $\sum_{i=1}^{n} m_i \wedge 1/\epsilon$ and additional random number of sampling queries in every iteration and the the expectation of this random number is finite.

7.1.3 Organization

The rest of the chapter is organized as follows. In section 2, we describe the problem formulations and introduce the notation that we will use. We then introduce our unbiased gradient simulation algorithms and the optimization algorithms that are based on these unbiased simulations. In section 3, we give concrete examples of SCO problems that arise in a variety of areas and explain how our algorithms are well-suited to solve them. In section 4, we prove several important

theoretical properties of our gradient simulation algorithm. In particular, its unbiasedness, finite variance and finite expected computational cost. We also show it has a certain "Lipschitz" property that makes it suitable for combining with variance reduction algorithms such as SVRG and SCSG. Finally, we prove the convergence properties of our algorithms. In section 5, we present numerical results obtained using our algorithms for maximizing Cox's partial likelihood and training conditional random fields.

7.2 Problem Description and Algorithms

7.2.1 Problem description and Notations

Throughout this chapter, we consider the following smooth stochastic composition optimization problem

$$\min_{x\in\mathcal{D}} F(x) \triangleq \mathbb{E}_v f_v(\mathbb{E}_w g_w(x)). \tag{7.5}$$

We define the support of the distributions v and w to be Ω_v and Ω_w. Note that the following two problems can be considered as special cases of (7.5); the first one is the finite sum problem:

$$\min_{x\in\mathcal{D}} F_n(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\frac{1}{m_i}\sum_{j=1}^{m_i} g_{ij}(x)), \tag{7.6}$$

and the second one is the mixed problem:

$$\min_{x\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbb{E}_w g_w(x)). \tag{7.7}$$

Later, we will discuss algorithms for these two special cases.

As for the notation, for a vector $v \in \mathbb{R}^n$, we use $[v]_i$ to denote the i-th entry for $1 \le i \le n$ and

use $\|v\|_p$ to denote its L_p-norm. For a matrix $A \in \mathbb{R}^{m\times n}$, we use $[A]_{ij}$, $[A]_{:j}$ and $[A]_{i:}$ to denote the (i,j)-th entry, j-th column and i-th row for every $1 \le i \le m$ and $1 \le j \le n$. We use $\|A\|_2$ and $\|A\|_F$ to denote its spectrum norm and Frobenius norm. We use $\|A\|_\infty$ to denote the maximum absolute value of the entries of A, that is, $\|A\|_\infty = \max\{|[A]_{ij}| \,|\, 1 \le i \le m, 1 \le j \le n\}$. For a multi-linear map $B \in \mathbb{R}^{m\times n\times p}$, we use $[B]_{ijk} \in \mathbb{R}$ to denote its (i,j,k)-th entry, use $[B]_{:jk} \in \mathbb{R}^m$,$[B]_{i:k} \in \mathbb{R}^n$, and $[B]_{ij:} \in \mathbb{R}^{1\times p}$ to denote its (j,k)-th column fiber,(i,k)-th row fiber, and (i,j)-th tube fiber, and use $[B]_{::k} \in \mathbb{R}^{m\times n}$, $[B]_{:j:} \in \mathbb{R}^{m\times p}$ and $[B]_{i::} \in \mathbb{R}^{n\times p}$ to denote its k-th frontal slice, j-th lateral slice and i-th horizontal slice, where $1 \le i \le m$, $1 \le j \le n$ and $1 \le k \le p$. We define $\|B\|_\infty = \{|[B]_{ijk}| \,|\, 1 \le i \le m, 1 \le j \le n, 1 \le k \le p\}$. Moreover, we use $\mathrm{vec}(\cdot)$ to denote the vectorize operation for one matrix or a multi-linear map. When there are multiple arguments in $\mathrm{vec}()$, it vectorize each component and stack them into another vector.

We write the Jacobian (with respect to x) of the vector valued $g_w(\cdot)$ as

$$\nabla g_w(x) = \begin{pmatrix} \frac{\partial [g_w]_1}{\partial [x]_1}(x) & \cdots & \frac{\partial [g_w]_1}{\partial [x]_p}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial [g_w]_d}{\partial [x]_1}(x) & \cdots & \frac{\partial [g_w]_d}{\partial [x]_p}(x) \end{pmatrix},$$

where

$$g_w(x) = ([g_w]_1(x), [g_w]_2(x), \ldots, [g_w]_d(x))^\top.$$

It then follows from the chain rule that the gradient (with respect of x) of $f_v(\cdot)$ for the stochastic problem is $\{\mathbb{E}_w \nabla g_w(x)\} \nabla f_v \{\mathbb{E}_w g_w(x)\}$ and

$$\nabla F(x) = \{\mathbb{E}_w \nabla g_w(x)\}^\top \mathbb{E}_v \{\nabla f_v(\mathbb{E}_w g_w(x))\}. \tag{7.8}$$

We use $\nabla^2 g_w(x) \in \mathbb{R}^{d\times p\times p}$ to denote the Hessian (with respect to x) of the vector valued $g_w(\cdot)$ and

use $\nabla^2 g_w(x)[u,v] \in \mathbb{R}^d$ to denote the vector that $\nabla^2 g_w(x)$ acting on $u, v \in \mathbb{R}^p$, that is,

$$[\nabla^2 g_w(x)[u,v]]_i = \sum_{j=1}^{p}\sum_{k=1}^{p}[\nabla^2 g_w(x)]_{ijk}[u]_j[v]_k = \sum_{j=1}^{p}\sum_{k=1}^{p}[\nabla^2[g_w]_i(x)]_{jk}[u]_j[v]_k.$$

Finally, we introduce the following notations used in our gradient simulation algorithms. Let $I_n(v_1) = \{w_i\}_{i=1}^n$ be a collection of random variables that are i.i.d. generated from the distribution of w given $v = v_1$, where v and w are the random variables in problem (7.5). Given the samples $I_n(v_1)$, let

$$\begin{aligned}
\overline{g}(x; n_1, n_2) &= \frac{1}{n_2 - n_1 + 1}\sum_{i=n_1}^{n_2} g_{w_i}(x), \\
\overline{\nabla g}(x; n_1, n_2) &= \frac{1}{n_2 - n_1 + 1}\sum_{i=n_1}^{n_2} \nabla g_{w_i}(x), \quad \text{and} \\
\overline{\nabla^2 g}(x; n_1, n_2) &= \frac{1}{n_2 - n_1 + 1}\sum_{i=n_1}^{n_2} \nabla^2 g_{w_i}(x),
\end{aligned}$$

for $x \in \mathcal{D} \subset \mathbb{R}^p$ and $1 \le n_1 \le n_2 \le n$. These quantities are *unbiased* estimates of $\mathbb{E}_w g_w(x)$, $\mathbb{E}_w \nabla g_w(x)$ and $\mathbb{E}_w \nabla^2 g_w(x)$. In addition, let

$$\bar{y}(x; n_1, n_2) = \overline{\nabla g}(x; n_0, n_1)^\top \nabla f_v(\bar{g}(x; n_1, n_2)),$$

which is the gradient of $f_{v_1}(\bar{g}(x; n_1, n_2))$. This is an estimate of $\nabla\{\mathbb{E}_v f_v(\mathbb{E}_w g_w(x))\}$ however, it is a biased estimate, that is,

$$\mathbb{E}\bar{y}(x; n_1, n_2) \neq \nabla\big(\mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}\big).$$

Since the samples are i.i.d., the expectation of $\bar{y}(x; n_1, n_2)$ only depends on the distribution of w condition on $v = v_1$, and the number of samples that are used to construct $\bar{y}(x; n_1, n_2)$. Then, we write

$$s(x; n_2 - n_1 + 1, v_1) = \mathbb{E}\{\bar{y}(x; n_1, n_2) | v = v_1\}.$$

We also let

$$[\bar{z}(x;n_1;n_2)]_i = \{[\overline{\nabla^2 g}(x;n_1,n_2)]_{::i}\}^\top \nabla f_{v_1}\{\bar{g}(x;n_1,n_2)\} + \{\overline{\nabla g}(x;n_1,n_2)\}^\top \nabla^2 f_{v_1}\{\bar{g}(x;n_1,n_2)\}[\overline{\nabla g}(x;n_1,n_2)]_{:i},$$

which is the i-th row of the Hessian of $f_{v_1}(\bar{g}(x;n_1,n_2))$ for $1 \le i \le p$. Similarly, it is also a *biased* estimate of $\nabla^2\big(\mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}\big)$.

7.2.2 Unbiased stochastic gradient simulation

We first present Algorithm 1 to simulate unbiased gradients for the stochastic problems (7.5) and (7.7) while fixing a component v_1 for $f_{v_1}(\mathbb{E}_w g_w(x))$. It can be considered as a variant of [192] based on a multilevel randomization technique.

Algorithm 9 UnbiasedGradient(x, v_1, n_0, γ)

procedure UNBIASEDGRADIENT$(x, v_1, n_0, \gamma)(.)$

Input: $x \in \mathcal{D}$,$v_1 \in \Omega_v$, base level $n_0 \ge 0 \in \mathbb{Z}$, rate parameter $1 < \gamma < 2$.

Output: $G(x, v_1) \in \mathbb{R}^p$, an unbiased estimate of the gradient of $f_{v_1}(\mathbb{E}_w g_w(x))$ at point x and component v_1.

Sample N from a geometric distribution with success probability $1 - p$ where $p = 0.5^\gamma$.

Independently sample $I_{2^{n_0+N+1}}(v_1) = \{w_i\}_{i=1}^{2^{n_0+N+1}}$ from the distribution of w given v_1.

Compute $Y_1(x) = \bar{y}(x; 1, 2^{n_0+N+1})$.

Compute $Y_2(x) = \bar{y}(x; 1, 2^{n_0+N})$.

Compute $Y_3(x) = \bar{y}(x; 2^{n_0+N} + 1, 2^{n_0+N+1})$.

Compute $Y_4(x) = \bar{y}(x; 1, 2^{n_0})$.

Compute $G(x, v_1) = \frac{Y_1(x) - 0.5(Y_2(x)+Y_3(x))}{\tilde{p}_N} + Y_4(x)$, where $\tilde{p}_N = (1-p)p^N$.

Output: $G(x, v_1)$

We shall prove in Section 4 that the output of Algorithm 1 is indeed an unbiased estimate of $f_{v_1}(\mathbb{E}_w g_w(x))$ for fixed v_1. It follows that if we sample $v_1 \sim v$, then $G(x, v_1)$ would be an unibased

estimate of the gradient of $\mathbb{E}_v f_v(\mathbb{E}_w g_w(x))$.

Remark: We note that Algorithm 1 requires conditional sampling of w given v. It is difficult to obtain such samples in a very general setting. However, in many applications, obtaining such samples can be relatively easy. We will discuss this in detail in Section 3. Moreover, Algorithm 1 uses a random number of samples to construct an unbiased estimate. We will show later that the number of samples needed is finite in expectation and free of the problem sample size. However, for problems such as (7.6), computing an unbiased estimate using this algorithm may need the same number of samples as computing the true gradient in a worst case scenario.

7.2.3 Optimization Algorithms

We now present our optimization algorithms to solve problem (7.5), (7.7) and (7.6) based on unbiased gradient simulation. First, in Algorithm 10, we present our SGD (SimGD) algorithm with a simple averaging techinique (see [225]). Convergence of our SimGD algorithm under different conditions will be analyzed in Section 4. It is worth noting that our SGD algorithm is an analogue of the standard stochastic gradient descent algorithm that substitutes simulated unbiased gradients for sampled stochastic gradients. Therefore, the unbiased gradient simulation algorithm enables us to solve SCO problems in the same way as generic stochastic optimization problems.

Algorithm 10 Simulated Gradient Descent (SimGD)

Input: Number of iterations T, step size $\{\lambda_t\}_{t=1}^{\infty}$, initial point x_0, base level n_0 and rate parameter $1 < \gamma < 2$.

for $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample v_t follows the distribution of v and let $\rho_t = \text{UnbiasedGradient}(x_t, v_t, n_0, \gamma)$
  $x_{t+1} = \Pi_{\mathcal{D}}(x_t - \lambda_t \rho_t)$

option I Output $\tilde{x}_T = \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} (t+1) x_t$

option II Output x_T

In contrast to SGD, where a diminishing step size is used, we also introduce an SVRG type of control variate variance reduced algorithm as mentioned in [205] with constant step size for SCO problems. As described in [205] for ERM problems (7.2) and in [226] for generic stochastic

optimization problems (7.1), a variance reduced stochastic gradient at point x with respect to the reference point $\tilde{x}$ is defined as $\nabla f_{v'}(x) - \nabla f_{v'}(\tilde{x}) + \nabla F(\tilde{x})$ where v' is sampled from v for the generic stochastic optimization problem (7.1) and defined similarly for the ERM problem. We adopt these variance reduction techniques in our setting of unbiased gradient simulation. **Specifically, we will simulate the unbiased gradients at x and $\tilde{x}$ simultaneously, using the same set of simulated data, to reduce variance.** The details of generating such variance reduced gradients are specified in Algorithm 11. For ease of presentation, Algorithm 11 is built on the setting of Algorithm 1 and it can be modified by using Algorithm 2 for solving problem (7.6).

Algorithm 11 SimulatedGradient(x, $\tilde{x}$, $G(\tilde{x})$, v_1, n_0, γ)

procedure SIMULATEDGRADIENT(x, $\tilde{x}$, $G(\tilde{x})$, v_1, n_0, γ)**Input:** $x \in \mathbb{R}^d$, $v_1 \in \Omega_v$, reference point $\tilde{x} \in \mathbb{R}^d$, an estimate of gradient at point $\tilde{x}$ $\hat{G}(\tilde{x}) \in \mathbb{R}^p$, base level $n_0 \geq 0$ and rate parameter $1 < \gamma < 2$.

Output: $W \in \mathbb{R}^p$, a variance reduced unbiased estimator of the gradient of $\mathbb{E}_v f(\mathbb{E}_w g_w(x), v)$ at point x.

Sample N from a geometric distribution with success rate $1 - p$ where $p = 0.5^\gamma$.

Compute $\tilde{p}_N = (1-p)p^N$.

Independently sample $I_{2^{n_0+N+1}}(v_1) = \{w_i\}_{i=1}^{2^{n_0+N+1}}$ from the conditional distribution of w given $v = v_1$.

Compute $Y_1(x) = \bar{y}(x; 1, 2^{n_0+N+1})$ and $Y_1(\tilde{x}) = \bar{y}(\tilde{x}; 1, 2^{n_0+N+1})$.

Compute $Y_2(x) = \bar{y}(x; 1, 2^{n_0+N})$ and $Y_2(\tilde{x}) = \bar{y}(\tilde{x}; 1, 2^{n_0+N})$.

Compute $Y_3(x) = \bar{y}(x; 2^{n_0+N} + 1, 2^{n_0+N+1})$ and $Y_3(\tilde{x}) = \bar{y}(\tilde{x}; 2^{n_0+N} + 1, 2^{n_0+N+1})$.

Compute $Y_4(x) = \bar{y}(x; 1, 2^{n_0})$ and $Y_4(\tilde{x}) = \bar{y}(\tilde{x}; 1, 2^{n_0})$.

Compute $W(x, v_1) = \frac{Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\}}{\tilde{p}_N} + Y_4(x)$.

Compute $W(\tilde{x}, v_1) = \frac{Y_1(\tilde{x}) - 0.5\{Y_2(\tilde{x}) + Y_3(\tilde{x})\}}{\tilde{p}_N} + Y_4(\tilde{x})$.

Set $W(x, \tilde{x}, v_1) = W(x, v_1) - W(\tilde{x}, v_1) + \hat{G}(\tilde{x})$.

Output: $W(x, \tilde{x}, v_1)$.

In Algorithm 11, the reference gradient $G(\tilde{x})$ can either be the full gradient at $\nabla F(\tilde{x})$ or an estimate of the full gradient $\nabla F(\tilde{x})$. For example, when it is efficient to compute full gradients of the objective function for problem (7.5) and (7.7), we propose to use the following method in Algorithm 5 to solve this problem. It can be considered as a variant of SVRG, we thus denote it by Simulated Variance Reduced Gradient Descent.

However, when the full gradients $\nabla F(\tilde{x})$ of the objective function (7.5) can be difficult to

Algorithm 12 Simulated Variance Reduced Gradient Descent(SimVRG)

Inputs: Number of epochs T, number of steps in each epoch M, step size λ and initial point $\tilde{x}_0$, base level $n_0 \geq 0$, and parameter $1 < \gamma < 2$.
for $s = 0, 1, 2, \ldots, T-1$ **do**
  Compute the full gradient $\nabla F(\tilde{x}_s)$
  $x_0 = \tilde{x}_s$
  for $t = 0, 1, 2, \ldots, M-1$ **do**
    Sample v_t from the distribution of v.
    Compute $\rho_t = \text{SimulatedGradient}(x_t, \tilde{x}_s, \hat{G}(x_s), v_t, n_0, \gamma)$.
    Update $x_{t+1} = \Pi_{\mathcal{D}}(x_t - \lambda \rho_t)$.
  option I Output $\tilde{x}_{s+1} = x_M$
  option II Output $\tilde{x}_{s+1} = x_t$ for randomly chosen $t \in \{1, ..., M\}$

compute, we estimate the full gradient $\nabla F(\tilde{x})$ by sampling the unbiased gradient within a batch of the indices and taking the average. This method is related to another variant of SVRG, namely SCGS in [222] and we summarize the details of this approach in Algorithm 13. Convergence properties of Algorithm 12 and Algorithm 13 will be analyzed in Section 4.

Algorithm 13 Stochastically Controlled Simulated Gradient Descent(SCSimG)

Inputs: Number of epochs T, number of steps in each epoch M, batch size B, sample size K, step size λ, initial point $\tilde{x}_0$, base level $n_0 \geq 0$ and parameter $1 < \gamma < 2$.
for $s = 0, 1, \ldots, T-1$ **do**
  $x_0 = \tilde{x}_s$
  Uniformly sample a batch $\mathcal{I}_s \subset \Omega_v$ according to the distribution of v with $|\mathcal{I}_s| = B$
  for $k = 1, 2, \ldots, K$ **do**
    Compute $h_k(\tilde{x}_s) = \frac{1}{B} \sum_{v_i \in \mathcal{I}_s} \text{UnbiasedGradient}(\tilde{x}_s, v_i, n_0, \gamma)$
  Compute $\tilde{h}(\tilde{x}_s) = \frac{1}{K} \sum_{i=1}^{K} h_i(\tilde{x}_s)$
  for $t = 0, 1, \ldots, M-1$ **do**
    Sample v_t from the distribution of v.
    Set $\rho_t = \text{SimulatedGradient}(x_t, \tilde{x}_s, \tilde{h}(\tilde{x}_s), v_t, n_0, \gamma)$.
    Update $x_{t+1} = \Pi_{\mathcal{D}}(x_t - \lambda \rho_t)$.
  option I Output $\tilde{x}_s = x_M$
  option II Output $\tilde{x}_s = x_t$ for randomly chosen $t \in \{1, ..., M\}$

7.3 Examples

We now present some important examples that can be formulated as SCO problems.

7.3.1 Conditional Random Fields (CRF)

Conditional random fields (CRF) [227] is a popular probabilistic model used for structural prediction. It has been used in a number of natural language processing (NLP) problems including part-of-speech tagging [227], noun-phrase chunking [228, 229] named identity recognition [230] and image segmentation in computer vision [231]. In the CRF models, the conditional probability of a structured outcome $y \in \mathcal{Y}$ given an observation $x \in \mathcal{X}$ is:

$$p(y \mid z; x) = \frac{\exp\{x^\top F(z, y)\}}{\sum_{y' \in \mathcal{Y}} \exp\{x^\top F(z, y')\}}, \tag{7.9}$$

where $x \in \mathbb{R}^p$ is the parameter for estimation and $F(z, y) \in \mathbb{R}^p$ is a vector of pre-specified feature functions depending on the underlying structure of $\mathcal{Y}$. Based on the set of training data $\{(z_i, y_i), i = 1, \ldots, n\}$, the parameter x can be estimated by maximizing the log likelihood function

$$\max_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid z_i, x). \tag{7.10}$$

As we shall see, the practical difficulty of computing the objective function value or its gradient lies in the exponential cardinality of $\mathcal{Y}$. The hardness of computing log-likelihood and gradients for CRFs has been considered in [232] and [233]. When the underlying structure of $\mathcal{Y}$ is a linear chain or a tree, both the objective function value and the gradient can be efficiently computed through dynamic programming (the Viterbi algorithm in [234]). For these structural cases, a number of methods can be used to solve (7.10); for example, deterministic methods such as the iterative scaling algorithm in [227] , L-BFGS in [229], stochastic methods such as stochastic gradient descent in [235] and SAG in [236]. However, when the underlying structure is more general (no linear chain or tree structure), computing a full gradient or even a stochastic gradient for problem (7.10) is difficult due to the exponential cardinality of $\mathcal{Y}$. In our setting, we can formulate (7.10) as a

composition optimization problem as in (7.5) by noticing that (7.10) is equivalent to

$$\min_x \frac{1}{n}\sum_{i=1}^{n}\big(\log\big[\sum_{y'\in\mathcal{Y}}\exp\{x^\top F(z_i,y')\}\big]-x^\top F(z_i,y_i),\big) \tag{7.11}$$

whose gradient can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{y'\in\mathcal{Y}}\exp\{x^\top F(z_i,y')\}F(z_i,y')}{\sum_{y'\in\mathcal{Y}}\exp\{x^\top F(z_i,y')\}}-F(z_i,y_i).$$

Note that this problem is equivalent to

$$\min_x \frac{1}{n}\sum_{i=1}^{n}\big(\log\big[\frac{1}{|\mathcal{Y}|}\sum_{y'\in\mathcal{Y}}\exp\{x^\top F(z_i,y')\}\big]-x^\top F(z_i,y_i)+\log|\mathcal{Y}|\big).$$

Therefore we can view it as a form of problem (7.5) and apply our optimization algorithms to solve (7.11).

To obtain a sample y' uniformly from $\mathcal{Y}$, we first let (V,E) be the underlying graph of the CRF. We assume that each vertex $v\in V$ takes value from $\{1,2,\ldots,K\}$. Under this setting, we can generate a discrete uniform random number over $\{1,2,\ldots,K\}$ for each vertex, and hence repeat this $|V|$ times to obtain a sample y' uniformly, where $|V|$ is the cardinality of V. This sampling scheme avoids sampling y' from a set of cardinality $K^{|V|}$ directly.

7.3.2 Softmax optimization

The Softmax optimization problems naturally arise when applying maximum likelihood estimation to the multinomial logistic model with application in many fields such as economics [237] and and network flows [238]. Specifically, the multinomial logistic model assumes the conditional probability mass of a discrete response $Y\in\{1,\ldots,K\}$ given covariates $X\in\mathbb{R}^p$ and parameters

$\beta = [\beta_1, \dots \beta_K] \in \mathbb{R}^{p \times K}$ satisfies

$$\mathbb{P}(Y = k|X, \beta) = \frac{\exp(x^\top \beta_k)}{\sum_{i=1}^K \exp(X^\top \beta_i)}.$$

Given n observations (X_i, Y_i), the log-likelihood function can be written as

$$l(\beta) = \sum_{i=1}^n \{X_i^\top \beta_{Y_i} - \log\{\sum_{j=1}^K \exp(X_j^\top \beta_j)\}\}.$$

Therefore, maximizing the log-likelihood function, which is known as the Softmax optimization problem, can be viewed as a compositional optimization problem, where the β here corresponds to the x in problem (7.6). To obtain a sample w_i in Algorithm 1 for this problem, we only need to generate a discrete uniform random variable over $\{1, \dots, K\}$.

7.3.3 Cox's partial likelihood

The Cox's partial likelihood model [239, 240] is a widely used in survival analysis for censored data. It belongs to a class of survival models in statistics called the propositional harzard models in [241]. In particular, the Cox's model assumes there is a hazard function for an observation with covariates $X \in \mathbb{R}^p$ and coefficient $\beta \in \mathbb{R}^p$ as:

$$\lambda(t|X) = \lambda_0(t) \exp(\beta^\top X),$$

where $\lambda_0(t)$ is the baseline hazard function. In Cox's model, for each data point, we have two variables T_i denoting the true life time and C_i denoting the censoring time independent of T_i which are not observed. Instead, we can only observe $(X_i, Y_i, \Delta_i)_{1 \le i \le n}$ assumed to be I.I.D. observations, where $X_i \in \mathbb{R}^p$ are the covariates, $Y_i \in \mathbb{R}$ are the observed times determined by $Y_i = \min(T_i, C_i)$, and $\Delta_i = \mathbb{I}\{Y_i = T_i\}$ are the indications for the censoring. Moreover, for a particular observation i, we define its risk set as the index set $\{j : Y_j \ge Y_i\}$. The Cox's model aims to maximize the partial

likelihood function as follows:

$$\max_{\beta\in\mathbb{R}^p} -\frac{1}{n}\sum_{i=1}^{n}\Delta_i[-X_i^\top\beta+\log\{\sum_{j=1}^{n}\mathbb{I}(Y_j\geq Y_i)\exp(X_j^\top\beta)\}], \tag{7.12}$$

which is equivalent to

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\Delta_i[-X_i^\top\beta+\log\{\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(Y_j\geq Y_i)\exp(X_j^\top\beta)\}],$$

whose gradient can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i[-X_i+\frac{\sum_{j=1}^{n}\mathbb{I}(Y_j\geq Y_i)\exp(X_j^\top\beta)X_j}{\sum_{j=1}^{n}\mathbb{I}(Y_j\geq Y_i)\exp(X_i^\top\beta)}]. \tag{7.13}$$

This problem as a form of (7.5) hence we can apply the proposed algorithms to solve it.

7.4 Theory

In this section we present the analysis of our algorithms applied to problem (7.5), that is, $\min_{x\in\mathcal{D}} F(x)\triangleq\mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}$. We omit the case for (7.6) and (7.7) as they can be analyzed similarly. We first give our assumptions.

7.4.1 Definitions, Assumptions and Lemmas

Assumption 1 In the compact set $\mathcal{D}$, each $f_v(\cdot)$ in the objective function of (7.5) is three times continuously differentiable. Its first order derivative is Lipschitz continuous with constant $L_{f,1}$, its second order derivative is Lipschitz continuous with constant $L_{f,2}$, and its third order derivative is Lipschitz continuous with constant $L_{f,3}$.

Assumption 2 In the compact set $\mathcal{D}$, each $g_w(\cdot)$ is twice continously differentiable. Its first order derivative is Lipschitz continuous with constant $L_{g,1}$ and its second order derivative is Lipschitz continuous with constant $L_{g,2}$.

Assumption 3 We assume $F(\cdot)$ in (7.5) is strongly convex with parameter μ and its gradient is Lipstchitz continuous with constant L.

Definition 2. *Define* $\mathcal{G} = \{y \in \mathbb{R}^d \mid y = g_w(x), x \in \mathcal{D}, w \in \Omega_w\}$ $\mathcal{H} = \{y \in \mathbb{R}^{d\times p} \mid y = \nabla g_w(x), x \in \mathcal{D}, w \in \Omega_w\}$ *and* $\mathcal{J} = \{z \in \mathbb{R}^{d\times p\times p} \mid z = \nabla^2 g_w(x), x \in \mathcal{D}, w \in \Omega_w\}$.

Assumption 4 We assume that $l_{g,0} = \sup\{\|y\|_\infty \mid y \in \mathcal{G} \subset \mathbb{R}^d\} < \infty$, $l_{g,1} = \sup\{\|y\|_\infty \mid y \in \mathcal{H} \subset \mathbb{R}^{d\times p}\} < \infty$, and $l_{g,2} = \sup\{\|z\|_\infty \mid z \in \mathcal{J} \subset \mathbb{R}^{d\times p\times p}\}$.

Assumption 5 We assume that $l_{f,0} = \sup\{|y| \mid y = f_v(x), x \in \mathcal{G}, v \in \Omega_v\} < \infty$, $l_{f,1} = \sup\{\|y\|_\infty \mid y = \nabla f_v(x), x \in \mathcal{G}, v \in \Omega_v\} < \infty$, $l_{f,2} = \sup\{\|y\|_\infty \mid y = \nabla^2 f_v(x), x \in \mathcal{G}, v \in \Omega_v\} < \infty$, and $l_{f,3} = \sup\{\|y\|_\infty \mid y = \nabla^3 f_v(x), x \in \mathcal{G}, v \in \Omega_v\} < \infty$.

Before we proceed, we state two elementary lemmas used in our proofs.

Lemma 30. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a continuously differentiable function with L-Lipschitz continuous gradients, then*

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \le \frac{L}{2}\|y - x\|_2^2.$$

We omit the proof of Lemma 30 since it is a well known result.

Lemma 31. *Given a positive integer N and a sequence of real number* a_i, $1 \le i \le N$, *we have, for all* $p \ge 1$, *that*

$$|\sum_{i=1}^{N} a_i|^p \le N^{p-1} \sum_{i=1}^{N} |a_i|^p, \tag{7.14}$$

Proof. Proof. This is a consequence of Jensen's inequality. □

7.4.2 Properties of the Unbiased Gradient Simulation Algorithm

In this subsection, we analysis the properties of Algorithm 1. We first prove the unbiasedness of $G(x, v_1)$.

Proposition 1 (Unbiasedness) For any $x \in \mathcal{D}$, sample $v_1 \sim v$, $G(x, v_1)$ is an unbiased estimate of $\nabla \mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}$, that is, $\mathbb{E}G(x, v_1) = \nabla \mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}$.

Proof. Proof of Proposition 1. Fix v_1 and $x \in \mathcal{D}$. We first show that the output $G(x, v_1)$ is an unbiased estimate of $\nabla f_{v_1}\{\mathbb{E}_w g_w(x)\}$. According to Algorithm 1, we have,

$$\begin{aligned}
\mathbb{E}G(x, v_1) &= \sum_{n=0}^{\infty} \mathbb{E}\{G(x, v_1)|N = n\}\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \frac{\mathbb{E}\{Y_1(x) - 0.5(Y_2(x) + Y_3(x))|N = n\}}{\tilde{p}_n}\tilde{p}_n + \mathbb{E}Y_4(x) \\
&= \sum_{n=0}^{\infty} \mathbb{E}\{Y_1(x) - 0.5(Y_2(x) + Y_3(x))|N = n\} + \mathbb{E}Y_4(x).
\end{aligned}$$

Note that condition on $N = n$, we assume there is hypothetically a set of i.i.d. samples $I_{2^{n_0+n+1}}(v_1) = \{w_i\}_{i=1}^{2^{n+n_0+1}}$ that follows the distribution of w given $v = v_1$ that $Y_1(x)$, $Y_2(x)$ and $Y_3(x)$ are constructed. Therefore

$$\begin{aligned}
\mathbb{E}\{Y_2(x)|N = n\} &= \mathbb{E}\{\bar{y}(x, 1, 2^{n_0+n})\} = s(x; 2^{n+n_0}) \\
&= \mathbb{E}\{\bar{y}(x; 2^{n_0+n} + 1, 2^{n_0+n+1})\} = \mathbb{E}\{Y_3(x)|N = n\}.
\end{aligned}$$

And $\mathbb{E}Y_4(x) = s(x; 2^{n_0}, v_1)$ and $\mathbb{E}\{Y_1(x)|N = n\} = s(x; 2^{n_0+n+1}, v_1)$. Therefore,

$$
\begin{aligned}
&\mathbb{E}G(x, v_1) \\
&= \sum_{n=0}^{\infty} \big(s(x; 2^{n_0+n+1}, v_1) - 0.5\{s(x; 2^{n_0+n}, v_1) + s(x; 2^{n_0+n}, v_1)\}\big) + s(x; 2^{n_0}, v_1) \\
&= \sum_{n=0}^{\infty} \{s(x; 2^{n_0+n+1}, v_1) - s(x; 2^{n_0+n}, v_1)\} + s(x; 2^{n_0}, v_1).
\end{aligned}
$$

Note that the above sum is a telescoping sum, therefore

$$
\begin{aligned}
\mathbb{E}G(x, v_1) &= \lim_{n\to\infty} s(x; 2^{n_0+n}, v_1) - s(x; 2^{n_0}, v_1) + s(x; 2^{n_0}, v_1) \\
&= \lim_{n\to\infty} s(x; 2^{n_0+n}, v_1) = \lim_{n\to\infty} \mathbb{E}\bar{y}(x; 1, 2^{n_0+n}, v_1) \\
&= \lim_{n\to\infty} \mathbb{E}\big(\overline{\nabla g}(x; 1, 2^{n_0+n})^\top \nabla f_{v_1}\{\bar{g}(x; 1, 2^{n_0+n})\}\big) \\
&= \lim_{n\to\infty} \mathbb{E}\big(\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}^\top f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\big).
\end{aligned}
$$

Since

$$
\begin{aligned}
&\|\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}^\top f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\|_2 \\
&\le \|\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}\|_F \|f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\|_2 \\
&\le \big(\sqrt{pd}\|\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}\|_\infty\big)\big(\sqrt{d}\|f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\|_\infty\big) \\
&\le \sqrt{p} d l_{g,0} l_{g,1},
\end{aligned}
$$

where the last inequality utilizes Assumption 4 and Assumption 5. Consequently, by the bounded convergence theorem, we can exchange the expectation and limit and hence

$$
\mathbb{E}G(x, v_1) = \mathbb{E} \lim_{n\to\infty} \big(\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}^\top f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\big).
$$

By continuity of $\nabla f_{v_1}(\cdot)$, we have

$$\lim_{n\to\infty} \nabla f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\} = \nabla f_{v_1}\{\lim_{n\to\infty} \frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}.$$

Since the samples are i.i.d., by the strong law of large numbers, we have

$$\lim_{n\to\infty} \frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x) = \mathbb{E}_w g_w(x) \text{ almost surely.}$$

By a similar argument,

$$\lim_{n\to\infty} \frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x) = \mathbb{E}_w \nabla g_w(x) \text{ almost surely.}$$

Therefore

$$\begin{aligned}
\mathbb{E}G(x, v_1) &= \mathbb{E} \lim_{n\to\infty} \big(\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} \nabla g_{w_i}(x)\}^\top f_{v_1}\{\frac{1}{2^{n_0+n}} \sum_{i=1}^{2^{n_0+n}} g_{w_i}(x)\}\big) \\
&= \mathbb{E}\big(\{\mathbb{E}_w \nabla g_w(x)\}^\top \nabla f_{v_1}\{\mathbb{E}_w g_w(x)\}\big) \\
&= \{\mathbb{E}_w \nabla g_w(x)\}^\top \nabla f_{v_1}\{\mathbb{E} g_w(x)\} = \nabla\{f_{v_1}(\mathbb{E}_w g_w(x))\}.
\end{aligned}$$

Finally, taking expectation w.r.t v_1, we obtain that

$$\mathbb{E}G(x, v_1) = \mathbb{E}_v \nabla(f_v\{\mathbb{E}_w g_w(x)\}) = \nabla \mathbb{E}_v f_v\{\mathbb{E}_w g_w(x)\}.$$

□

Next, we will state two ancillary lemmas that will be used in proving the finite variance of $G(x, v_1)$. Proof of these two lemmas can be found in the Supplementary.

Lemma 32. *For every $s \in \mathcal{H} \subset R^{d\times p}$, $t \in \mathcal{G} \subset R^d$, and $v_1 \in \Omega_v$, define $H : \mathcal{H} \times \mathcal{G} \to \mathbb{R}^p$ by $H(s,t) = s^\top \nabla f_{v_1}(t)$. Then every component function of $H(s,t)$ has a Lipschitz continuous*

gradient with constant $L_H = \sqrt{L_{f,1}^2 + 2dl_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2}$*, i.e., for every* $1 \le i \le p$*, we have*

$$\|\nabla[H]_i(s_1, t_1) - \nabla[H]_i(s_2, t_2)\|_F \le L_H \|vec([s_1]_{:i}, t_1) - vec([s_2]_{:i}, t_2)\|_2.$$

Lemma 33. *For every* $s, s_0 \in \mathcal{H} \subset \mathbb{R}^{d\times p}$ *and* $t, t_0 \in \mathcal{G} \subset \mathbb{R}^p$*, define*

$$R(s, s_0, t, t_0) = H(s, t) - H(s_0, t_0) - \nabla H(s_0, s_0)[s - s_0, t - t_0].$$

Then we have

$$\|R(s, s_0, t, t_0)\| \le \frac{L_H}{2}(\|s - s_0\|_F^2 + p\|t - t_0\|_2^2).$$

Proposition 2 (Finite second order moment) Fix any $x \in \mathcal{D}$ and $v_1 \in \Omega_v$, we have

$$\mathbb{E}\|G(x, v_1)\|_2^2 \le C'_{\mathcal{D}},$$

where

$$C'_{\mathcal{D}} = 2pd^2 l_{g,1}^2 l_{f,1}^2 + \frac{108p^2d^2(L_{f,1}^2 + 2df_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2)(l_{g,0}^4 + l_{g,1}^4)}{4^{n_0}(1 - 0.5^{\gamma})(1 - 0.5^{2-\gamma})}$$

and $1 < \gamma < 2$ is from the unbiased gradient simulation algorithm. Therefore $G(x, v_1)$ has finite variance.

Proof. Proof of Proposition 2. First, by (7.14),

$$\begin{aligned}\|G(x, v_1)\|_2^2 &= \|\frac{\big(Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\}\big)}{\tilde{p}_N} + Y_4(x)\|_2^2 \\ &\le 2\|\frac{\big(Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\}\big)}{\tilde{p}_N}\|_2^2 + 2\|Y_4(x)\|_2^2.\end{aligned}$$

To obtain an upper bound of $\mathbb{E}\|G(x, v_1)\|_2^2$, we first take expectation with respect to N, therefore

$$\begin{aligned}
&\mathbb{E}\|G(x, v_1)\|_2^2 \\
&\leq 2\sum_{n=0}^{\infty} \mathbb{E}\Big(\frac{\|Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\}\|_2^2}{\tilde{p}_n^2}|N = n\Big)\mathbb{P}(N = n) + 2\mathbb{E}\|Y_4(x)\|_2^2 \\
&\leq 2\sum_{n=0}^{\infty} \frac{\mathbb{E}\big(\|Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\}\|_2^2|N = n\big)}{\tilde{p}_n} + 2\mathbb{E}\|Y_4(x)\|_2^2. \qquad (7.15)
\end{aligned}$$

To proceed with equation (7.15), we first upper bound $\|Y_4(x)\|_2^2$ by

$$\begin{aligned}
\|Y_4(x)\|_2^2 &= \|\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} \nabla g_{w_i}(x)\}^\top \nabla f_{v_1}\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} g_{w_i}(x)\}\|_2^2 \\
&\leq \|\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} \nabla g_{w_i}(x)\|_2^2 \, \|\nabla f_{v_1}\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} g_{w_i}(x)\}\|_2^2 \\
&\leq \|\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} \nabla g_{w_i}(x)\|_F^2 \, \|\nabla f_{v_1}\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} g_{w_i}(x)\}\|_2^2.
\end{aligned}$$

Note that by Assumption 4 and 5,

$$\|\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} \nabla g_{w_i}(x)\|_F \leq \sqrt{pd}\|\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} \nabla g_{w_i}(x)\|_\infty \leq \sqrt{pd}l_{g,1}, \text{ and}$$

$$\|\nabla f_{v_1}\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} g_{w_i}(x)\}\|_2 \leq \sqrt{d}\|\nabla f_{v_1}\{\frac{1}{2^{n_0}}\sum_{i=1}^{2^{n_0}} g_{w_i}(x)\}\|_\infty \leq \sqrt{d}l_{f,1}.$$

Therefore

$$\mathbb{E}\|Y_4(x)\|_2^2 \leq pd^2 l_{g,1}^2 l_{f,1}^2. \qquad (7.16)$$

To bound the second term on the right hand side of (7.15), we first define the following vector-valued function: for $s \in \mathcal{H} \subseteq \mathbb{R}^{d\times p}$ and $t \in \mathcal{G} \subseteq \mathbb{R}^d$, define $H : \mathcal{H} \times \mathcal{G} \to \mathbb{R}^p$ by $H(s, t) \triangleq s^\top \nabla f_{v_1}(t)$. Moreover, to simplify the notation, let $\bar{n}_0 = n_0 + n$ and $\bar{n}_0^+ = n_0 + n + 1$. Therefore given

that $N = n$, we can write

$$
\begin{aligned}
& Y_1(x) - 0.5\{Y_2(x) + Y_3(x)\} \\
&= \bar{y}(x; 1, 2^{\bar{n}_0^+}) - 0.5\{\bar{y}(x; 1, 2^{\bar{n}_0}) + \bar{y}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+})\} \\
&= H\{\overline{\nabla g}(x; 1, 2^{\bar{n}_0^+}), \bar{g}(x; 1, 2^{\bar{n}_0^+})\} - 0.5H\{\overline{\nabla g}(x; 1, 2^{\bar{n}_0}), \bar{g}(x; 1, 2^{\bar{n}_0})\} \\
&- 0.5H\{\overline{\nabla g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \bar{g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}). \qquad (7.17)
\end{aligned}
$$

Since $\bar{g}(x; 1, 2^{\bar{n}_0^+}) = 0.5\{\bar{g}(x; 1, 2^{\bar{n}_0}) + \bar{g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+})\}$, and $\overline{\nabla g}(x; 1, 2^{\bar{n}_0^+}) = 0.5\{\overline{\nabla g}(x; 1, 2^{\bar{n}_0}) + \overline{\nabla g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+})\}$, when expanding the three functions in (7.17) at $(\mathbb{E}_w \nabla g_w(x), \mathbb{E}_w g_w(x))$, the zeroth order terms and first order terms vanish. Therefore condition on $N = n$,

$$
\begin{aligned}
& Y_1(x) - 0.5(Y_2(x) + Y_3(x)) \\
&= R\{\overline{\nabla g}(x; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w g_w(x)\} \\
&- 0.5R\{\overline{\nabla g}(x; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x; 1, 2^{\bar{n}_0}), \mathbb{E}_w g_w(x)\} \\
&- 0.5R\{\overline{\nabla g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \mathbb{E}_w g_w(x)\}.
\end{aligned}
$$

As a result, using (7.14) and (7.44), we have

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \frac{\mathbb{E}[\|Y_1 - 0.5(Y_2+Y_3)\|_2^2 | N=n]}{\tilde{p}_n} \\
&\le \sum_{n=0}^{\infty} \frac{3}{\tilde{p}_n}\Big(\mathbb{E}\|R\{\overline{\nabla g}(x;1,2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x;1,2^{\bar{n}_0^+}), \mathbb{E}_w g_w(x)\}\|_2^2 \\
&\quad + \frac{1}{4}\mathbb{E}\|R\{\overline{\nabla g}(x;1,2^{\bar{n}_0}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x;1,2^{\bar{n}_0}), \mathbb{E}_w g_w(x)\}\|_2^2 \\
&\quad + \frac{1}{4}\mathbb{E}\|R\{\overline{\nabla g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(x), \bar{g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}), \mathbb{E}_w g_w(x)\}\|_2^2\Big) \\
&\le \frac{3L_H^2}{4}\sum_{n=0}^{\infty}\frac{1}{\tilde{p}_n}\Big(\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \\
&\quad + \frac{1}{4}\mathbb{E}(\|\overline{\nabla g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2\Big) \\
&\quad + \frac{1}{4}\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w g_w(x)\|_2^2)^2. \qquad (7.18)
\end{aligned}
$$

Then, by (7.14),

$$
\begin{aligned}
&\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \\
&\le 2\mathbb{E}\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^4 + 2p^2\mathbb{E}\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^4.
\end{aligned}
$$

Next, we will analyze the two terms on the right hand side of the inequality above. Since $\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) = \frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}} \nabla g_{w_i}(x)$, and $\mathbb{E}\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) = \mathbb{E}_w \nabla g_w(x)$, we can write

$$
\begin{aligned}
&\mathbb{E}\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^4 \\
&= \mathbb{E}\{\sum_{k=1}^{d}\sum_{h=1}^{p}(\frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}}\{[\nabla g_{w_i}(x)]_{kh} - \mathbb{E}_w[\nabla g_w(x)]_{kh}\})^2\}^2 \\
&\le pd\sum_{k=1}^{d}\sum_{h=1}^{p}\mathbb{E}(\frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}}\{[\nabla g_{w_i}(x)]_{kh} - \mathbb{E}_w[\nabla g_w(x)]_{kh}\})^4,
\end{aligned}
$$

where the last inequality is obtained by using (7.14). Note that for I.I.D. $X_{i=1}^n$'s that $\mathbb{E}X_i = 0$, and

$|X| \le c_0$ we have

$$\begin{aligned}
&\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i)^4 \\
&= \frac{1}{n^4}\mathbb{E}\{\sum_{i=1}^{n} X_i^4 + \sum_{i\neq j}(4X_i^3X_j + 3X_i^2X_j^2) + \sum_{i\neq j\neq k} 6X_i^2X_jX_k + \sum_{i\neq j\neq k\neq h} X_iX_jX_kX_h\} \\
&= \frac{1}{n^4}\{n\mathbb{E}X_1^4 + 3n(n-1)\mathbb{E}X_1^2X_2^2\} \le \frac{3c_0^4}{n^2}.
\end{aligned}$$

Since $|[\nabla g_{w_i}(x)]_{kh} - \mathbb{E}_w[\nabla g_w(x)]_{kh}| \le 2l_{g,1}$ and $\mathbb{E}\{[\nabla g_{w_i}(x)]_{kh} - \mathbb{E}_w[\nabla g_w(x)]_{kh}\} = 0$, we have $\mathbb{E}(\frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}}\{[\nabla g_{w_i}(x)]_{kh} - \mathbb{E}_w[\nabla g_w(x)]_{kh}\})^4 \le \frac{48l_{g,1}^4}{4^{\bar{n}_0^+}}$ and hence $\mathbb{E}\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w\nabla g_w(x)\|_F^4 \le \frac{48p^2d^2l_{g,1}^4}{4^{\bar{n}_0^+}}$. By the same argument, we also have $\mathbb{E}\|\overline{\nabla g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w\nabla g_w(x)\|_F^4 \le \frac{48p^2d^2l_{g,1}^4}{4^{\bar{n}_0}}$ and $\mathbb{E}\|\overline{\nabla g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w\nabla g_w(x)\|_F^4 \le \frac{48p^2d^2l_{g,1}^4}{4^{\bar{n}_0}}$. Similarly, since $\mathbb{E}\bar{g}(x;1,2^{\bar{n}_0^+}) = \mathbb{E}_w g_w(x)$ and $|[g_{w_i}(x)]_j - \mathbb{E}_w[g_w(x)]_j| \le 2l_{g,0}$, we have

$$\begin{aligned}
&\mathbb{E}\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^4 = \mathbb{E}(\sum_{j=1}^{d}\{\frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}}([g_{w_i}(x)]_j - \mathbb{E}_w[g_w(x)]_j)\}^2)^2 \\
&\le d\sum_{j=1}^{d}\mathbb{E}\{\frac{1}{2^{\bar{n}_0^+}}\sum_{i=1}^{2^{\bar{n}_0^+}}([g_{w_i}(x)]_j - \mathbb{E}_w[g_w(x)]_j)\}^4 \le \frac{48d^2l_{g,0}^4}{4^{\bar{n}_0^+}}.
\end{aligned}$$

Using the same argument, we also have $\mathbb{E}\|\bar{g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w g_w(x)\|_2^4 \le \frac{48d^2l_{g,0}^4}{4^{\bar{n}_0}}$, and $\mathbb{E}\|\bar{g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^4 \le \frac{48d^2l_{g,0}^4}{4^{\bar{n}_0}}$. Therefore

$$\begin{aligned}
&\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w\nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \\
&\le 2\mathbb{E}\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w\nabla g_w(x)\|_F^4 + 2p^2\mathbb{E}\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^4 \\
&\le \frac{96p^2d^2(l_{g,0}^4 + l_{g,1}^4)}{4^{\bar{n}_0^+}}.
\end{aligned}$$

Hence

$$\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \leq \frac{96p^2d^2(l_{g,0}^4 + l_{g,1}^4)}{4^{\bar{n}_0}} \quad \text{and}$$

$$\mathbb{E}(\|\overline{\nabla g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \leq \frac{96p^2d^2(l_{g,0}^4 + l_{g,1}^4)}{4^{\bar{n}_0}}.$$

Now we continue with the analysis of (7.18)

$$\begin{aligned}
&\sum_{n=0}^{\infty} \frac{1}{\tilde{p}_n} \mathbb{E}\{(Y_1(x) - 0.5(Y_2(x) + Y_3(x)))^2 \mid N = n\} \\
&\leq \frac{3L_H}{4} \sum_{n=0}^{\infty} \frac{1}{\tilde{p}_n} \Big(\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \\
&\quad + \frac{1}{4}\mathbb{E}(\|\overline{\nabla g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;1,2^{\bar{n}_0}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \\
&\quad + \frac{1}{4}\mathbb{E}(\|\overline{\nabla g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(x)\|_F^2 + p\|\bar{g}(x;2^{\bar{n}_0}+1,2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(x)\|_2^2)^2 \Big) \\
&\leq 72(L_{f,1}^2 + 2df_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2) p^2 d^2 (l_{g,0}^4 + l_{g,1}^4) \sum_{n=0}^{\infty} \frac{3}{\tilde{p}_n 4^{n+n_0+1}},
\end{aligned}$$

since $L_H = \sqrt{L_1^2 + 2df_{f,2}^2 + 2dl_{g,1}^2 L_2^2}$. Note that $\tilde{p}_n = (1 - 0.5^{\gamma})0.5^{\gamma n}$ and $1 < \gamma < 2$; therefore

$$\sum_{n=0}^{\infty} \frac{3}{\tilde{p}_n 4^{n+n_0+1}} = \frac{3}{4^{n_0+1}(1-0.5^{\gamma})} \sum_{n=0}^{\infty} 2^{n(\gamma-2)} = \frac{3}{4^{n_0+1}(1-0.5^{\gamma})(1-0.5^{2-\gamma})} < \infty.$$

Hence

$$\sum_{n=0}^{\infty} \frac{\{\|Y_1(x) - 0.5(Y_2(x) + Y_3(x))\|_2^2}{\tilde{p}_n} \leq \frac{54p^2d^2(L_{f,1}^2 + 2df_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2)(l_{g,0}^4 + l_{g,1}^4)}{4^{n_0}(1-0.5^{\gamma})(1-0.5^{2-\gamma})} \quad (7.19)$$

Combining (7.16) and (7.19), we can bound (7.15) by

$$\begin{aligned}
\mathbb{E}\|G(x,v_1)\|_2^2 &\leq 2\mathbb{E}\|Y_4(x)\|_2^2 + 2\sum_{n=0}^{\infty} \frac{\{\|Y_1(x) - 0.5(Y_2(x) + Y_3(x))\|_2^2}{\tilde{p}_n} \\
&\leq 2pd^2 l_{g,1}^2 l_{f,1}^2 + \frac{108p^2d^2(L_{f,1}^2 + 2df_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2)(l_{g,0}^4 + l_{g,1}^4)}{4^{n_0}(1-0.5^{\gamma})(1-0.5^{2-\gamma})} = C'_{\mathcal{D}}.
\end{aligned}$$

□

Proposition 3 (Finite expected computational cost) For any $x \in \mathcal{D}$ and $v_1 \in \Omega_v$, the number of random numbers one needs to generate (simulation cost) to construct $G(x, v_1)$ has finite expectation.

Proof. Proof of Proposition 3. Fix $v_1 \in \Omega_v$ and $x \in \mathcal{D}$, and denote by $cost_G$ the number of random variables one needs to generate to construct $G(x, v_1)$. In Algorithm 1, we generate one geometric random variable N and 2^{n_0+n+1} number of w_i that follows the distribution of w conditioned on $v = v_1$. Thus we have $cost_G = 1 + 2^{n_0+N+1}$. Taking expectation w.r.t. N, we conclude

$$\begin{aligned}\mathbb{E}(cost_G) = \mathbb{E}\{\mathbb{E}(cost_G|N)\} &= \sum_{n=0}^{\infty} \mathbb{E}(cost_G|N=n)\mathbb{P}(N=n) \\ &= \sum_{n=0}^{\infty} (1 + 2^{n_0+n+1})(1 - 0.5^{\gamma})0.5^{\gamma n} \\ &= 1 + 2^{n_0+1}(1 - 0.5^{\gamma})(1 - 2^{1-\gamma})^{-1} < \infty,\end{aligned}$$

where the convergence of the series above relies on $\gamma > 1$. □

Remark: Note that the choices of both the base level n_0 and γ affect both the variance of the simulated estimator and its computational cost. By choosing a larger n_0, the variance of the simulated gradient will be lower but it will also have a higher computational cost. Similarly, choosing a smaller γ will result in an estimator that has lower variance but higher computational cost.

7.4.3 Convergence of the Simulated Gradient Descent Algorithm

In this subsection, we establish the convergence properties of Algorithm 10 under different conditions. Note that with the unbiasedness and finite second order moment properties of the sim-

ulated gradients, convergence properties of the Simulated Gradient Descent (SimGD) algorithm for SCO problems follow from the classical theory of the stochastic gradient descent algorithm for generic stochastic optimization problems. For completeness, we include the proof of the convergence properties in the Supplementary.

Lemma 34. *[Almost Sure Convergence] If $F(\cdot)$ is μ-strongly convex, assume $\mathbb{E}\|x_t - x_\star\|_2^2 \leq D$ for all $t \geq 0$. When $\sum_t \lambda_t = \infty$ and $\sum_t \lambda_t^2 < \infty$, $\|x_t - x_\star\|_2^2$ converges to 0 almost surely.*

The techniques of our proof for the Lemma below come mostly from [225]. We include a proof in the Supplementary.

Lemma 35. *[Rate of Convergence] In the presence of μ-strong convexity for $F(\cdot)$, with $\lambda_t = \frac{2}{\mu(t+1)}$, we can show that $\mathbb{E}\|x_T - x_\star\|_2^2 \leq \frac{4C'_{\mathcal{D}}}{\mu^2(T+1)}$ and $\mathbb{E}\|\tilde{x}_T - x_\star\|_2^2 \leq \frac{4C'_{\mathcal{D}}}{\mu^2(T+1)}$. In the case where $F(\cdot)$ is not strongly convex, if we have $\mathbb{E}\|x_t - x_\star\|_2^2 \leq D$ for all t, then with $\lambda_t = \frac{c}{\sqrt{t+1}}$ and $c > 0$, we can show that $\mathbb{E}F(\tilde{x}_T) - F(x_\star) \leq \frac{2\sqrt{2}C'_{\mathcal{D}} + c^{-1}4\sqrt{2}D}{\sqrt{T}}$.*

Corollary 7.4.0.1. *The iteration complexity of Algorithm 3 is $O(\epsilon^{-1})$ when $F(\cdot)$ is μ-strongly convex and the iteration complexit is $O(\epsilon^{-2})$ when $F(\cdot)$ is not strongly convex.*

7.4.4 Lipschitz Continuity of the Simulated Variance Reduced Gradient

In this subsection, we will present the convergence properties of the Simulated Variance Reduced Gradient (SVRG) algorithm. In contrast to the stochastic variance reduced gradient algorithm for ERM problem (7.2) , the property that

$$\mathbb{E}\|\nabla f_i(x) - f_i(\tilde{x}) + \nabla F_n(\tilde{x})\|_2^2 \leq 4L\{F_n(x) - F_n(x_\star) + F_n(\tilde{x}) - F_n(x_\star)\},$$

where i is uniformly sampled form $\{1, \ldots, n\}$ and L is the Lipschitz constant of $\nabla F_n(x)$ **may no longer hold because of the variance introduced by the simulation procedure.** Instead, we establish a Lipschitz continuity property of the output $W = W(x, v_1) - W(\tilde{x}, v_1) + G(\tilde{x})$, where

$G(\tilde{x})$ can be full gradient or a subsampled gradient at $\tilde{x}$, from Algorithm 3 that is important in the proof of the convergence rate of Algorithm 4 and 5. We need the following two lemmas to prove the results.

Lemma 36 (Azuma-Hoeffding)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables such that $|X_i| \leq B$ for all $1 \leq i \leq n$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $t > 0$, we have*

$$\mathbb{P}(|\sum_{i=1}^{n} X_i - n\mathbb{E}[X]| > t) \leq 2\exp\Big(-\frac{t^2}{2nB^2}\Big), \tag{7.20}$$

and

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[X]| > t) \leq 2\exp\Big(-\frac{nt^2}{2B^2}\Big). \tag{7.21}$$

Lemma 37. *For all $n \geq 1$, we have*

$$\mathbb{E}\Big[\sup_{x\in\mathcal{D}}|[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}|^4\Big] \leq C_1\big(\frac{log(4n^2)}{n}\big)^2 \tag{7.22}$$

$$\mathbb{E}\Big[\sup_{x\in\mathcal{D}}|[\bar{g}(x;1,n)]_h - [\mathbb{E}_w g_w(x)]_h|^4\Big] \leq C_0\big(\frac{log(4n^2)}{n}\big)^2 \tag{7.23}$$

$$\mathbb{E}\Big[\sup_{x\in\mathcal{D}}|[\overline{\nabla^2 g}(x;1,n)]_{kij} - [\mathbb{E}_w \nabla^2 g_w(x)]_{kij}|^4\Big] \leq C_2\big(\frac{log(4n^2)}{n}\big)^2 \tag{7.24}$$

for any $n \geq 1$, $1 \leq k, h \leq d$ and $1 \leq i, j \leq p$, where $C_1 = 8l_{g,1}^{4-p}(4diam(\mathcal{D})^p)p^{p/2}L_{g,1}^p + 64l_{g,1}^4(p+1)^2$, $C_0 = 8l_{g,0}^{4-p}(4diam(\mathcal{D})^p)p^{p/2}L_{g,0}^p + 64l_{g,0}^4(p+1)^2$ and $C_2 = 8l_{g,2}^{4-p}(4diam(\mathcal{D})^p)p^{p/2}L_{g,2}^p + 64l_{g,2}^4(p+1)^2$.

Proof of Lemma 37 can be found in the Supplementary.

In this subsection, we need the following ancillary functions to develop our theory. For $x \in \mathcal{H} \subset \mathbb{R}^{d\times p}$, $y \in \mathcal{G} \subset \mathbb{R}^d$ and $z \in \mathcal{J} \subset \mathbb{R}^{d\times p\times p}$, for every $1 \leq i \leq p$ and $1 \leq j \leq p$, define

$J(x, y, z) : \mathcal{H} \times \mathcal{G} \times \mathcal{J} \to \mathbb{R}^{p \times p}$ that

$$[J]_{ij}(x, y, z) = z_{:ij}^{\top} \nabla f_v(y) + [x]_{i:} \nabla^2 f_v(y) [x]_{:j}.$$

Lemma 38. *Then* $[J]_{ij}(x, y, z)$ *has Lipschitz continuous gradient with constant* L_J*, that is, for* $x_1, x_2 \in \mathcal{H}$*,* $y_1, y_2 \in \mathcal{G}$ *and* $z_1, z_2 \in \mathcal{J}$*,*

$$\|\nabla [J]_{ij}(x_1, y_1, z_1) - \nabla [J]_{ij}(x_2 y_2, z_2)\|_F \leq L_J \|vec(x_1, y_1, z_1) - vec(x_2, y_2, z_1)\|_2,$$

where

$$L_J = \{12d^2 L_{f,2}^2 l_{g,1}^2 + 4d(\sqrt{d}_{g,2} L_{f,2} + d^2 l_{g,1}^2 L_{f,3})^2 + dL_{f,1}^2 + 4d^2 l_{f,2}^2 L_{g,2}^2 + 2d^2 l_{f,2}^2 + 4d^3 l_{f,3}^2\}^{1/2}.$$

Proof of Lemma 38 can be found in the Supplementary.

Base on the ancillary function $J(x, y, z)$, for $x, x_0 \in \mathcal{H} \subset \mathbb{R}^{d \times p}$, $y, y_0 \in \mathcal{G} \subset \mathbb{R}^d$ and $z, z_0 \in \mathcal{J} \subset \mathbb{R}^{d \times p \times p}$, we define

$$\begin{aligned}
&[R]_{ij}(x, x_0, y, y_0, z, z_0) \\
&= [J]_{ij}(x, y, z) - [J]_{ij}(x_0, y_0, z_0) - \{\nabla [J]_{ij}(x_0, y_0, z_0)\}[x - x_0, y - y_0, z - z_0],
\end{aligned}$$

where

$$\begin{aligned}
&\{\nabla [J]_{ij}(x_0, y_0, z_0)\}[x - x_0, y - y_0, z - z_0] \\
&= \left(\text{vec}\{\nabla [J]_{ij}(x_0, y_0, z_0)\}\right)^{\top} \text{vec}(x - x_0, y - y_0, z - z_0) \\
&= \sum_{k'=1}^{d} \sum_{j'=1}^{d} \frac{\partial [J]_{ij}}{\partial [x]_{k'j'}}(x_0, y_0, z_0)([x]_{k'j'} - [x_0]_{k'j'}) + \sum_{h'=1}^{d} \frac{\partial [J]_{ij}}{\partial [y]_{h'}}([y]_{h'} - [y_0]_{h'}) \\
&\quad + \sum_{k''=1}^{d} \sum_{i''=1}^{p} \sum_{j''=1}^{p} \frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_0, y_0, z_0)([z]_{k''i''j''} - [z_0]_{k''i''j''})
\end{aligned}$$

Lemma 39. *For all $x, x_0 \in \mathcal{H}$, $y, y_0 \in \mathcal{G}$ and $z, z_0 \in \mathcal{J}$, we have*

$$|[R]_{ij}(x, x_0, y, y_0, z, z_0)| \leq \frac{L_J}{2}\|vec(x, y, z) - vec(x_0, y_0, z_0)\|_2^2,$$

where

$$L_J = \{12d^2L_{f,2}^2 l_{g,1}^2 + 4d(\sqrt{d}_{g,2}L_{f,2} + d^2 l_{g,1}^2 L_{f,3})^2 + dL_{f,1}^2 + 4d^2 l_{f,2}^2 L_{g,2}^2 + 2d^2 l_{f,2}^2 + 4d^3 l_{f,3}^2\}^{1/2}.$$

Proof. Proof. This result is a direct consequence of Lemma 30. □

Now we proceed with the main lemma of this section and this lemma will be used for proving convergence results for SimVRG and SCSimG algorithms.

Lemma 40. *There exist a constant $C_{\mathcal{D}} < \infty$ such that for any $v_1 \in \Omega_v$ and $x, \tilde{x} \in \mathcal{D}$, $W(x, v_1)$ and $W(\tilde{x}, v_1)$ from the variance reduced unbiased gradient $W(x, \tilde{x}, v_1) = W(x, v_1) - W(\tilde{x}, v_1) + \nabla G(\tilde{x})$ in Algorithm 3 satisfies*

$$\mathbb{E}\|W(x, v_1) - W(\tilde{x}, v_1)\|_2^2 \leq C_{\mathcal{D}}\|x - \tilde{x}\|_2^2, \tag{7.25}$$

where

$$\begin{aligned} C_{\mathcal{D}} = {} & 4p^2d^2f_{g,2}^2 l_{f,1}^2 + 4p^2d^4 l_{g,1}^4 l_{f,2}^2 \\ & + 9L_J^2 p^2 (C_0 + C_1 + C_2)\Big(\frac{(n_0+1)^2}{1-2^{\gamma-2}} + \frac{2(n_0+1)2^{\gamma-2}}{(1-2^{\gamma-2})^2} + \frac{2^{3\gamma-6} + 2^{\gamma-2}}{(1-2^{\gamma-2})^3}\Big). \end{aligned}$$

Proof of this lemma can be found in the Supplementary.

7.4.5 Convergence of the Simulated Variance Reduced Gradient Algorithm

In this section we prove the convergence of Algorithm 12. We make use of the constant $C_{\mathcal{D}}$ defined in Lemma 40 and Assumption 3 that $F(\cdot)$ is μ-strongly convex.

Lemma 41. *Let $F : \mathbb{R}^p \to \mathbb{R}$ be a convex function with L-Lipschitz gradient and denote $x_\star = \arg\min_{x\in\mathbb{R}^p} F(x)$ to be the global minimizer of $F(\cdot)$. Then for any $x \in \mathbb{R}^p$,*

$$\frac{1}{2L}\|\nabla F(x)\|_2^2 \leq F(x) - F(x_\star).$$

We omit the proof for this lemma since it is a well known result.

Theorem 1 Consider Algorithm 12 with options II. Let λ be sufficiently small and M be sufficiently large so that

$$\alpha = \frac{1}{\mu(1-\frac{4}{\mu}C_{\mathcal{D}}\lambda)\lambda M} + \frac{(\frac{4}{\mu}C_{\mathcal{D}} + 2L)\lambda}{1-\frac{4}{\mu}C_{\mathcal{D}}\lambda} < 1. \tag{7.26}$$

Then under Assumptions 1-5, we have geometric convergence in expectation for the SimVRG :

$$\mathbb{E}[F(\tilde{x}_s)] \leq F(x_\star) + \alpha^s[F(\tilde{x}_0) - F(x_\star)]$$

Proof. Proof of Theorem 1. It follows from Lemma 41 that

$$\|\nabla F(x) - \nabla F(x_\star)\|_2^2 = \|\nabla F(x)\|_2^2 \leq 2L[F(x) - F(x_\star)] \tag{7.27}$$

. Now conditioning on x_t, we can take expectation with respect to $v_t \in \Omega_v$ to obtain

$$\begin{aligned}
\mathbb{E}[\|\rho_t\|_2^2 \mid x_t] \leq& 2\mathbb{E}[\|W(x_t, v_t) - W(\tilde{x}_s, v_t)\|_2^2 \mid x_t] + 2\nabla\|F(\tilde{x}_s)\|_2^2 \\
\leq& 2C_{\mathcal{D}}\|x_t - \tilde{x}_s\|_2^2 + 4L[F(\tilde{x}_s) - F(x_\star)] \\
\leq& 4C_{\mathcal{D}}(\|x_t - x_\star\|_2^2 + \|\tilde{x}_s - x_\star\|_2^2) + 4L[F(\tilde{x}_s) - F(x_\star)] \\
\leq& \frac{8}{\mu}C_{\mathcal{D}}[F(x_t) - F(x_\star)] + (\frac{8}{\mu}C_{\mathcal{D}} + 4L)[F(\tilde{x}_s) - F(x_\star)].
\end{aligned} \tag{7.28}$$

where the second inequality follows from Theorem 40 and equation (7.27). The last inequality follows from the strong convexity of $F(\cdot)$. Thus, by the contraction property of the projection

operator $\Pi_{\mathcal{D}}$,

$$
\begin{aligned}
&\mathbb{E}[\|x_{t+1} - x_\star\|_2^2 \mid x_t] \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda(x_t - x_\star)^\intercal \mathbb{E}[\rho_t \mid x_t] + \lambda^2 \mathbb{E}[\|\rho_t\|_2^2 \mid x_t] \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda(x_t - x_\star)^\intercal \nabla F(x_t) + \frac{8}{\mu} C_{\mathcal{D}} \lambda^2 [F(x_t) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)] \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda[F(x_t) - F(x_\star)] + \frac{8}{\mu} C_{\mathcal{D}} \lambda^2 [F(x_t) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)] \\
= &\|x_t - x_\star\|_2^2 - 2\lambda(1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda)[F(x_t) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)]. \qquad (7.29)
\end{aligned}
$$

where the third line follows from the unbiasedness of the simulated gradient and the fourth line follows from the convexity of $F()$. Since $\tilde{x}_{s+1}$ is selected uniformly after all M updates are completed and $x_0 = \tilde{x}_s$. Summing over the previous inequality over $t = 0, ..., M-1$, taking expectation and using option II at stage s, we obtain

$$
\begin{aligned}
&\mathbb{E}[\|x_M - x_\star\|_2^2] + 2\lambda(1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda) M \mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)] \\
\leq &\mathbb{E}[\|x_0 - x_\star\|_2^2] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
= &\mathbb{E}[\|\tilde{x}_s - x_\star\|_2^2] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
\leq &\frac{2}{\mu} \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
= &(\frac{2}{\mu} + (\frac{8}{\mu} C_{\mathcal{D}} + 4L)\lambda^2 M) \mathbb{E}[F(\tilde{x}) - F(x_\star)] \qquad (7.30)
\end{aligned}
$$

Thus we obtain

$$
\mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)] \leq \left[\frac{1}{\mu(1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda)\lambda M} + \frac{(\frac{4}{\mu} C_{\mathcal{D}} + 2L)\lambda}{1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda} \right] \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \qquad (7.31)
$$

This implies that $\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \leq \alpha^s \mathbb{E}[F(\tilde{x}_0) - F(x_\star)]$. The conclusion follows. □

As we mentioned, the sample complexity becomes difficult to analyze in the presence of batch size randomization. However, the corollary below provides an estimate of the total number of

samples that are needed to achieve and ϵ-accurate solution for the finite sample SCO problems using Algorithm 4.

Corollary 7.4.0.2. *In Algorithm 12, let* $T_\epsilon = \min\{n \geq 0 \mid F(\tilde{x}_k) - F(x_\star) \leq \epsilon\}$ *and let* $N_{k,t}$ *be the geometric random number that is generated when calling SimulatedGradient procedure at t-th epoch and k-th iteration. Then we have*

$$\mathbb{E}\{\sum_{k=1}^{T_\epsilon}\sum_{t=1}^{M}(2^{n_0+N_{k,t}+1}+1)\} = O(\log(1/\epsilon)).$$

Proof. Proof of Corrollary 7.4.0.2. Since T_ϵ is a stopping time, by Wald's identity and Proposition 3, we have

$$\begin{aligned}\mathbb{E}\{\sum_{k=1}^{T_\epsilon}\sum_{t=1}^{M}2^{N_{k,t}+1}\} &= M\ \mathbb{E}T_\epsilon\ \mathbb{E}(2^{n_0+N_{k,t}+1}+1)\\ &= M\{1+2^{n_0+1}(1-0.5^\gamma)(1-2^{1-\gamma})^{-1}\}\mathbb{E}T_\epsilon.\end{aligned}$$

Next, we analyze $\mathbb{E}T_\epsilon$. Since T_ϵ is non- negative, we have

$$\begin{aligned}\exp(\mathbb{E}T_\epsilon) \leq \mathbb{E}\exp(T_\epsilon) &= \int_0^\infty \mathbb{P}\{\exp(T_\epsilon)\geq x\}dx = 1+\int_1^\infty \mathbb{P}\{T_\epsilon \geq \log(x)\}dx\\ &\leq 1+\int_1^\infty \mathbb{P}\{T_\epsilon \geq \lfloor\log(x)\rfloor\}dx \leq 3+\int_3^\infty \mathbb{P}\{T_\epsilon \geq \lfloor\log(x)\rfloor\}dx.\end{aligned}$$

By the definition of T_ϵ,Markov's inequality and Theorem 2 , we have

$$\mathbb{P}(T_\epsilon \geq k) \leq \mathbb{P}\{F(\tilde{x}_k)-F(x_\star)\geq\epsilon\} \leq \frac{1}{\epsilon}\mathbb{E}\{F(\tilde{x}_k)-F(x_\star)\} \leq \frac{1}{\epsilon}\alpha^k\{F(\tilde{x}_0)-F(x_\star)\}.$$

Therefore,

$$\exp\{\mathbb{E}T_\epsilon\} \leq 3+\frac{1}{\epsilon}\int_3^\infty\{F(\tilde{x}_0)-F(x_\star)\}\alpha^{\lfloor\log(x)\rfloor}dx \leq 3+\frac{F(\tilde{x}_0)-F(x_\star)}{\alpha\epsilon}\int_3^\infty x^{\log(\alpha)}dx.$$

If we choose M and λ in Algorithm 4 such that $\log \alpha < -1$, we have

$$\exp\{\mathbb{E}T_\epsilon\} \leq 3 + \frac{F(\tilde{x}_0) - F(x_\star)}{\alpha\epsilon(-\log\alpha - 1)} 3^{\log\alpha + 1}.$$

Therefore $\mathbb{E}T_\epsilon = O(\log(1/\epsilon))$. Consequently, $\mathbb{E}\{\sum_{k=1}^{T_\epsilon} \sum_{t=1}^{M} (2^{n_0+N_{k,t}+1} + 1)\} = O(1/\epsilon)$. □

Corollary 7.4.0.3. *Let $\{\tilde{x}_s\}_{s\geq 0}$ be the sequence of outputs from each epoch of the Simulated SVRG algorithm. Then, with probability 1, $\tilde{x}_s$ converges exponentially fast to $x_\star$.*

Proof. Proof of Corrollary 3. It follows from Theorem 7.4 that we can find $0 < \alpha < 1$ such that $\mathbb{E}[F(\tilde{x}_s)] \leq F(\tilde{x}_\star) + \alpha^s[F(\tilde{x}_0) - F(\tilde{x}_\star)]$. Pick any $\alpha < \rho < 1$. Define the set $\mathcal{A}_s = \{F(\tilde{x}_s) - F(x_\star) > \rho^s\}$ in probability space, we have $\mathbb{P}(\mathcal{A}_s) \leq (\frac{\alpha}{\rho})^s \mathbb{E}[F(\tilde{x}_0) - F(x_\star)]$ which implies that $\sum_{s\geq 0} \mathbb{P}(\mathcal{A}_s) < \infty$. It then follows from Borel-Cantelli lemma that

$$\begin{aligned}\mathbb{P}(\mathcal{A}_s \text{ occurs infinitely often}) = \mathbb{P}\Big(\limsup_{s\to\infty} \mathcal{A}_s\Big) = \mathbb{P}(\bigcap_{t=0}^{\infty} \bigcup_{s=t}^{\infty} \mathcal{A}_s) = \inf_{t\geq 0} \mathbb{P}(\bigcup_{s=t}^{\infty} \mathcal{A}_s) \\ \leq \inf_{t\geq 0} \sum_{s\geq t} \mathbb{P}(\mathcal{A}_s) = 0. \end{aligned} \tag{7.32}$$

Thus with probability 1, $F(\tilde{x}_s) - F(x_\star) < \rho^s$ for s large enough (depending on each the probability path), which implies $\|\tilde{x}_s - x_\star\|_2^2 \leq \frac{2}{\mu}\rho^s$ in the presence of μ-strong convexity. □

7.4.6 Convergence of the Stochastically Controlled Simulated Gradient Algorithm

In this section we prove the convergence of Algorithm 13.

Lemma 42. *Fix $x \in \mathcal{D}$ and $K, B \geq 1$, sample a batch $\mathcal{I} \subset \Omega_v$ with $|\mathcal{I}| = B$ following the distribution of v and independently generate*

$$h_k(x) = \frac{1}{B} \sum_{v_i \in \mathcal{I}} UnibasedGradient(x, v_i, n_0, \gamma)$$

for $1 \leq k \leq K$. Let $C'_{\mathcal{D}}$ be the constant in the proof of Proposition 2, where $\mathbb{E}\|W(x,v)\|_2^2 \leq C'_{\mathcal{D}}$

for arbitary $v \in \Omega_v$. *Defining* $\tilde{h}(x) = \frac{1}{K}\sum_{i=1}^{K} h_i(x)$, *we have*

$$\mathbb{E}[\tilde{h}(x)] = \nabla F(x) \qquad \textit{and} \qquad Var[\tilde{h}(x)] \leq \frac{C'_{\mathcal{D}}}{KB} + 4pd^2 l_{\mathcal{D}}^4 (\frac{1}{K} + \frac{1}{B}), \tag{7.33}$$

so $Var[\tilde{h}(x)]$ *can be made arbitrarily small for any* $x \in \mathcal{D}$ *by making* K *and* B *sufficiently large.*

Proof of this Lemma can be found in the Supplementary.

Theorem 2 Consider the Simulated SCSG Algorithm 6 with options II. Fix $\epsilon > 0$ as the level of accuracy. Let λ be sufficiently small and M be sufficiently large so that

$$\alpha = \frac{2}{\mu(1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda)\lambda M} + \frac{(\frac{8}{\mu}C_{\mathcal{D}} + 8L)\lambda}{1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda} < 1, \tag{7.34}$$

while making either K or B large enough so that

$$\frac{4(\lambda + \frac{1}{2\mu})}{1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda} Var[\tilde{h}(\tilde{x}_s)] < \epsilon \tag{7.35}$$

Then

$$\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \leq \alpha^s \mathbb{E}[F(\tilde{x}_0) - F(x_\star)] + \frac{1}{1-\alpha}\epsilon \tag{7.36}$$

Proof. Proof of Theorem 2 Conditioning on x_t, we can take expectation with respect to $v_t \in \Omega_v$ to obtain

$$\begin{aligned}
&\mathbb{E}[\|\rho_t\|_2^2 \mid x_t] \\
\leq& 2\mathbb{E}[\|W(x_t, v_t) - W(\tilde{x}_s, v_t)\|_2^2 \mid x_t] + 4\|\nabla F(\tilde{x}_s)\|_2^2 + 4\|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 \\
\leq& 2C_{\mathcal{D}}\|x_t - \tilde{x}_s\|_2^2 + 8L[F(\tilde{x}_s) - F(x_\star)] + 4\|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 \\
\leq& 4C_{\mathcal{D}}(\|x_t - x_\star\|_2^2 + \|\tilde{x}_s - x_\star\|_2^2) + 8L[F(\tilde{x}_s) - F(x_\star)] + 4\|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 \\
\leq& \frac{8}{\mu}C_{\mathcal{D}}[F(x_t) - F(x_\star)] + (\frac{8}{\mu}C_{\mathcal{D}} + 8L)[F(\tilde{x}_s) - F(x_\star)] + 4\|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2.
\end{aligned} \tag{7.37}$$

where the second inequality follows from Lemma 40 and equation (7.27). The last inequality

follows from the strong convexity of $F(\cdot)$. Now following (7.37), using the distance contraction property of projection operator $\Pi_{\mathcal{D}}(\cdot)$ we can write

$$
\begin{aligned}
&\mathbb{E}[\|x_{t+1} - x_\star\|_2^2 \mid x_t] \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda(x_t - x_\star)^\intercal \mathbb{E}[\rho_t \mid x_t] + \lambda^2 \mathbb{E}[\|\rho_t\|_2^2 \mid x_t] \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda(x_t - x_\star)^\intercal (\nabla F(x_t) - \nabla F(\tilde{x}_s) + \tilde{h}(\tilde{x}_s)) + \frac{8}{\mu} C_{\mathcal{D}} \lambda^2 [F(x_t) - F(x_\star)] \\
&+ (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)] + 4\lambda^2 \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 \\
\leq &\|x_t - x_\star\|_2^2 - 2\lambda[F(x_t) - F(x_\star)] + 2\lambda(x_t - x_\star)^\intercal (\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)) \\
&+ \frac{8}{\mu} C_{\mathcal{D}} \lambda^2 [F(x_t) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)] + 4\lambda^2 \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 \\
= &\|x_t - x_\star\|_2^2 - 2\lambda(1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda)[F(x_t) - F(x_\star)] + (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 [F(\tilde{x}_s) - F(x_\star)] \\
&+ 4\lambda^2 \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 + 2\lambda(x_t - x_\star)^\intercal (\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)),
\end{aligned} \tag{7.38}
$$

where the third line follows from the convexity of $F(\cdot)$. Now we consider a fixed stage s, so that $x_0 = \tilde{x}_s$ and $\tilde{x}_{s+1}$ is selected uniformly after all M updates are completed. Summing the previous inequality over $t = 1, ..., M$, taking expectation and using option II at stage s, we obtain

$$
\begin{aligned}
&\mathbb{E}[\|x_M - x_\star\|_2^2] + 2\lambda(1 - \frac{4}{\mu} C_{\mathcal{D}} \lambda) M \mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)] \\
\leq &\mathbb{E}[\|x_0 - x_\star\|_2^2] + (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
&+ 4\lambda^2 M \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 + 2\lambda M \mathbb{E}[(\tilde{x}_{s+1} - x_\star)^\intercal (\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s))] \\
\leq &\mathbb{E}[\|\tilde{x}_s - x_\star\|_2^2] + (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
&+ 4\lambda M(\lambda + \frac{1}{2\mu}) \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 + \frac{\mu}{2} \lambda M \mathbb{E}[\|\tilde{x}_{s+1} - x_\star\|_2^2] \\
\leq &\mathbb{E}[\|\tilde{x}_s - x_\star\|_2^2] + (\frac{8}{\mu} C_{\mathcal{D}} + 8L)\lambda^2 M \mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \\
&+ 4\lambda M(\lambda + \frac{1}{2\mu}) \|\tilde{h}(\tilde{x}_s) - \nabla F(\tilde{x}_s)\|_2^2 + \lambda M \mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)],
\end{aligned} \tag{7.39}
$$

where the second inequality follows from $2a^\intercal b \leq \beta\|a\|_2^2 + \frac{1}{\beta}\|b\|_2^2$ while $\beta = \frac{\mu}{2}$. The last inequality follows from the strong convexity of $F(\cdot)$. Finally, taking expectation over the randomness of $\tilde{h}(\tilde{x}_s)$, we have

$$\begin{aligned}
&\lambda(1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda)M\mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)] \\
\leq&\mathbb{E}[\|\tilde{x}_s - x_\star\|_2^2] + (\frac{8}{\mu}C_{\mathcal{D}} + 8L)\lambda^2 M\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + 4\lambda M(\lambda + \frac{1}{2\mu})Var[\tilde{h}(\tilde{x_s})] \\
\leq&\frac{2}{\mu}\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + (\frac{8}{\mu}C_{\mathcal{D}} + 8L)\lambda^2 M\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + 4\lambda M(\lambda + \frac{1}{2\mu})Var[\tilde{h}(\tilde{x}_s)] \\
=&(\frac{2}{\mu} + (\frac{8}{\mu}C_{\mathcal{D}} + 8L)\lambda^2 M)\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + 4\lambda M(\lambda + \frac{1}{2\mu})Var[\tilde{h}(\tilde{x}_s)] \qquad (7.40)
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
&\mathbb{E}[F(\tilde{x}_{s+1}) - F(x_\star)] \\
\leq&\left[\frac{2}{\mu(1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda)\lambda M} + \frac{(\frac{8}{\mu}C_{\mathcal{D}} + 8L)\lambda}{1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda}\right]\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + \frac{4(\lambda + \frac{1}{2\mu})}{1 - \frac{8}{\mu}C_{\mathcal{D}}\lambda}Var[\tilde{h}(\tilde{x}_s)] \\
\leq&\alpha\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] + \epsilon \qquad (7.41)
\end{aligned}$$

This implies that $\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \leq \alpha^s\mathbb{E}[F(\tilde{x}_0) - F(x_\star)] + \frac{\epsilon}{1-\alpha}$. The conclusion follows. □

Corollary 7.4.0.4. *Let $\{\tilde{x}_s\}_{s\geq 0}$ be the sequence of outputs from each epoch of the Simulated SCSG algorithm and define $\tilde{y}_s = \min_{t\leq s}\{F(\tilde{x}_t) - F(x_\star)\}$ for $s \geq 0$ to be the lowest objective value after epoch s. Then, with probability 1, we have $\inf_{s\geq 0}\tilde{y}_s \leq \frac{\epsilon}{1-\alpha}$.*

Proof. Proof of Corollary 4. It follows from Theorem 3 that we can find $0 < \alpha < 1$ where $\mathbb{E}[F(\tilde{x}_s) - F(x_\star)] \leq \alpha^s\mathbb{E}[F(\tilde{x}_0) - F(x_\star)] + \frac{\epsilon}{1-\alpha}$. We also have $\sup_{x\in\mathcal{D}}\{F(x) - F(x_\star)\} \leq 2l_{\mathcal{D}}$ from the definition of $l_{\mathcal{D}}$. It follows that for any $\tilde{x}_0 \in \mathcal{D}$, we have that $\mathbb{E}[F(\tilde{x}_s) - F(x_\star)|\tilde{x}_0] \leq \alpha^s \cdot 2l_{\mathcal{D}} + \frac{\epsilon}{1-\alpha}$. For any $\rho > 0$, picking N large enough so that $\delta = (\alpha^N \cdot 2l_{\mathcal{D}} + \frac{\epsilon}{1-\alpha})(\frac{\epsilon}{1-\alpha} + \rho)^{-1} < 1$, we have

$$\mathbb{P}(\tilde{y}_N \geq \frac{\epsilon}{1-\alpha} + \rho) \leq \mathbb{P}(F(\tilde{x}_N) - F(x) \geq \frac{\epsilon}{1-\alpha} + \rho) \leq \mathbb{E}[F(\tilde{x}_0) - F(x)](\frac{\epsilon}{1-\alpha} + \rho)^{-1} \leq \delta.$$

However, if we denote $\mathcal{X}_N$ to be the distribution of $\tilde{x}_N$ conditioning on $\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho$, then it follows from the Markov Property that

$$
\begin{aligned}
&\mathbb{P}(\tilde{y}_{2N} \geq \frac{\epsilon}{1-\alpha}+\rho) \\
=&\mathbb{P}(\tilde{y}_{2N} \geq \frac{\epsilon}{1-\alpha}+\rho|\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho)\mathbb{P}(\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho) \\
=&\mathbb{P}(\min_{N+1\leq s\leq 2N}\{F(\tilde{x}_s)-F(x_\star)\} \geq \frac{\epsilon}{1-\alpha}+\rho|\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho)\mathbb{P}(\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho) \\
=&(\mathbb{P}_{\tilde{x}_N\sim\mathcal{X}_N}\mathbb{P}(\min_{N+1\leq s\leq 2N}\{F(\tilde{x}_s)-F(x_\star)\} \geq \frac{\epsilon}{1-\alpha}+\rho|\tilde{x}_N))\cdot\mathbb{P}(\tilde{y}_N \geq \frac{\epsilon}{1-\alpha}+\rho) \\
\leq&(\mathbb{P}_{\tilde{x}_N\sim\mathcal{X}_N}\mathbb{P}(F(\tilde{x}_{2N})-F(x_\star) \geq \frac{\epsilon}{1-\alpha}+\rho|\tilde{x}_N))\cdot\delta \\
\leq&(\mathbb{P}_{\tilde{x}_N\sim\mathcal{X}_N}\mathbb{E}[F(\tilde{x}_{2N})-F(x_\star)|\tilde{x}_N])\cdot(\frac{\epsilon}{1-\alpha}+\rho)^{-1}\cdot\delta \\
=&(\mathbb{P}_{\tilde{x}_0\sim\mathcal{X}_N}\mathbb{E}[F(\tilde{x}_N)-F(x_\star)|\tilde{x}_0])\cdot(\frac{\epsilon}{1-\alpha}+\rho)^{-1}\cdot\delta \\
\leq&\mathbb{P}_{\tilde{x}_0\sim\mathcal{X}_N}(\alpha^N\cdot 2l_{\mathcal{D}}+\frac{\epsilon}{1-\alpha})\cdot(\frac{\epsilon}{1-\alpha}+\rho)^{-1}\cdot\delta \leq \delta^2
\end{aligned}
$$

Continue on, we can prove that $\mathbb{P}(\tilde{y}_{kN} \geq \frac{\epsilon}{1-\alpha}+\rho) \leq \delta^k$. Thus if we define the set $\mathcal{A}_\rho = \{\inf_{s\geq 0}\tilde{y}_s \geq \frac{\epsilon}{1-\alpha}+\rho\}$ and $\mathcal{A} = \{\inf_{s\geq 0}\tilde{y}_s > \frac{\epsilon}{1-\alpha}\}$ in probability space, we have

$$
\mathbb{P}(\mathcal{A}_\rho) = \mathbb{P}(\inf_{s\geq 0}\tilde{y}_s \geq \frac{\epsilon}{1-\alpha}+\rho) \leq \mathbb{P}(\tilde{y}_{kN} \geq \frac{\epsilon}{1-\alpha}+\rho) \leq \delta^k, \tag{7.42}
$$

for any $k \geq 1$. Since $\delta < 1$, we have $\mathbb{P}(\mathcal{A}_\rho) = 0$ for any $\rho > 0$ which implies $\mathbb{P}(\mathcal{A}) = \mathbb{P}(\bigcup_{n\geq 1}\mathcal{A}_{\frac{1}{n}}) \leq \sum_{n=1}^{\infty}\mathbb{P}(\mathcal{A}_{\frac{1}{n}}) = 0$. So, with probability 1, $\inf_{s\geq 0}\tilde{y}_s \leq \frac{\epsilon}{1-\alpha}$. □

7.5 Numerical Experiments

In our numerical experiments, all algorithms were implemented in C++, and all experiments were performed on an Intel i5-5200U processor using Ubuntu 16.04.

7.5.1 Cox's partial likelihood

We implemented Algorithms 2(SGD),4(SimVRG) and 5(SCSimG) to minimize a regularized Cox's negative partial log- likelihood and compared their performance with the Compositional-SVRG-1 algorithm (Comp-SVRG-1) in [207], the Stochastic Compositional Gradient Descent algorithm (SCGD) in [206] and Gradient Descent(GD) algorithm. The optimization problem in Section 3.2 combined with L_2 regulation can be written as:

$$\min_{\beta\in\mathbb{R}^p} \quad \frac{1}{n}\sum_{i=1}^{n}\Delta_i[-X_i^\top\beta + \log\{\sum_{j=1}^{n}\mathbb{I}(Y_j \geq Y_i)\exp(X_j^\top\beta)\}] + \frac{1}{2}\|\beta\|_2^2,, \tag{7.43}$$

where (X_i, Y_i, Δ_i) and T_i, C_i for $i = 1, \ldots, n$ come from the Cox's model as in the setting of Section 3.2. Here, we generated our dataset by seting $n = 10^4$, $p = 10^3$ and letting X_i follow i.i.d. standard normal distribution. Moreover, T_i was generated according to the standard exponential base line hazard function and C_i was generated independent of T_i with a 30% censoring rate. One can check that each component function is strongly convex with Lipschitz continuous gradients. The numerical results are presented below in Figure 1.

In Figure 1, the left plot is the logarithm of the objective value minus the optimal value versus the number of iterations while the right plot is the logarithm of the same difference versus the CPU running time. We compare both the running time and the iteration number to give a more comprehensive review of each algorithm since the iteration time for each algorithm could be drastically different due to different update rules. Moreover, the parameters in each algorithm were selected and tuned to achieve a relatively optimal performance without heavily increasing the computational cost. In Algorithm 5, we set $\lambda = 0.01$, $\gamma = 3/2$ $M = 100$ and $n_0 = 0$, in Algorithm 6, $\lambda = 0.0005$, $M = 100$, $B = 100$, $K = 50$ and $n_0 = 2$, in Compositional SVRG-1, $\lambda = 0.001$, $M = 100$ and $B = 500$, in Gradient Descent $\lambda = 0.01$.

As we can see, in the left plot, the SimVRG and Compositional-SVRG-1 algorithm performed best amongst all algorithms while SimVRG also had better performance in the right plot. Algorithm 6, SCSimG was slightly less effective due to the lack of full gradient computation, but, as

expected from the theorems in Section 4, Algorithm 6 also converged linearly to the optimal solution. SimGD algorithm is ploted for every 50 iterations for the sake of fairness (to account for the inner loop in the other algorithms) and it also showed satisfactory performance without the presence of variance reducetion techinques.

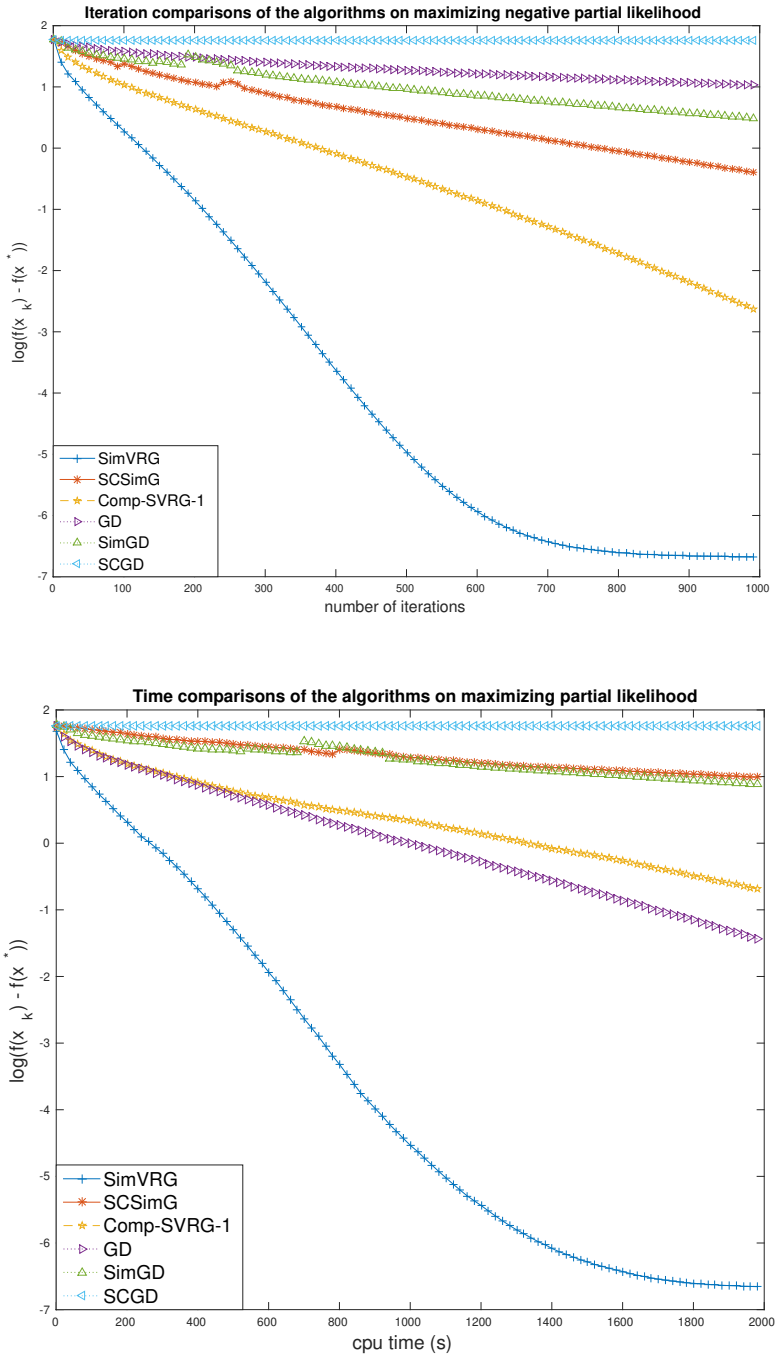


Figure 7.1: Performance plots for different algorithms on Cox's partial likelihood dataset. For both plots, the y-axis is the logarithm of the objective value minus the optimal value. For the plot on the left, the x-axis is number of iterations while for the right plot, the x-axis is the running time of the algorithms.

7.5.2 Conditional Random Fields

We implemented Algorithms 3(SimGD),5(SimVRG) and 6(SCSimG) to train conditional random field models and compared their performance with the Compositional-SVRG-1 algorithm (Comp-SVRG-1) in [207], the Stochastic Compositional Gradient Descent algorithm (SCGD) in [206] and Gradient Descent(GD) algorithm. However, we used the optical character recognition (OCR) data in [242]. Specifically, the ORC dataset provides labelling for letters in a image composed of words. The numerical results are summarized in Figure 2.

Once again, to make comparisons fair, the performance of algorithms are measured both in number of iterations and CPU time. For the parameters, in Algorithm 5, we have $\lambda = 0.001$,$\gamma = 3/2$ $M = 200$ and $n_0 = 0$, in Algorithm 6, $\lambda = 0.0001$, $M = 200$, $B = 100$, $K = 10$ and $n_0 = 2$, in Gradient Descent, $\lambda = 0.01$. In other algorithms, the parameters are chosen according to their convergence theorem with scaling factor 0.5. For example, basic SCGD corresponds to Theorem 6 in [206].

As we can see from the figures, once again, the SimVRG of Algorithm 5 has the best performance amongst the group. However, in this example, the gradient descent algorithm actually outperforms Algorithm 6, SimVRG in terms iteration complexity. This is possibly due to the lack of accurate gradient estimation in Algorithm 6. Specifically, as the dataset grows large, it becomes more costly to obtain accurate gradient estimate. On the other hand, the SimGD in Algorithm 3 outperforms SCGD in terms of iterations and CPU time for both datasets. We note that the occasional increase of function value in some executions of the SimGD algorithm is caused by the variance of our gradient simulation.

7.6 Conclusion and Future Work.

In this chapter, we introduced unbiased gradient simulation algorithms that are based on a multilevel Monte Carlo technique for solving stochastic compositional optimization (SCO) problems and proved convergence of our algorithms and applied them on a number of different statistical

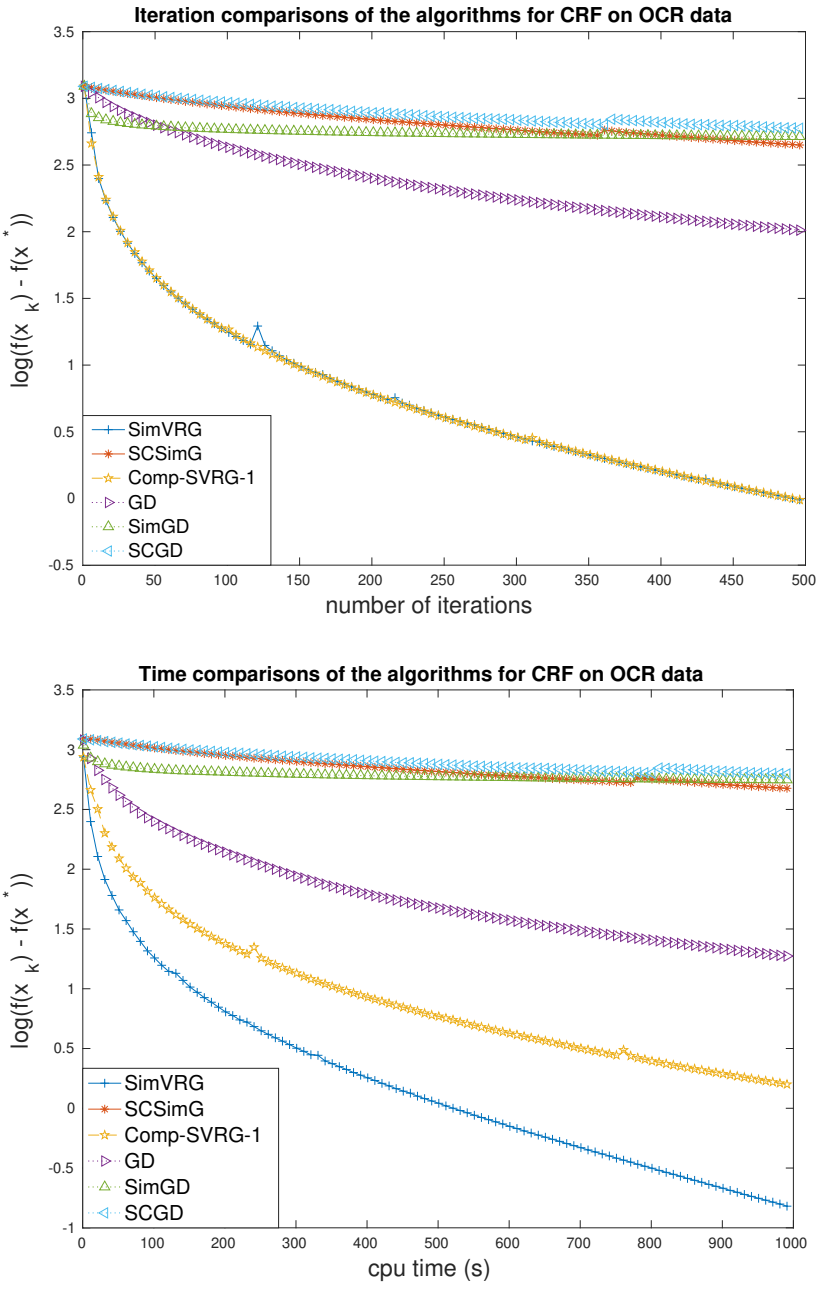


Figure 7.2: Performance plots for different algorithms on the OCR dataset. For both plots, the y-axis is the logarithm of the objective value minus the optimal value. For the plot on the left, the x-axis is number of iteration while for the right plot, the x-axis is the running time of the algorithms.

and machine learning problems.

There are several directions where we can expand upon our work. For example, different accelerating schemes and second order methods usually show fast convergence in practice, and can be extended using simulated gradients for SCO problems. Another direction is to extend our approach to adaptive step size schemes. A limitation of our unbiased gradient simulation algorithm is the requirement for smoothness of the objective function. Therefore, developing unbiased simulation of sub-gradient methods and utilizing them for optimizing non-smooth functions is also of great interest. Analyzing the sample complexity of our algorithms and the optimal choice of the parameters are also interesting problems for future work.

7.7 Supplementary A:Proof of Lemma 32

Proof. Proof. Before proving this lemma, we introduce the notation for partial derivatives of $H(s,t)$, i.e., each component of the gradient $\nabla H(s,t) \in \mathbb{R}^p \times (\mathbb{R}^{d\times p} \times \mathbb{R}^d)$. Let

$$\frac{\partial [H]_i}{\partial [s]_{kj}}(s,t) = \delta_{ij}\frac{\partial f_{v_1}}{\partial [t]_k}(t), \text{ and} \frac{\partial [H]_i}{\partial [t]_h}(s,t) = \sum_{k=1}^{d}[s]_{ki}\frac{\partial [\nabla f_{v_1}]_k}{\partial [t]_h} = \sum_{k=1}^{d}[s]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t),$$

where $1 \le i \le p, 1 \le j \le p$, $1 \le k \le d$, $1 \le h \le d$, and δ_{ij} is the Kronecker delta, i.e., $\delta_{ij} = 1$ when $i = j$; $\delta_{ij} = 0$ otherwise. Note that by Assumption 1, ∇f_{v_1} is Lipschitz continuous with constant $L_{f,1}$; therefore $\frac{\partial [H]_i}{\partial [s]_{kj}}(s,t)$, which is the partial derivative of ∇f_{v_1}, is Lipschitz continuous with constant $L_{f,1}$. By Assumption 1, $\nabla^2 f_{v_1}$ is Lipschitz continuous with constant $L_{f,2}$; therefore $\frac{\partial [H]_i}{\partial [t]_h}(s,t)$ is Lipschitz continuous with constant $L_{f,2}$. Therefore

$$\begin{aligned}
&\|\nabla [H]_i(s_1,t_1) - \nabla [H]_i(s_2,t_2)\|_F \\
&\le \sqrt{\sum_{k=1}^{d}\sum_{j=1}^{p}(\delta_{ij}\frac{\partial f_{v_1}}{\partial [t]_k}(t_1) - \delta_{ij}\frac{\partial f_{v_1}}{\partial [t]_k}(t_2))^2 + \sum_{h=1}^{d}(\sum_{k=1}^{d}[s_1]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1) - \sum_{k=1}^{d}[s_2]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_2))^2}.
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{k=1}^{d}\sum_{j=1}^{p}\{\delta_{ij}\frac{\partial f_{v_1}}{\partial [t]_k}(t_1) - \delta_{ij}\frac{\partial f_{v_1}}{\partial [t]_k}(t_2)\}^2 &= \sum_{k=1}^{d}\{\frac{\partial f_{v_1}}{\partial [t]_k}(t_1) - \frac{\partial f_{v_1}}{\partial [t]_k}(t_2)\}_2^2 = \|\nabla f_{v_1}(t_1) - \nabla f_{v_2}(t_2)\|_2^2 \\
&\le L_{f,1}^2\|t_1 - t_2\|_2^2
\end{aligned}$$

using the fact that $|[s_2]_{ki}| \le l_{g,1}$ $|\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_2)| \le l_{f,2}$ for all k and h,

$$
\begin{aligned}
&\sum_{h=1}^{d}\{\sum_{k=1}^{d}[s_1]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1) - \sum_{k=1}^{d}[s_2]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_2)\}^2 \\
\le\ & 2\sum_{h=1}^{d}\{\sum_{k=1}^{d}[s_1]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1) - \sum_{k=1}^{d}[s_2]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1)\}^2 \\
&+ 2\sum_{h=1}^{d}\{\sum_{k=1}^{d}[s_2]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1) - \sum_{k=1}^{d}[s_2]_{ki}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_2)\}^2 \\
\le\ & 2l_{f,2}^2\sum_{h=1}^{d}\{\sum_{k=1}^{d}[s_1]_{ki} - [s_2]_{ki}\}^2 + 2l_{g,1}^2\sum_{h=1}^{d}\{\sum_{k=1}^{d}\frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_1) - \frac{\partial^2 f_{v_1}}{\partial [t]_k \partial [t]_h}(t_2)\}^2 \\
\le\ & 2l_{f,2}^2 d\|[s_1]_{:i} - [s_2]_{:i}\|_2^2 + 2l_{g,1}^2 dL_{f,2}^2\|t_1 - t_2\|_2^2.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\|\nabla [H]_i(s_1,t_1) - \nabla [H]_i(s_2,t_2)\|_F &\le \sqrt{L_{f,1}^2\|t_1 - t_2\|_2^2 + 2l_{f,2}^2 d\|[s_1]_{\cdot i} - [s_2]_{\cdot i}\|_2^2 + 2dl_{g,1}^2 L_{f,2}^2\|t_1 - t_2\|_2^2} \\
&\le \sqrt{L_{f,1}^2 + 2dl_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2}\|\mathrm{vec}([s_1]_{:i}, t_1) - \mathrm{vec}([s_2]_{:i}, t_2)\|_2 \\
&= L_H\|\mathrm{vec}([s_1]_{:i}, t_1) - \mathrm{vec}([s_2]_{:i}, t_2)\|_2.
\end{aligned}
$$

□

7.8 Supplementary B: Proof of Lemma 33

Proof. Proof. Recall that $\nabla H(s,t)[u,v] \in \mathbb{R}^p$, $u \in \mathbb{R}^{d\times p}$, $v \in \mathbb{R}^d$ and each component $\nabla H(s,t)$ is defined as

$$
[\nabla H(s,t)[u,v]]_i = \nabla [H]_i(s,t)[u,v] = \sum_{k=1}^{d}\sum_{j=1}^{p}\frac{\partial [H]_i}{\partial [s]_{kj}}(s,t)\cdot [u]_{kj} + \sum_{h=1}^{d}\frac{\partial [H]_i}{\partial [t]_h}(s,t)\cdot [v]_h.
$$

Note that $R(s, s_0, t, t_0)$ can be considered as the remainder of the first order Taylor expansion of $H(s,t)$ at (s_0, t_0). Now using Lemma 30, we have

$$
\begin{aligned}
\|R(s,s_0,t,t_0)\|_2 &= \|H(s,t) - H(s_0,t_0) - \nabla H(s_0,t_0)[(s-s_0),(t-t_0)]\|_2 \\
&= \sqrt{\sum_{i=1}^{p} |[H]_i(s,t) - [H]_i(s_0,t_0) - \nabla [H]_i(s_0,t_0)[s-s_0,t-t_0]|^2} \\
&\leq \sum_{i=1}^{p} |[H]_i(s,t) - [H]_i(s_0,t_0) - \nabla [H]_i(s_0,t_0)[s-s_0,t-t_0]| \\
&\leq \sum_{i=1}^{p} \frac{1}{2}\sqrt{L_{f,1}^2 + 2dl_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2}\, \|\mathrm{vec}([s]_i, t) - \mathrm{vec}([s_0]_i, t_0)\|_2^2 \\
&= \frac{1}{2}\sqrt{L_{f,1}^2 + 2dl_{f,2}^2 + 2dl_{g,1}^2 L_{f,2}^2}(\|s-s_0\|_F^2 + p\|t-t_0\|_2^2) \\
&= \frac{L_H}{2}(\|s-s_0\|_F^2 + p\|t-t_0\|_2^2)
\end{aligned} \tag{7.44}
$$

for any $x, x_0 \in \mathcal{H}$ and $y, y_0 \in \mathcal{G}$.

□

7.9 Supplementary C: Proof of Lemma 34

Proof. Proof. Define $Y_t = \|x_t - x_\star\|_2^2$. By the contraction property of projection operators, we have $Y_{t+1} = \|x_{t+1} - x_\star\|_2^2 = \|\Pi_{\mathcal{D}}(x_t - \lambda_t \rho_t) - \Pi_{\mathcal{D}}(x_\star)\|_2^2 \leq \|x_t - \lambda_t\rho_t - x_\star\|_2^2$. Thus

$$
Y_{t+1} - Y_t \leq \|x_{t+1} - x_\star\|^2 = \|x_t - x_\star\|^2 = -2\lambda_t (x_t - x_\star)^\intercal \rho_t + \lambda_t^2 \|\rho_t\|_2^2, \tag{7.45}
$$

Moreover, with respect to the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$, we can obtain, using Proposition 1 and 2, $\mathbb{E}\{\rho_t \,|\, \mathcal{F}_t\} = \nabla F(x_t)$ and $\mathbb{E}\{\|\rho_t\|_2^2 \,|\, \mathcal{F}_t\} \leq C'_{\mathcal{D}}$ and by convexity of $F(\cdot)$, we have $0 \geq F(x_\star) - F(x) \geq (x_\star - x)^\intercal \nabla F(x)$. Therefore

$$
\mathbb{E}[Y_{t+1} - Y_t | \mathcal{F}_t] \leq -2\lambda_t (x_t - x_\star)^\intercal \nabla F(x_t) + \lambda_t^2 C'_{\mathcal{D}} \leq \lambda_t^2 C'_{\mathcal{D}}. \tag{7.46}
$$

Define $M_t = Y_t + \sum_t^\infty \lambda_s^2 C'_{\mathcal{D}}$ with respect to the natural filtration $\mathcal{F}_t$. Then it can be checked that M_t is a positive supermartingale with finite expected values. Thus, it follows from the martingale convergence theorem that M_t and consequently $Y_t = \|x_t - x_\star\|_2^2$ converges almost surely. To show that $\|x_t - x_\star\|_2^2 \to 0$, we define $Z_t = \sum_{s=0}^t 2\lambda_t (x_t - x_\star)^\intercal \nabla F(x_t)$, and notice that $0 \leq Z_t \leq Z_{t+1}$ due to convexity of $F(\cdot)$. Therefore, using the monotone convergence theorem and (7.46) we have

$$\begin{aligned}\mathbb{E}[\sum_t 2\lambda_t (x_t - x_\star)^\intercal \nabla F(x_t)] &\leq \sum_t \mathbb{E}[2\lambda_t (x_t - x_\star)^\intercal \nabla F(x_t)] \\ &= \sum_t \mathbb{E}[Y_t] - \mathbb{E}[Y_{t+1}] + \lambda_t^2 \mathbb{E}[\rho_t^2] \leq D + \sum_t \lambda_t^2 C'_{\mathcal{D}} < \infty. \end{aligned} \tag{7.47}$$

Thus the monotone series $Z_t = \sum_{s \leq t} 2\lambda_s (x_s - x_\star)^\intercal \nabla F(x_s)$ converges almost surely. It follows from $\sum_t \lambda_t = \infty$ and $(x_t - x_\star)^\intercal \nabla F(x_t) \geq 0$ that $(x_t - x_\star)^\intercal \nabla F(x_t) \to 0$. Since $F(\cdot)$ is μ-strongly convex, we have $(x_t - x_\star)^\intercal \nabla F(x_t) \geq \mu \|x_t - x_\star\|_2^2$, which implies $\|x_t - x_\star\|_2^2 \to 0$.

□

7.10 Supplementary D: Proof of Lemma 35

Proof. Proof. By the contraction property of projection operators, we have

$$\begin{aligned}\mathbb{E}[\|x_t - x_\star\|^2 | x_{t-1}] &\leq \mathbb{E}[\|x_{t-1} - \lambda_t \rho_{t-1} - x_\star\|^2 | x_{t-1}] \\ &= \|x_{t-1} - x_\star\|_2^2 + \lambda_t^2 \mathbb{E}[\|\rho_{t-1}\|_2^2 | x_{t-1}] - 2\lambda_t (x_{t-1} - x_\star)^\intercal \mathbb{E}[\rho_{t-1} | x_{t-1}] \\ &= \|x_{t-1} - x_\star\|_2^2 + \lambda_t^2 \mathbb{E}[\|\rho_{t-1}\|_2^2 | x_{t-1}] - 2\lambda_t (x_{t-1} - x_\star)^\intercal \nabla F(x_{t-1}) \\ &\leq \|x_{t-1} - x_\star\|_2^2 + \lambda_t^2 C'_{\mathcal{D}} - 2\lambda_t (F(x_{t-1}) - F(x_\star) + \frac{\mu}{2} \|x_{t-1} - x_\star\|_2^2). \end{aligned} \tag{7.48}$$

The third line follows from the Proposition 1 and the fourth line follows from Proposition 2 and strong convexity. Now we have

$$\mathbb{E}[F(x_{t-1})] - F(x_\star) \leq \frac{\lambda_t C'_{\mathcal{D}}}{2} + \frac{\lambda_t^{-1} - \mu}{2} \mathbb{E}\|x_{t-1} - x_\star\|_2^2 - \frac{\lambda_t^{-1}}{2} \mathbb{E}\|x_t - x_\star\|_2^2. \tag{7.49}$$

Finally, with $\lambda_t = \frac{2}{\mu(t+1)}$, it follows from the convexity of $F(\cdot)$ that

$$
\begin{aligned}
0 \leq \mathbb{E}[F(\tilde{x}_T)] - F(x_\star) &\leq \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} (t+1)\big(\mathbb{E}[F(x_t)] - F(x_\star)\big) \\
&\leq \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} \frac{t+1}{t+2} \frac{C'_{\mathcal{D}}}{\mu} + \frac{\mu}{4}\big((t)(t+1)\mathbb{E}\|x_{t-1} - x_\star\|_2^2 - (t+1)(t+2)\mathbb{E}\|x_t - x_\star\|_2^2\big) \\
&\leq \frac{2C'_{\mathcal{D}}}{\mu(T+1)} - \frac{\mu}{2}\mathbb{E}\|x_T - x_\star\|_2^2. \qquad (7.50)
\end{aligned}
$$

The last inequality implies that both $\mathbb{E}\|x_T - x_\star\|_2^2 \leq \frac{4C'_{\mathcal{D}}}{\mu^2(T+1)}$ and $\mathbb{E}\|\tilde{x}_T - x_\star\|_2^2 \leq \frac{4C'_{\mathcal{D}}}{\mu^2(T+1)}$ (using strong convexity).

When $F(\cdot)$ is non-strongly convex, we can use the convexity of $F(\cdot)$ so that the last inequality of (7.48) becomes

$$
\mathbb{E}[\|x_t - x_\star\|^2 | x_{t-1}] \leq \|x_{t-1} - x_\star\|_2^2 + \lambda_t^2 C'_{\mathcal{D}} - 2\lambda_t(F(x_{t-1}) - F(x_\star)), \qquad (7.51)
$$

Thus we have

$$
\mathbb{E}[F(x_{t-1})] - F(x_\star) \leq \frac{\lambda_t C'_{\mathcal{D}}}{2} + \frac{\lambda_t^{-1}}{2}\mathbb{E}\|x_{t-1} - x_\star\|_2^2 - \frac{\lambda_t^{-1}}{2}\mathbb{E}\|x_t - x_\star\|_2^2. \qquad (7.52)
$$

Finally, with $\lambda_t = \frac{c}{\sqrt{(t+1)}}$, it follows from the convexity of $F()$ and the assumption that $\mathbb{E}\|x_t - x_\star\|_2^2 \leq D$ that

$$
\begin{aligned}
0 \leq \mathbb{E}[F(\tilde{x}_T)] - F(x_\star) &\leq \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} (t+1)\big(\mathbb{E}[F(x_t)] - F(x_\star)\big) \\
&\leq \frac{\sqrt{2}}{2c(T)(T+1)} D + \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} (t+1) \frac{cC'_{\mathcal{D}}}{2\sqrt{t+2}} + \sum_{t=1}^{T-1} \Big(\frac{\sqrt{t+2}(t+1)}{2c} - \frac{\sqrt{t+1}(t)}{2c}\Big)\mathbb{E}\|x_t - x_\star\|_2^2 \\
&\leq \frac{\sqrt{2}}{2c(T)(T+1)} D + \frac{2}{(T)(T+1)} \sum_{t=0}^{T-1} \sqrt{t+1} \frac{cC'_{\mathcal{D}}}{2} + \sum_{t=1}^{T-1} \frac{\sqrt{t+1}}{2c} \Big(\frac{3t+2}{\sqrt{(t+2)(t+1)} + t}\Big) D \\
&\leq \frac{\sqrt{2}}{2c(T)(T+1)} D + \frac{2}{(T)(T+1)} (T+1)^{\frac{3}{2}} \Big(\frac{cC'_{\mathcal{D}}}{2} + \frac{3D}{2c}\Big) \leq \frac{2\sqrt{2}C'_{\mathcal{D}} + c^{-1}4\sqrt{2}D}{\sqrt{T}} \qquad (7.53)
\end{aligned}
$$

□

7.11 Supplementary E: Proof of Lemma 37

Proof. Proof. We start by proving (7.22). Since $\nabla g_w(x)$ is Lipschitz continuous with constant $L_{g,1}$, then every $\partial[g_w]_k/\partial[x]_j(x)$ is Lipschitz continuous with constant $L_{g,1}$ for every $1 \le k \le d$ and $1 \le j \le p$. It follows from Definition 2 that for any $w \in \Omega_w$,

$$\max_{\substack{1\le j\le p\\ 1\le k\le d}} \{|\frac{\partial[g_w]_k}{\partial[x]_j}(x) - \frac{\partial[g_w]_k}{\partial[x]_j}(\tilde{x})|\} \le L_{g,1}\|x-\tilde{x}\|_2. \tag{7.54}$$

It also follows from Definition 2 that $\mathrm{diam}(\mathcal{D}) < \infty$. Consequently, we can find a set $\Gamma \subset \mathbb{R}^p$ with cardinality $|\Gamma| \le \big(\frac{2\mathrm{diam}(\mathcal{D})}{\epsilon/\sqrt{p}}\big)^p$ such that for any $x \in \mathcal{D}$, there exists $z \in \Gamma$ with $\|x-z\|_2 \le \epsilon$, and hence

$$\begin{aligned}
|[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}| =& |\frac{1}{n}\sum_{i=1}^n \frac{\partial[g_{w_i}]_k}{\partial[x]_j}(x) - \mathbb{E}_w \frac{\partial[g_w]_k}{\partial[x]_j}(x)| \\
\le& |\frac{1}{n}\sum_{i=1}^n \frac{\partial[g_{w_i}]_k}{\partial[x]_j}(z) - \mathbb{E}_w \frac{\partial[g_w]_k}{\partial[x]_j}(z)| + 2\epsilon L_{g,1} \\
=& |[\overline{\nabla g}(z;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(z)]_{kj}| + 2\epsilon L_{g,1}.
\end{aligned}$$

Fixing $\delta > 0$ and $0 < \epsilon < \min\{\mathrm{diam}(\mathcal{D}), \frac{\delta}{2L_{g,1}}\}$, we have

$$\begin{aligned}
\mathbb{P}\Big\{|[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}| \ge \delta\Big\} &\le \mathbb{P}\Big\{\max_{z\in\Gamma} |[\overline{\nabla g}(z;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(z)]_{kj}| + 2\epsilon L_{g,1} \ge \delta\Big\} \\
&\le \sum_{z\in\Gamma} \mathbb{P}\Big\{|[\overline{\nabla g}(z;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(z)]_{kj}| \ge \delta - 2\epsilon L_{g,1}\Big\}.
\end{aligned}$$

By Assumption 4, $|\frac{\partial[g_{w_i}]_k}{\partial[x]_j}(x)| < l_{g,1}$ for every $x \in \mathcal{D}$, $1 \le k \le d$ and $1 \le j \le p$. Therefore, by

applying Azuma-Hoeffding inequality and using the cardinality bound of Γ, we have

$$\mathbb{P}\Big\{|[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}| \geq \delta\Big\} \leq \sum_{z\in\Gamma} 2\exp\{-\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2}\}$$
$$\leq 2\big(\frac{2\mathrm{diam}(\mathcal{D})}{\epsilon/\sqrt{p}}\big)^p \exp\{-\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2}\}. \quad (7.55)$$

Noticing that $\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}| \leq 2l_{g,1}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}|^4\Big] \leq (2l_{g,1})^4 \mathbb{P}\{\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}|| \geq \delta\} \\
&\quad + \delta^4 \mathbb{P}\{\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}| < \delta\} \\
\leq& 32 l_{g,1}^4 \big(\frac{2\mathrm{diam}(\mathcal{D})}{\epsilon/\sqrt{p}}\big)^p \exp\{-\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2}\} + \delta^4 \\
=& 32 l_{g,1}^4 (2\mathrm{diam}(\mathcal{D}))^p p^{p/2} \exp\{-\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2} + p\log(\frac{1}{\epsilon})\} + \delta^4, \quad (7.56)
\end{aligned}$$

where the second inequality above follows from (7.55). Letting $\delta = \frac{\sqrt{2}l_{g,1}\sqrt{4(p+1)\log(4n^2)}}{\sqrt{n}}$ and $\epsilon = \frac{\sqrt{2}l_{g,1}}{2L_{g,1}\sqrt{n}}$, we have $\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2} = (\sqrt{4(p+1)\log(4n^2)} - 1)^2$. Note that $(x-1)^2 \geq x^2/4$ for all $x \geq 2$. Since, $\sqrt{4(p+1)\log(4n^2)} \geq 2$ for $p \geq 1$, $n \geq 1$, we have

$$\frac{n(\delta-2\epsilon L_{g,1})^2}{2l_{g,1}^2} = (\sqrt{4(p+1)\log(4n^2)} - 1)^2 \geq (p+1)\log(4n^2).$$

Hence

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}|^4\Big] \\
\leq& 32 l_{g,1}^4 (2\mathrm{diam}(\mathcal{D}))^p p^{p/2} \exp\{-(p+1)\log(4n^2) + p\log\sqrt{n} + p\log(\sqrt{2}L_{g,1}/l_{g,1})\} \\
&+ \frac{64 l_{g,1}^4 (p+1)^2 \{\log(4n^2)\}^2}{n^2}.
\end{aligned}$$

Since $\log(4n^2) > \log\sqrt{n}$ for every $n \geq 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{x\in\mathcal{D}} |[\overline{\nabla g}(x;1,n)]_{kj} - [\mathbb{E}_w \nabla g_w(x)]_{kj}|^4\right] \\
&\leq 32 l_{g,1}^4 (2\mathrm{diam}(\mathcal{D}))^p p^{p/2} \exp\{-\log(4n^2) + p\log(\sqrt{2}L_{g,1}/l_{g,1})\} + \frac{64 l_{g,1}^4 (p+1)^2 \{\log(4n^2)\}^2}{n^2} \\
&= \frac{8 l_{g,1}^{4-p} (4\mathrm{diam}(\mathcal{D})^p) p^{p/2} L_{g,1}^p}{n^2} + \frac{64 l_{g,1}^4 (p+1)^2 \{\log(4n^2)\}^2}{n^2} \leq C_1 \frac{(\log(4n))^2}{n^2}
\end{aligned}$$

where $C_1 = 8 l_{g,1}^{4-p} (4\mathrm{diam}(\mathcal{D})^p) p^{p/2} L_{g,1}^p + 64 l_{g,1}^4 (p+1)^2$.

To prove (7.23), we notice that $g_w(x)$ is Lipschitz continuous with constant $L_{g,0}$ and for all $x \in \mathcal{D}$, $|[g_w]_k| \leq l_{g,0}$. Therefore, we can apply exactly the same argument to derive (7.23). Finally, (7.24) can be proved in the same way.

□

7.12 Supplementary F: Proof of Lemma 38

Proof. Proof. Note that

$$[J]_{ij}(x,y,z) = z_{:ij}^\top \nabla f_v(y) + [x]_{i:} \nabla^2 f_v(y) [x]_{:j} = \sum_{k=1}^d \left([z]_{kij} \frac{\partial f_{v_1}}{\partial [y]_k}(y) + [x]_{ki} \Big(\sum_{h=1}^d \frac{\partial f_{v_1}}{\partial [y]_k \partial [y]_h}(y) [x]_{hj} \Big) \right).$$

We can then compute each component of the gradient $\nabla[J]_{ij}(x,y,z) \in \mathbb{R}^{(d\times p)\times d\times(d\times p\times p)}$ as

$$\begin{aligned}
\frac{\partial[J]_{ij}}{\partial[x]_{k'j'}}(x,y,z) &= \delta_{ij'}\sum_{h=1}^{d}\frac{\partial f_{v_1}}{\partial[y]_{k'}\partial[y]_h}(y)[x]_{hj} + \delta_{jj'}\sum_{k=1}^{d}\frac{\partial f_{v_1}}{\partial[y]_k\partial[y]_{k'}}(y)[x]_{ki} \\
&= \delta_{ij'}[\nabla^2 f_{v_1}]_{k':}(y)[x]_{:j} + \delta_{jj'}[\nabla^2 f_{v_1}]_{k':}(y)x_{:i} \\
\frac{\partial[J]_{ij}}{\partial[y]_{h'}}(x,y,z) &= \sum_{k=1}^{d}\left([z]_{kij}\frac{\partial f_{v_1}}{\partial[y]_k\partial[y]_{h'}}(y) + [x]_{ki}\Big(\sum_{h=1}^{d}\frac{\partial f_{v_1}}{\partial[y]_k\partial[y]_h\partial[y]_{h'}}(y)[x]_{hj}\Big)\right) \\
&= [z]_{:ij}^{\top}[\nabla^2 f_{v_1}(y)]_{:h'} + [x]_{:i}^{\top}[\nabla^3 f_{v_1}(y)]_{::h'}[x]_{:j} \\
\frac{\partial[J]_{ij}}{\partial[z]_{k''i''j''}}(x,y,z) &= \delta_{ii''}\delta_{jj''}\frac{\partial f_{v_1}}{\partial[y]_{k''}}(y) = \delta_{ii''}\delta_{jj''}[\nabla f_{v_1}(y)]_{k''}.
\end{aligned}$$

where $1 \le i', j', i'', j'' \le p, 1 \le k', h', k'' \le d$ and δ_{ij} is the Kronecker delta. Note that by Assumptions 1, 2, 4, and 5, we have

$$\begin{aligned}
&|\frac{\partial[J]_{ij}}{\partial[x]_{k'j'}}(x_1,y_1,z_1) - \frac{\partial[J]_{ij}}{\partial[x]_{k'j'}}(x_2,y_2,z_2)| \\
&\le \delta_{ij'}|[\nabla^2 f_{v_1}]_{k':}(y_1)[x_1]_{:j} - [\nabla^2 f_{v_1}]_{k':}(y_2)[x_2]_{:j}| + \delta_{jj'}|[\nabla^2 f_{v_1}]_{k':}(y_1)[x_1]_{:i} - [\nabla^2 f_{v_1}]_{k':}(y_2)[x_2]_{:i}| \\
&\le \delta_{ij'}\sqrt{d}\{l_{f,2}\|[x_1]_{:j} - [x_2]_{:j}\|_2 + L_{f,2}l_{g,1}\|y_1 - y_2\|_2\} + \delta_{jj'}\sqrt{d}\{l_{f,2}\|[x_1]_{:i} - [x_2]_{:i}\|_2 + L_{f,2}l_{g,1}\|y_1 - y_2\|_2\} \\
&= (\delta_{ij'} + \delta_{jj'})\sqrt{d}L_{f,2}l_{g,1}\|y_1 - y_2\|_2 + \delta_{ij'}\sqrt{d}l_{f,2}\|[x_1]_{:j} - [x_2]_{:j}\|_2 + \delta_{jj'}\sqrt{d}l_{f,2}\|[x_1]_{:i} - [x_2]_{:i}\|_2
\end{aligned}$$

$$\begin{aligned}
&|\frac{\partial[J]_{ij}}{\partial[y]_{h'}}(x_1,y_1,z_1) - \frac{\partial[J]_{ij}}{\partial[y]_{h'}}(x_2,y_2,z_2)| \\
&\le |[z_1]_{:ij}^{\top}[\nabla^2 f_{v_1}(y_1)]_{:h'} - [z_2]_{:ij}^{\top}[\nabla^2 f_{v_1}(y_2)]_{:h'}| + |[x_1]_{:i}^{\top}[\nabla^3 f_{v_1}(y_1)]_{::h'}[x_1]_{:j} - [x_2]_{:i}^{\top}[\nabla^3 f_{v_1}(y_2)]_{::h'}[x_2]_{:j}| \\
&\le \sqrt{d}l_{g,2}L_{f,2}\|y_1 - y_2\|_2 + \sqrt{d}l_{f,2}L_{g,2}\|[z_1]_{:ij} - [z_2]_{:ij}\|_2 + dl_{g,1}^2L_{f,3}\|y_1 - y_2\|_2 \\
&\quad + dl_{g,1}l_{f,3}\|[x_1]_{:j} - [x_2]_{:j}\|_2 + dl_{g,1}l_{f,3}\|[x_1]_{:i} - [x_2]_{:i}\|_2 \\
&= (\sqrt{d}l_{g,2}L_{f,2} + dl_{g,1}^2L_{f,3})\|y_1 - y_2\|_2 + \sqrt{d}l_{f,2}L_{g,2}\|[z_1]_{:ij} - [z_2]_{:ij}\|_2 \\
&\quad + dl_{g,1}l_{f,3}\|[x_1]_{:j} - [x_2]_{:j}\|_2 + dl_{g,1}l_{f,3}\|[x_1]_{:i} - [x_2]_{:i}\|_2
\end{aligned}$$

$$|\frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_1, y_1, z_1) - \frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_2, y_2, z_2)| \leq |\delta_{ii''}\delta_{jj''}[\nabla f_{v_1}(y_1)]_{k''} - \delta_{ii''}\delta_{jj''}[\nabla f_{v_1}(y_1)]_{k''}|$$
$$\leq \delta_{ii''}\delta_{jj''}L_{f,1}\|y_1 - y_2\|_2.$$

Note that

$$\begin{aligned}
&\|\nabla [J]_{ij}(x_1, y_1, z_1) - \nabla [J]_{ij}(x_2, y_2, z_2)\|_F^2 \\
&= \sum_{k'=1}^{d}\sum_{j'=1}^{p} |\frac{\partial [J]_{ij}}{\partial [x]_{k'j'}}(x_1, y_1, z_1) - \frac{\partial [J]_{ij}}{\partial [x]_{k'j'}}(x_2, y_2, z_2)|^2 + \sum_{h'=1}^{d} |^2\frac{\partial [j]_{ij}}{\partial [y]_{h'}}(x_1, y_1, z_1) - \frac{\partial [j]_{ij}}{\partial [y]_{h'}}(x_2, y_2, z_2)|^2 \\
&\quad + \sum_{k''=1}^{d}\sum_{i''=1}^{p}\sum_{j''=1}^{p} |\frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_1, y_1, z_1) - \frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_2, y_2, z_2)|^2.
\end{aligned}$$

Then based on our previous computation and using (7.14), we have

$$\begin{aligned}
&\sum_{k'=1}^{d}\sum_{j'=1}^{p} |\frac{\partial [J]_{ij}}{\partial [x]_{k'j'}}(x_1, y_1, z_1) - \frac{\partial [J]_{ij}}{\partial [x]_{k'j'}}(x_2, y_2, z_2)|^2 \\
&\leq \sum_{k'=1}^{d}\sum_{j'=1}^{p} 3\{(2\delta_{ij'} + 2\delta_{jj'})dL_{f,2}^2 l_{g,1}^2\|y_1 - y_2\|_2^2 + \delta_{ij'}dl_{f,2}^2\|[x_1]_{:j} - [x_2]_{:j}\|_2^2 + \delta_{jj'}dl_{f,2}^2\|[x_1]_{:i} - [x_2]_{:i}\|_2^2\} \\
&= 12d^2L_{f,2}^2 l_{g,1}^2\|y_1 - y_2\|_2^2 + d^2l_{f,2}^2\|[x_1]_{:j} - [x_2]_{:j}\|_2^2 + d^2l_{f,2}^2\|[x_1]_{:i} - [x_2]_{:i}\|_2^2\},
\end{aligned}$$

$$\begin{aligned}
&\sum_{h'=1}^{d} |\frac{\partial [j]_{ij}}{\partial [y]_{h'}}(x_1, y_1, z_1) - \frac{\partial [j]_{ij}}{\partial [y]_{h'}}(x_2, y_2, z_2)|^2 \\
&\leq 4d(\sqrt{d}l_{g,2}L_{f,2} + dl_{g,1}^2 L_{f,3})^2\|y_1 - y_2\|_2^2 + 4d^2l_{f,2}^2L_{g,2}^2\|[z_1]_{:ij} - [z_2]_{:ij}\|_2^2 \\
&\quad + 4d^3l_{f,3}^2\|[x_1]_{:j} - [x_2]_{:j}\|_2^2 + 4d^3l_{g,1}^2l_{f,3}^2\|[x_1]_{:i} - [x_2]_{:i}\|, \text{ and}
\end{aligned}$$

$$\sum_{k''=1}^{d}\sum_{i''=1}^{p}\sum_{j''=1}^{p} |\frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_1, y_1, z_1) - \frac{\partial [J]_{ij}}{\partial [z]_{k''i''j''}}(x_2, y_2, z_2)|^2 \leq dL_{f,1}^2\|y_1 - y_2\|_2^2.$$

Therefore

$$
\begin{aligned}
&\|\nabla[J]_{ij}(x_1,y_1,z_1)-\nabla[J]_{ij}(x_2,y_2,z_2)\|_F^2\\
&\le \{12d^2L_{f,2}^2l_{g,1}^2+4d(\sqrt{d}_{g,2}L_{f,2}+d^2l_{g,1}^2L_{f,3})^2+dL_{f,1}^2\}\|y_1-y_2\|_2^2+4d^2l_{f,2}^2L_{g,2}^2\|[z_1]_{:ij}-[z_2]_{:ij}\|_2^2\\
&\quad+(d^2l_{f,2}^2+4d^3l_{f,3}^2)\|[x_1]_{:j}-[x_2]_{:j}\|_2^2+(d^2l_{f,2}^2+4d^3l_{f,3}^2)\|[x_1]_{:i}-[x_2]_{:i}\|_2^2\\
&\le \{12d^2L_{f,2}^2l_{g,1}^2+4d(\sqrt{d}_{g,2}L_{f,2}+d^2l_{g,1}^2L_{f,3})^2+dL_{f,1}^2+4d^2l_{f,2}^2L_{g,2}^2\\
&\quad+2d^2l_{f,2}^2+4d^3l_{f,3}^2\}\|\mathrm{vec}(x_1-x_2,y_1-y_2,z_1-z_2)\|_2^2\\
&=L_J^2\|\mathrm{vec}(x_1-x_2,y_1-y_2,z_1-z_2)\|_2^2
\end{aligned}
$$

□

7.13 Supplementary G: Proof of Lemma 40

Proof. Proof. Fixing $v_1\in\Omega_v$ and $x,\tilde{x}\in\mathcal{D}$, we have

$$
W(x,v_1)-W(\tilde{x},v_1)=\frac{1}{\tilde{p}_N}\Big(Y_1(x)-Y_1(\tilde{x})-\frac{1}{2}\Big(Y_2(x)-Y_2(\tilde{x})+Y_3(x)-Y_3(\tilde{x})\Big)\Big)+Y_4(x)-Y_4(\tilde{x}).
$$

Similar to the proof of Proposition 2, we first take expectation with respect to N. Then,

$$
\begin{aligned}
&\mathbb{E}\|W(x,v_1)-W(\tilde{x},v_1)\|_2^2=\sum_{n=0}^{\infty}\mathbb{E}\{\|W(x,v_1)-W(\tilde{x},v_1)\|_2^2|N=n\}\tilde{p}_n\\
&=\sum_{n=0}^{\infty}\sum_{i=1}^{p}\mathbb{E}\big(\{[W(x,v_1)]_i-[W(\tilde{x},v_1)]_i\}^2|N=n\big)\tilde{p}_n\le\sum_{i=1}^{p}2\mathbb{E}\{[Y_4(x)]_i-[Y_4(\tilde{x})]_i\}^2\\
&\quad+\sum_{n=0}^{\infty}\sum_{i=1}^{p}\frac{2}{\tilde{p}_n}\mathbb{E}\{\big([Y_1(x)]_i-[Y_1(\tilde{x})]_i-0.5\{[Y_2(x)]_i-[Y_2(\tilde{x})]_i+[Y_3(x)]_i-[Y_3(\tilde{x})]_i\}\big)^2|N=n\},
\end{aligned}
$$

where the last inequality comes from (7.14). Since $[Y_4(\cdot)]_i$ and $[Y_1(\cdot)]_i-0.5\{[Y_2(\cdot)]_i+[Y_3(\cdot))]_i\}$ are continuous for every $1\le i\le p$. By the mean value theorem, there exist ζ_i and ξ_i that lie

between x and $\tilde{x}$ such that $[Y_4(x)]_i - [Y_4(\tilde{x})]_i = \nabla[Y_4(\zeta_i)]_i^\top (x - \tilde{x})$ and

$$\begin{aligned}
&\left([Y_1(x)]_i - 0.5\{[Y_2(x)]_i + [Y_3(x)]_i\}\right) - \left([Y_1(\tilde{x})]_i - 0.5\{[Y_2(\tilde{x})]_i + [Y_3(\tilde{x})]_i\}\right) \\
&= \{\nabla\left([Y_1(\xi_i)]_i - 0.5\{[Y_2(\xi_i)]_i + [Y_3(\xi_i)]_i\}\right)\}^\top (x - \tilde{x}).
\end{aligned}$$

Therefore, we may write

$$\begin{aligned}
&\mathbb{E}\|W(x, v_1) - W(\tilde{x}, v_1)\|_2^2 = \sum_{i=1}^{p} 2\mathbb{E}\{\nabla[Y_4(\zeta_i)]_i^\top (x - \tilde{x})\}^2 \\
&\quad + \sum_{n=0}^{\infty}\sum_{i=1}^{p} \frac{2}{\tilde{p}_n}\mathbb{E}\left(\left\{\nabla\left([Y_1(\xi_i)]_i - 0.5\{[Y_2(\xi_i)]_i + [Y_3(\xi_i)]_i\}\right)^\top (x - \tilde{x})\right\}^2 | N = n\right) \\
&\leq \sum_{i=1}^{p} 2\|x - \tilde{x}\|_2^2 \mathbb{E}\|\nabla[Y_4(\zeta_i)]_i\|_2^2 \\
&\quad + \sum_{n=0}^{\infty}\sum_{i=1}^{p} \frac{2\|x - \tilde{x}\|_2^2}{\tilde{p}_n}\mathbb{E}\left\{\|\nabla\left([Y_1(\xi_i)]_i - 0.5\{[Y_2(\xi_i)]_i + [Y_3(\xi_i)]_i\}\right)\|_2^2 | N = n\right\} \\
&= \sum_{i=1}^{p} 2\|x - \tilde{x}\|_2^2 \mathbb{E}\|\nabla[Y_4(\zeta_i)]_i\|_2^2 \\
&\quad + \sum_{n=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{2\|x - \tilde{x}\|_2^2}{\tilde{p}_n}\mathbb{E}\left\{\|[\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\}\|_2^2 | N = n\right\} \qquad (7.57)
\end{aligned}$$

where the last inequality uses the Cauchy-Shwartz inequality. Next, we first obtain an upper bound for $\mathbb{E}\|\nabla[Y_4(\zeta_i)]_i\|_2^2$ and then bound $\mathbb{E}\left\{\|\nabla\left([Y_1(\xi_i)]_i - 0.5\{[Y_2(\xi_i)]_i + [Y_3(\xi_i)]_i\}\right)\|_2^2 | N = n\right\}$ using a function of n in order to analyze the infinite sum above.

To obtain an upper bound for $\mathbb{E}\|[\nabla Y_4(\zeta_i)]_i\|_2^2$, we first note that

$$\begin{aligned}
&\nabla\{[Y_4(\zeta_i)]_i\} \\
&= \{[\overline{\nabla^2 g}(x; 1, 2^{n_0})]_{::i}\}^\top \nabla f_{v_1}\{\bar{g}(x; 1, 2^{n_0})\} + \{\overline{\nabla g}(x; 1, 2^{n_0})\}^\top \nabla^2 f_{v_1}(\bar{g}(x; 1, 2^{n_0}))[\overline{\nabla g}(x; 1, 2^{n_0})]_{i:}.
\end{aligned}$$

Therefore by (7.14),

$$\begin{aligned}\|\nabla\{[Y_4(\zeta_i)]_i\}\|_2^2 &\le 2\|\{[\overline{\nabla^2 g}(x;1,2^{n_0})]_{::i}\}^\top \nabla f_{v_1}\{\bar{g}(x;1,2^{n_0})\}\|_2^2 \\ &\quad + 2\|\{\overline{\nabla g}(x;1,2^{n_0})\}^\top \nabla^2 f_{v_1}(\bar{g}(x;1,2^{n_0}))[\overline{\nabla g}(x;1,2^{n_0})]_{i:}\|_2^2 \\ &\le 2\|[\overline{\nabla^2 g}(x;1,2^{n_0})]_{::i}\|_F^2 \|\nabla f_{v_1}\{\bar{g}(x;1,2^{n_0})\}\|_2^2 \\ &\quad + 2\|\overline{\nabla g}(x;1,2^{n_0})\|_F^2 \|\nabla^2 f_{v_1}(\bar{g}(x;1,2^{n_0}))\|_F^2 \|[\overline{\nabla g}(x;1,2^{n_0})]_{i:}\|_2^2.\end{aligned}$$

By Assumptions 4 and 5,

$$\begin{aligned}&\|[\overline{\nabla^2 g}(\zeta_i;1,2^{n_0})]_{::i}\|_F^2 \le pd\|[\overline{\nabla^2 g}(\zeta_i;1,2^{n_0})]_{::i}\|_\infty^2 \le pdl_{g,2}^2, \\ &\|\nabla f_{v_1}\{\bar{g}(\zeta_i;1,2^{n_0})\}\|_2^2 \le p\|f_{v_1}\{\bar{g}(\zeta_i;1,2^{n_0})\}\|_\infty^2 \le dl_{f,1}^2, \\ &\|\overline{\nabla g}(\zeta_i;1,2^{n_0})\|_F^2 \le pd\|\overline{\nabla g}(\zeta_i;1,2^{n_0})\|_\infty^2 \le pdl_{g,1}^2, \\ &\|\nabla^2 f_{v_1}(\bar{g}(\zeta_i;1,2^{n_0}))\|_F^2 \le d^2\|\nabla^2 f_{v_1}(\bar{g}(\zeta_i;1,2^{n_0}))\|_\infty^2 \le d^2 l_{f,2}^2, \text{ and} \\ &\|[\overline{\nabla g}(\zeta_i;1,2^{n_0})]_{i:}\|_2^2 \le d\|[\overline{\nabla g}(\zeta_i;1,2^{n_0})]_{i:}\|_\infty^2 \le dl_{g,1}^2.\end{aligned}$$

Therefore

$$\|\nabla\{[Y_4(\zeta_i)]_i\}\|_2^2 \le 2pd^2 f_{g,2}^2 l_{f,1}^2 + 2pd^4 l_{g,1}^4 l_{f,2}^2.$$

Hence

$$2\|x-\tilde{x}\|_2^2 \mathbb{E}\|\nabla[Y_4(\zeta_i)]_i\|_2^2 \le \{4pd^2 f_{g,2}^2 l_{f,1}^2 + 4pd^4 l_{g,1}^4 l_{f,2}^2\}\|x-\tilde{x}\|_2^2. \tag{7.58}$$

To bound the second term in (7.57), we let $\bar{n}_0 = n + n_0$ and $\bar{n}_0^+ = n + n_0 + 1$ and note that

conditioned on $N = n$,

$$
\begin{aligned}
[\nabla Y_1(\xi_i)]_{ij} &= [J]_{ij}\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0^+}), \bar{g}(\xi_i; 1, 2^{\bar{n}_0^+}), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0^+})\} \\
&= [J]_{ij}\{\mathbb{E}_w \nabla g_w(\xi_i), \mathbb{E}_w g_w(\xi_i), \mathbb{E}_w \nabla^2 g_w(\xi_i)\} \\
&\quad + \nabla[J]_{ij}\{\mathbb{E}_w \nabla g_w(\xi_i), \mathbb{E}_w g_w(\xi_i), \mathbb{E}_w \nabla^2 g_w(\xi_i)\}[\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(\xi_i), \\
&\quad \bar{g}(\xi_i; 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla^2 g_w(\xi_i)] + \\
&\quad R\big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\big\},
\end{aligned}
$$

$$
\begin{aligned}
[\nabla Y_2(\xi_i)]_{ij} &= [J]_{ij}\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0}), \bar{g}(\xi_i; 1, 2^{\bar{n}_0}), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0})\} \\
&= [J]_{ij}\{\mathbb{E}_w \nabla g_w(\xi_i), \mathbb{E}_w g_w(\xi_i), \mathbb{E}_w \nabla^2 g_w(\xi_i)\} \\
&\quad + \nabla[J]_{ij}\{\mathbb{E}_w \nabla g_w(\xi_i), \mathbb{E}_w g_w(\xi_i), \mathbb{E}_w \nabla^2 g_w(\xi_i)\}[\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0}) - \mathbb{E}_w \nabla g_w(\xi_i), \\
&\quad \bar{g}(\xi_i; 1, 2^{\bar{n}_0}) - \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0}) - \mathbb{E}_w \nabla^2 g_w(\xi_i)] + \\
&\quad R\big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\big\}, \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
[\nabla Y_3(\xi_i)]_{ij} &= [J]_{ij}\{\overline{\nabla g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \bar{g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \overline{\nabla^2 g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+})\} \\
&\quad + \nabla[J]_{ij}\{\mathbb{E}_w \nabla g_w(\xi_i), \mathbb{E}_w g_w(\xi_i), \mathbb{E}_w \nabla^2 g_w(\xi_i)\}[\overline{\nabla g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla g_w(\xi_i), \\
&\quad \bar{g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w g_w(\xi_i),\ \overline{\nabla^2 g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}) - \mathbb{E}_w \nabla^2 g_w(\xi_i)] \\
&\quad + R\{\overline{\nabla g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w \nabla g_w(\xi_i),\ \bar{g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w g_w(\xi_i), \\
&\quad \overline{\nabla^2 g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w \nabla^2 g_w(\xi_i)\}.
\end{aligned}
$$

Therefore, condition on $N = n$, we have

$$
\begin{aligned}
&[\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\} \\
&= R\Big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\Big\} \\
&- \frac{1}{2} R\Big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\Big\} \\
&- \frac{1}{2} R\{\overline{\nabla g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w \nabla g_w(\xi_i),\ \bar{g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w \nabla^2 g_w(\xi_i)\}.
\end{aligned}
$$

Then, by (7.14)

$$
\begin{aligned}
&\mathbb{E}\Big\{\Big([\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\}\Big)^2 \Big| N = n\Big\} \\
&\le 3\mathbb{E}\Big(R\Big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\Big\}^2\Big) \\
&\quad + \frac{3}{4}\mathbb{E}\Big(R\Big\{\overline{\nabla g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla g_w(\xi_i), \bar{g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 1, 2^{\bar{n}_0}), \mathbb{E}_w \nabla^2 g_w(\xi_i)\Big\}^2\Big) \\
&+ \frac{3}{4}\mathbb{E}\Big(R\{\overline{\nabla g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}), \mathbb{E}_w \nabla g_w(\xi_i),\ \bar{g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w g_w(\xi_i), \overline{\nabla^2 g}(\xi_i; 2^{\bar{n}_0} + 1, 2^{\bar{n}_0^+}),\ \mathbb{E}_w \nabla^2 g_w(\xi_i)\}^2
\end{aligned}
$$

Now, applying Lemma 39 on the three terms on the right-hand-side of the inequality above,

$$
\begin{aligned}
&\mathbb{E}\Big\{\Big([\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\}\Big)^2 \Big| N = n\Big\} \\
&\le \frac{3L_J^2}{4}\{\mathbb{E}\|\overline{\nabla g}(\xi_i; 1, 2^{n_0+n+1}) - \mathbb{E}_w \nabla g_w(\xi_i)\|_F^4 + \mathbb{E}\|\bar{g}(\xi_i; 1, 2^{n_0+n+1}) - \mathbb{E}_w g_w(\xi_i)\|_F^4 \\
&\quad + \mathbb{E}\|\overline{\nabla^2 g}(\xi_i; 1, 2^{n_0+n+1}) - \mathbb{E}_w g_w(\xi_i)\|_F^4\} + \frac{3L_J^2}{16}\{\mathbb{E}\|\overline{\nabla g}(\xi_i; 1, 2^{n_0+n}) - \mathbb{E}_w \nabla g_w(\xi_i)\|_F^4 + \\
&\quad \mathbb{E}\|\bar{g}(\xi_i; 1, 2^{n_0+n}) - \mathbb{E}_w g_w(\xi_i)\|_F^4 + \mathbb{E}\|\overline{\nabla^2 g}(\xi_i; 1, 2^{n_0+n}) - \mathbb{E}_w \nabla^2 g_w(\xi_i)\|_F^4\} + \\
&\quad \frac{3L_J^2}{16}\{\mathbb{E}\|\overline{\nabla g}(\xi_i; 2^{n_0+n} + 1, 2^{n_0+n+1}) - \mathbb{E}_w \nabla g_w(\xi_i)\|_F^4 + \mathbb{E}\|\bar{g}(\xi_i; 2^{n_0+n} + 1, 2^{n_0+n+1}) - \mathbb{E}_w g_w(\xi_i)\|_F^4 \\
&+ \mathbb{E}\|\overline{\nabla^2 g}(\xi_i; 2^{n_0+n} + 1, 2^{n_0+n+1}) - \mathbb{E}_w \nabla^2 g_w(\xi_i)\|_F^4\}.
\end{aligned}
$$

Then using Lemma 37 to the right-hand-side of the above inequality, we have

$$
\begin{aligned}
&\mathbb{E}\Big\{\Big([\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\}\Big)^2 \big| N = n\Big\} \\
&\leq \frac{3L_J^2}{4}\big((C_0 + C_1 + C_2)\frac{\{\log(4^{n+n_0+1})\}^2}{4^{n+n_0+1}}\big) + \frac{3L_J^2}{8}\big((C_0 + C_1 + C_2)\frac{\{\log(4^{n_0+n})\}^2}{4^{n_0+n}}\big) \\
&= \frac{3L_J^2(C_0 + C_1 + C_2)}{4^{n_0+n+1}}\{\frac{1}{4}\{\log(4^{n+n_0+1})\}^2 + \frac{1}{2}\{\log(4^{n_0+n})\}^2\} \\
&\leq \frac{3L_J^2(C_0 + C_1 + C_2)}{4^{n_0+n+1}}\{\frac{3}{2}(n_0 + n + 1)^2\}, \qquad (7.59)
\end{aligned}
$$

where the last inequality is the result of $\log 4 < 2$.

Now we are ready to obtain a bound for (7.57). Using (7.58) and (7.59), we have

$$
\begin{aligned}
&\mathbb{E}\|W(x, v_1) - W(\tilde{x}, v_1)\|_2^2 \\
&\leq \sum_{i=1}^{p} 2\|x - \tilde{x}\|_2^2 \mathbb{E}\|\nabla[Y_4(\zeta_i)]_i\|_2^2 \\
&+ \sum_{n=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{2\|x - \tilde{x}\|_2^2}{\tilde{p}_n}\mathbb{E}\Big\{\|[\nabla Y_1(\xi_i)]_{ij} - 0.5\{[\nabla Y_2(\xi_i)]_{ij} + [\nabla Y_3(\xi_i)]_{ij}\}\|_2^2 | N = n\Big\} \\
&\leq \sum_{i=1}^{p}\{4pd^2 f_{g,2}^2 l_{f,1}^2 + 4pd^4 l_{g,1}^4 l_{f,2}^2\}\|x - \tilde{x}\|_2^2 \\
&\quad + \sum_{n=0}^{\infty}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{2\|x - \tilde{x}\|_2^2}{\tilde{p}_n}\frac{3L_J^2(C_0 + C_1 + C_2)}{4^{n_0+n+1}}\{\frac{3}{2}(n_0 + n + 1)^2\} \\
&= \|x - \tilde{x}\|_2^2\Big\{4p^2d^2 f_{g,2}^2 l_{f,1}^2 + 4p^2 d^4 l_{g,1}^4 l_{f,2}^2 + 9L_J^2 p^2(C_0 + C_1 + C_2)\sum_{n=0}^{\infty}\frac{(n_0 + n + 1)^2}{\tilde{p}_n 4^{n_0+n+1}}\Big\}.
\end{aligned}
$$

Since $\tilde{p}_n = (1 - 0.5^{\gamma})0.5^{\gamma n}$ and $1 < \gamma < 2$, we have

$$
\begin{aligned}
\sum_{n=0}^{\infty}\frac{(n_0 + n + 1)^2}{\tilde{p}_n 4^{n_0+n+1}} &= \frac{1}{(1 - 0.5^{\gamma})4^{n_0+1}}\sum_{n=0}^{\infty}\frac{(n_0 + n + 1)^2}{2^{(2-\gamma)n}} \\
&= \frac{(n_0 + 1)^2}{1 - 2^{\gamma-2}} + \frac{2(n_0 + 1)2^{\gamma-2}}{(1 - 2^{\gamma-2})^2} + \frac{2^{3\gamma-6} + 2^{\gamma-2}}{(1 - 2^{\gamma-2})^3}.
\end{aligned}
$$

Therefore

$$\mathbb{E}\|W(x, v_1) - W(\tilde{x}, v_1)\|_2^2 \leq C_{\mathcal{D}}\|x - \tilde{x}\|_2^2,$$

where

$$\begin{aligned} C_{\mathcal{D}} = & 4p^2d^2f_{g,2}^2l_{f,1}^2 + 4p^2d^4l_{g,1}^4l_{f,2}^2 \\ & + 9L_J^2p^2(C_0 + C_1 + C_2)\big(\frac{(n_0+1)^2}{1-2^{\gamma-2}} + \frac{2(n_0+1)2^{\gamma-2}}{(1-2^{\gamma-2})^2} + \frac{2^{3\gamma-6}+2^{\gamma-2}}{(1-2^{\gamma-2})^3}\big). \end{aligned}$$

□

7.14 Supplementary H: Proof of Lemma 42

Proof. Proof. First we have

$$\begin{aligned} \mathbb{E}[\tilde{h}(x)] = \mathbb{E}[h_1(x)] = \mathbb{E}[\mathbb{E}[h_1(x)|\mathcal{I}]] =& \frac{1}{B}\mathbb{E}[\mathbb{E}[\sum_{v_i\in\mathcal{I}} \text{UnibasedGradient}(x, v_i)|\mathcal{I}]] \\ =& \frac{1}{B}\mathbb{E}[\sum_{v_i\in\mathcal{I}} \nabla(f_{v_i}(\mathbb{E}_w g_w(x)))] = \nabla F(x). \end{aligned}$$

Secondly, for any $v \in \Omega_v$, denote $W_i = \text{UnbiasedGradient}(x, v_i)$, $h_v = \nabla(f_v(\mathbb{E}_w g_w(x)))$ and $h(\mathcal{I}) = \mathbb{E}[h_1(x)|\mathcal{I}] = \frac{1}{B}\sum_{v_i\in\mathcal{I}} h_{v_i}$, we have

$$\begin{aligned} Var[\tilde{h}(x)] =& \mathbb{E}[Var[\tilde{h}(x)|\mathcal{I}]] + Var[\mathbb{E}[\tilde{h}(x)|\mathcal{I}]] = \frac{1}{K}\mathbb{E}[Var[h_1(x)|\mathcal{I}]] + Var_{\mathcal{I}}[h(\mathcal{I})] \\ =& \frac{1}{K}\mathbb{E}\big[\mathbb{E}[(h_1(x) - h(\mathcal{I}))^\intercal(h_1(x) - h(\mathcal{I}))|\mathcal{I}]\big] + \frac{1}{B}Var_v[h_v] \\ =& \frac{1}{KB^2}\mathbb{E}\big[\mathbb{E}[(\sum_{i=1}^B W_i - h_{v_i} + h_{v_i} - h(\mathcal{I}))^\intercal(\sum_{i=1}^B W_i - h_{v_i} + h_{v_i} - h(\mathcal{I}))|\mathcal{I}]\big] + \frac{1}{B}Var_v[h_v] \\ =& \frac{1}{KB^2}\mathbb{E}\big[\mathbb{E}[\sum_{i=1}^B \|W_i - h_{v_i}\|_2^2 + \sum_{i=1}^B\sum_{j=1}^B (h_{v_i} - h(\mathcal{I}))^\intercal(h_{v_j} - h(\mathcal{I}))|\mathcal{I}]\big] + \frac{1}{B}Var_v[h_v] \\ \leq& \frac{C'_{\mathcal{D}}}{KB} + 4pd^2l_{f,1}^2l_{g,1}^2(\frac{1}{K} + \frac{1}{B}) \end{aligned}$$

where the last inequality follows from the definition of $C'_{\mathcal{D}}$ and the fact that each component of h_v is bounded by $dl_{f,1}l_{g,1}$ for any $v \in \Omega_v$, according to the definition of $l_{\mathcal{D}}$ and h_v. The equality above it follows from the independence between the W_i's given $\mathcal{I}$.

□

References

[1] J. Blanchet, D. Goldfarb, G. Iyengar, F. Li, and C. Zhou, "Unbiased simulation for optimizing stochastic function compositions," *arXiv preprint arXiv:1711.07564*, 2017.

[2] H. Lam and F. Li, "Sampling uncertain constraints under parametric distributions," in *2018 Winter Simulation Conference (WSC)*, 2018, pp. 2072–2083.

[3] H. Lam and F. Li, "Parametric scenario optimization under limited data: A distributionally robust optimization view," *ACM Transactions on Modeling and Computer Simulation*, To appear.

[4] J. Blanchet, F. Li, and X. Li, "Unbiased sampling of multidimensional partial differential equations with random coefficients," *arXiv preprint arXiv:1806.03362*,

[5] F. Li, H. Lam, and S. Prusty, "Robust importance weighting for covariate shift," S. Chiappa and R. Calandra, Eds., ser. In proceeding of the 23rd International Conference of Artificial Intelligence and Statistics (AISTATS) 2020, Proceedings of Machine Learning Research, Online: PMLR, 2020, pp. 352–362.

[6] F. Li, H. Lam, H. Chen, and A. Meisami, "Constrained reinforcement learning via policy splitting," S. J. Pan and M. Sugiyama, Eds., ser. In proceeding of the 12th Asian Conference on Machine Learning (ACML) 2020, Proceedings of Machine Learning Research, Online: PMLR.

[7] H. Lam and F. Li, *General feasibility bounds for sample average approximation via vapnik-chervonenkis dimension*, 2021. arXiv: `2103.01324 [math.OC]`.

[8] R. P. Liu, "On feasibility of sample average approximation solutions," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 2026–2052, 2020.

[9] R. Chen and J. Luedtke, "On sample average approximation for two-stage stochastic programs without relatively complete recourse," *arXiv preprint arXiv:1912.13078*, 2019.

[10] Q. Cai, A. Filos-Ratsikas, P. Tang, and Y. Zhang, "Reinforcement mechanism design for e-commerce," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1339–1348.

[11] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie, "A reinforcement learning framework for explainable recommendation," in *2018 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2018, pp. 587–596.

[12] D. Wu, X. Chen, X. Yang, H. Wang, Q. Tan, X. Zhang, J. Xu, and K. Gai, "Budget constrained bidding by model-free reinforcement learning in display advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1443–1451.

[13] A. Prékopa, "Probabilistic programming," in *Handbooks in Operations Research& Management Science*, A. Ruszczynski and A. Shapiro, Eds., Amsterdam, Netherlands: Elsevier, 2003.

[14] M. R. Murr and A. Prékopa, "Solution of a product substitution problem using stochastic programming," in *Probabilistic Constrained Optimization*, U. Stanislav, Ed., Manhattan, New York: Springer, 2000, pp. 252–271.

[15] M. A. Lejeune and A. Ruszczynski, "An efficient trajectory method for probabilistic production-inventory-distribution problems," *Operations Research*, vol. 55, no. 2, pp. 378–394, 2007.

[16] A. Prékopa and T. Szántai, "Flood control reservoir system design using stochastic programming," in *Mathematical Programming in Use*, M. Balinski and C. Lemarechal, Eds., Manhattan, New York: Springer, 1978, pp. 138–151.

[17] A. Prékopa, T. Rapcsák, and I. Zsuffa, "Serially linked reservoir system design using stochastic programing," *Water Resources Research*, vol. 14, no. 4, pp. 672–678, 1978.

[18] Y. Shi, J. Zhang, and K. B. Letaief, "Optimal stochastic coordinated beamforming for wireless cooperative networks with csi uncertainty," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 960–973, 2015.

[19] L. J. Hong, J. Luo, and B. L. Nelson, "Chance constrained selection of the best," *INFORMS Journal on Computing*, vol. 27, no. 2, pp. 317–334, 2015.

[20] M. C. Campi and S. Garatti, "A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality," *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 257–280, 2011.

[21] A. Nemirovski and A. Shapiro, "Scenario approximations of chance constraints," in *Probabilistic and randomized methods for design under uncertainty*, Springer, 2006, pp. 3–47.

[22] M. C. Campi and S. Garatti, "The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1211–1230, 2008.

[23] D. P. De Farias and B. Van Roy, "On Constraint Sampling in the Linear Programming Approach to Approximate Dynamic Programming," *Mathematics of Operations Research*, vol. 29, no. 3, pp. 462–478, 2004.

[24] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 674–699, 2008.

[25] G. Schildbach, L. Fagiano, and M. Morari, "Randomized solutions to convex programs with multiple chance constraints," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2479–2501, 2013.

[26] M. C. Campi and S. Garatti, "Wait-and-Judge Scenario Optimization," *Mathematical Programming*, vol. 167, no. 1, pp. 155–189, 2018.

[27] M. C. Campi and A. Carè, "Random Convex Programs with L_1-Regularization: Sparsity and Generalization," *SIAM Journal on Control and Optimization*, vol. 51, no. 5, pp. 3532–3557, 2013.

[28] A. Carè, S. Garatti, and M. C. Campi, "FAST-- Fast Algorithm for the Scenario Technique," *Operations Research*, vol. 62, no. 3, pp. 662–671, 2014.

[29] G. C. Calafiore, F. Dabbene, and R. Tempo, "Research on Probabilistic Methods for Control System Design," *Automatica*, vol. 47, no. 7, pp. 1279–1293, 2011.

[30] M. Chamanbaz, F. Dabbene, R. Tempo, V. Venkataramanan, and Q.-G. Wang, "Sequential Randomized Algorithms for Convex Optimization in the Presence of Uncertainty," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2565–2571, 2016.

[31] G. C. Calafiore, "Repetitive Scenario Design," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1125–1137, 2017.

[32] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014.

[33] E. Delage and Y. Ye, "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[34] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations research*, vol. 58, no. 4-part-1, pp. 902–917, 2010.

[35] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, "A distributionally robust perspective on uncertainty quantification and chance constrained programming," *Mathematical Programming*, vol. 151, no. 1, pp. 35–62, 2015.

[36] S. Zymler, D. Kuhn, and B. Rustem, "Distributionally robust joint chance constraints with second-order moment information," *Mathematical Programming*, pp. 1–32, 2013.

[37] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, "Ambiguous joint chance constraints under mean and dispersion information," *Operations Research*, vol. 65, no. 3, pp. 751–767, 2017.

[38] B. Li, R. Jiang, and J. L. Mathieu, "Ambiguous risk constraints with moment and unimodality information," *Mathematical Programming*, vol. 173, no. 1-2, pp. 151–192, 2019.

[39] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, vol. 158, no. 1-2, pp. 291–327, 2016.

[40] Y. Zhang, R. Jiang, and S. Shen, "Ambiguous chance-constrained bin packing under mean-covariance information," *arXiv preprint arXiv:1610.00035*, 2016.

[41] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, 2013.

[42] J. Cheng, E. Delage, and A. Lisser, "Distributionally robust stochastic knapsack problem," *SIAM Journal on Optimization*, vol. 24, no. 3, pp. 1485–1506, 2014.

[43] W. Xie and S. Ahmed, "On deterministic reformulations of distributionally robust joint chance constrained optimization problems," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1151–1182, 2018.

[44] W. Chen, M. Sim, J. Sun, and C.-P. Teo, "From CVaR to uncertainty set: Implications in joint chance-constrained optimization," *Operations research*, vol. 58, no. 2, pp. 470–485, 2010.

[45] Z. Chen, D. Kuhn, and W. Wiesemann, "Data-driven chance constrained programs over wasserstein balls," *arXiv preprint arXiv:1809.00210*, 2018.

[46] R. Ji and M. Lejeune, "Data-driven distributionally robust chance-constrained optimization with wasserstein metric," *Available at SSRN 3201356*, 2018.

[47] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

[48] I. R. Petersen, M. R. James, and P. Dupuis, "Minimax optimal control of stochastic uncertain systems with relative entropy constraints," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 398–412, 2000.

[49] L. P. Hansen and T. J. Sargent, *Robustness*. Princeton university press, 2008.

[50] D. Love and G. Bayraksan, "Phi-divergence constrained ambiguous stochastic programs for data-driven optimization," Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio, Tech. Rep., 2015.

[51] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and P. Plechác, "Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 80–111, 2016.

[52] H. Lam and E. Zhou, "The empirical likelihood approach to quantifying uncertainty in sample average approximation," *Operations Research Letters*, vol. 45, no. 4, pp. 301–307, 2017.

[53] H. Lam, "Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization," *Operations Research*, vol. 67, no. 4, pp. 1090–1105, 2019.

[54] J.-y. Gotoh, M. J. Kim, and A. E. Lim, "Robust empirical optimization is almost the same as mean–variance optimization," *Operations Research Letters*, vol. 46, no. 4, pp. 448–452, 2018.

[55] J. Duchi, P. Glynn, and H. Namkoong, "Statistics of robust optimization: A generalized empirical likelihood approach," *arXiv preprint arXiv:1610.03425*, 2016.

[56] P. M. Esfahani and D. Kuhn, "Data-Driven Distributionally Robust Optimization using the wasserstein Metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, pp. 1–52, 2015.

[57] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *arXiv preprint arXiv:1604.01446*, 2016.

[58] J. Blanchet, Y. Kang, and K. Murthy, "Robust wasserstein profile inference and applications to machine learning," *arXiv preprint arXiv:1610.05627*, 2016.

[59] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[60] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

[61] Z. Hu, J. Cao, and L. J. Hong, "Robust simulation of global warming policies using the dice model," *Management Science*, vol. 58, no. 12, pp. 2190–2206, 2012.

[62] P. Glasserman and X. Xu, "Robust risk measurement and model risk," *Quantitative Finance*, vol. 14, no. 1, pp. 29–58, 2014.

[63] H. Lam, "Robust sensitivity analysis for stochastic systems," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1248–1275, 2016.

[64] Z. Hu and L. J. Hong, "Robust simulation of stochastic systems with input uncertainties modeled by statistical divergences," in *Proceedings of the 2015 Winter Simulation Conference*, L. Yilmaz et al., Ed., IEEE, Piscataway, New Jersey, 2015, pp. 643–654.

[65] H. Lam, "Sensitivity to serial dependency of input processes: A robust approach," *Management Science*, vol. 64, no. 3, pp. 1311–1327, 2018.

[66] S. Ghosh and H. Lam, "Robust analysis in stochastic simulation: Computation and performance guarantees," *Operations Research*, vol. 67, no. 1, pp. 232–249, 2019.

[67] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming*, vol. 107, no. 1, pp. 37–61, 2006.

[68] H. Lam and F. Li, "Sampling uncertain constraints under parametric distributions," in *2018 Winter Simulation Conference (WSC)*, IEEE, 2018, pp. 2072–2083.

[69] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[70] E. L. Lehmann, *Elements of Large-sample Theory*. Springer Science & Business Media, 2004.

[71] L. Pardo, *Statistical Inference Based on Divergence Measures*. New York: Chapman and Hall/CRC, 2005.

[72] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2014.

[73] A. P. Korostelev and O. Korosteleva, *Mathematical Statistics: Asymptotic Minimax Theory*. Providence, Rhode Island: American Mathematical Society, 2011.

[74] D. P. Bertsekas, "Control of uncertain systems with a set-membership description of the uncertainty.," Ph.D. dissertation, Massachusetts Institute of Technology, 1971.

[75] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar, *et al.*, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[76] F. H. Clarke, "Generalized gradients and applications," *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.

[77] A. Marandi, A. Ben-Tal, D. den Hertog, and B. Melenberg, "Extending the scope of robust quadratic optimization," *Available on Optimization Online*, 2017.

[78] A. Ben-Tal, E. G. Laurent, and A. Nemirovski, *Robust Optimization*, ser. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[79] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2007.

[80] M. A. Lejeune and F. Margot, "Solving chance-constrained optimization problems with stochastic quadratic inequalities," *Operations Research*, vol. 64, no. 4, pp. 939–957, 2016.

[81] G. Folland, *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.

[82] A. Winkelbauer, "Moments and absolute moments of the normal distribution," *arXiv preprint arXiv:1209.4340*, 2012.

[83] A. O. Daalhuis, "Confluent hypergeometric functions," *NIST Handbook of Mathematical Functions, FWJ Olver, DW Lozier, RF Boisvert, and CW Clark, eds., Cambridge University, New York*, pp. 321–349, 2010.

[84] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014, vol. 16.

[85] J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming," *Annals of Operations Research*, vol. 142, no. 1, pp. 215–241, 2006.

[86] J. L. Higle and S. Sen, *Stochastic decomposition: a statistical method for large scale stochastic linear programming*. Springer Science & Business Media, 2013, vol. 8.

[87] P. Jirutitijaroen and C. Singh, "Reliability constrained multi-area adequacy planning using stochastic programming with sample-average approximations," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 504–513, 2008.

[88] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, "Sample average approximation method for chance constrained programming: Theory and applications," *Journal of optimization theory and applications*, vol. 142, no. 2, pp. 399–416, 2009.

[89] W. Wang and S. Ahmed, "Sample average approximation of expected value constrained stochastic programs," *Operations Research Letters*, vol. 36, no. 5, pp. 515–519, 2008.

[90] G. Barbarosoğlu and Y. Arda, "A two-stage stochastic programming framework for transportation planning in disaster response," *Journal of the operational research society*, vol. 55, no. 1, pp. 43–53, 2004.

[91] C. Liu, Y. Fan, and F. Ordóñez, "A two-stage stochastic programming model for transportation network protection," *Computers & Operations Research*, vol. 36, no. 5, pp. 1582–1590, 2009.

[92] N. Noyan, "Risk-averse two-stage stochastic programming with an application to disaster management," *Computers & Operations Research*, vol. 39, no. 3, pp. 541–559, 2012.

[93] G. Huang and D. P. Loucks, "An inexact two-stage stochastic programming model for water resources management under uncertainty," *Civil Engineering Systems*, vol. 17, no. 2, pp. 95–118, 2000.

[94] M. Dillon, F. Oliveira, and B. Abbasi, "A two-stage stochastic programming model for inventory management in the blood supply chain," *International Journal of Production Economics*, vol. 187, pp. 27–41, 2017.

[95] X. Chen, A. Shapiro, and H. Sun, "Convergence analysis of sample average approximation of two-stage stochastic generalized equations," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 135–161, 2019.

[96] M. Anthony and N. Biggs, *Computational learning theory*. Cambridge University Press, 1997, vol. 30.

[97] M. J. Kearns, U. V. Vazirani, and U. Vazirani, *An introduction to computational learning theory*. MIT press, 1994.

[98] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[99] A. Van Der Vaart and J. A. Wellner, "A note on bounds for vc dimensions," *Institute of Mathematical Statistics collections*, vol. 5, p. 103, 2009.

[100] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, 1989.

[101] S. Hanneke, "The optimal sample complexity of pac learning," *Journal of Machine Learning Research*, vol. 17, no. 38, pp. 1–15, 2016.

[102] H. U. Simon, "An almost optimal pac algorithm," in *Proceedings of The 28th Conference on Learning Theory*, P. Grünwald, E. Hazan, and S. Kale, Eds., ser. Proceedings of Machine Learning Research, vol. 40, Paris, France: PMLR, 2015, pp. 1552–1563.

[103] R. M. Dudley, "Central limit theorems for empirical measures," *The Annals of Probability*, pp. 899–929, 1978.

[104] A. W. Van Der Vaart and J. A. Wellner, "Weak convergence," in *Weak convergence and empirical processes*, Springer, 1996, pp. 16–28.

[105] R.-J. Jing, M. Moreno-Maza, and D. Talaashrafi, "Complexity estimates for fourier-motzkin elimination," in *International Workshop on Computer Algebra in Scientific Computing*, Springer, 2020, pp. 282–306.

[106] M. J. Panik, "Extreme points and directions for convex sets," in *Fundamentals of Convex Analysis: Duality, Separation, Representation, and Resolution*. Dordrecht: Springer Netherlands, 1993, pp. 189–234, ISBN: 978-94-015-8124-0.

[107] M. Terzer, "Large scale methods to enumerate extreme rays and elementary modes," Ph.D. dissertation, ETH Zurich, 2009.

[108] G. Calafiore and M. C. Campi, "Uncertain convex programs: Randomized solutions and confidence levels," *Mathematical Programming*, vol. 102, no. 1, pp. 25–46, 2005.

[109] S. Ahmed, "Two-stage stochastic integer programming: A brief introduction," *Wiley encyclopedia of operations research and management science*, 2010.

[110] S. Ahmed, A. Shapiro, and E. Shapiro, "The sample average approximation method for stochastic programs with integer recourse," *Submitted for publication*, pp. 1–24, 2002.

[111] S. Küçükyavuz and S. Sen, "An introduction to two-stage stochastic mixed-integer programming," in *Leading Developments from INFORMS Communities*, INFORMS, 2017, pp. 1–27.

[112] H. M. Bidhandi and J. Patrick, "Accelerated sample average approximation method for two-stage stochastic programming with binary first-stage variables," *Applied Mathematical Modelling*, vol. 41, pp. 582–595, 2017.

[113] I. Borosh and L. B. Treybig, "Bounds on positive integral solutions of linear diophantine equations," *Proceedings of the American Mathematical Society*, vol. 55, no. 2, pp. 299–304, 1976.

[114] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Advances in neural information processing systems*, 2007, pp. 601–608.

[115] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[116] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

[117] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 114.

[118] M. Sugiyama, M. Krauledat, and K.-R. MÃžller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 985–1005, 2007.

[119] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[120] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[121] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 264–271.

[122] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, e49–e57, 2006.

[123] M. Sugiyama and M. Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[124] H. Hachiya, T. Akiyama, M. Sugiyama, and J. Peters, "Adaptive importance sampling with automatic model selection in value function approximation.," in *AAAI*, 2008, pp. 1351–1356.

[125] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[126] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[127] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2006, pp. 120–128.

[128] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Advances in neural information processing systems*, 2008, pp. 1433–1440.

[129] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," in *Proceedings of the 24th international conference on Machine learning*, ACM, 2007, pp. 81–88.

[130] T. Kanamori, T. Suzuki, and M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation," *Machine Learning*, vol. 86, no. 3, pp. 335–367, 2012.

[131] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh, "Sample selection bias correction theory," in *International conference on algorithmic learning theory*, Springer, 2008, pp. 38–53.

[132] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 1855–1862.

[133] D. Pardoe and P. Stone, "Boosting for regression transfer," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress, 2010, pp. 863–870.

[134] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[135] Y.-L. Yu and C. Szepesvári, "Analysis of kernel mean matching under covariate shift," in *ICML*, Omnipress, 2012, pp. 1147–1154.

[136] F. Cucker and D. X. Zhou, *Learning theory: an approximation theory viewpoint*. Cambridge University Press, 2007, vol. 24.

[137] J. Blanchet and H. Lam, "State-dependent importance sampling for rare-event simulation: An overview and recent advances," *Surveys in Operations Research and Management Science*, vol. 17, no. 1, pp. 38–59, 2012.

[138] J. Wen, C.-N. Yu, and R. Greiner, "Robust learning under uncertain test distributions: Relating covariate shift to model misspecification.," in *ICML*, 2014, pp. 631–639.

[139] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 4, pp. 699–746, 2008.

[140] E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small, "Non-parametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 1229–1245, 2017.

[141] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[142] H. Sun and Q. Wu, "A note on application of integral operator in learning theory," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 416–421, 2009.

[143] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in computational mathematics*, vol. 13, no. 1, p. 1, 2000.

[144] B. L. Nelson, "Control variate remedies," *Operations Research*, vol. 38, no. 6, pp. 974–992, 1990.

[145] P. W. Glynn and R. Szechtman, "Some new perspectives on the method of control variates," in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, Springer, 2002, pp. 27–49.

[146] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *International conference on computational learning theory*, Springer, 2001, pp. 416–426.

[147] I. Pinelis *et al.*, "Optimum bounds for the distributions of martingales in banach spaces," *The Annals of Probability*, vol. 22, no. 4, pp. 1679–1706, 1994.

[148] H. Sun and Q. Wu, "Regularized least square regression with dependent samples," *Advances in Computational Mathematics*, vol. 32, no. 2, pp. 175–189, 2010.

[149] M. A. Lifshits, *Gaussian random functions*. Springer Science & Business Media, 2013, vol. 322.

[150] P. Geibel, "Reinforcement learning for mdps with constraints," in *European Conference on Machine Learning*, Springer, 2006, pp. 646–653.

[151] J. Lee, Y. Jang, P. Poupart, and K.-E. Kim, "Constrained bayesian reinforcement learning via approximate linear programming."

[152] E. A. Feinberg and U. G. Rothblum, "Splitting randomized stationary policies in total-reward markov decision processes," *Mathematics of Operations Research*, vol. 37, no. 1, pp. 129–153, 2012.

[153] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 22–31.

[154] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, 2015, pp. 1889–1897.

[155] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[156] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, 2018.

[157] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.

[158] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[159] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[160] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[161] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997, vol. 6.

[162] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[163] R. Bellman, *Dynamic programming*. Courier Corporation, 2013.

[164] H. Robbins and S. Monro, "A stochastic approximation method," in *Herbert Robbins Selected Papers*, Springer, 1985, pp. 102–109.

[165] E. Even-Dar and Y. Mansour, "Convergence of optimistic and incremental q-learning," in *Advances in neural information processing systems*, 2002, pp. 1499–1506.

[166] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine learning*, vol. 16, no. 3, pp. 185–202, 1994.

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

[168] N. Archak, V. Mirrokni, and S Muthukrishnan, "Budget optimization for online campaigns with positive carryover effects," in *International Workshop on Internet and Network Economics*, Springer, 2012, pp. 86–99.

[169] C. Bayer, P. Friz, S. Riedel, and J. Schoenmakers., "From rough path estimates to multilevel Monte Carlo," *SIAM Journal on Numerical Analysis*, vol. 54, no. 3, pp. 1449–1483, 2016.

[170] S. Whitaker., "Flow in porous media I: A theoretical derivation of Darcy's law," *Transport in porous media*, vol. 1, no. 1, pp. 3–25, 1986.

[171] D. Duffie., *Dynamic asset pricing theory*. Princeton University Press, 2010.

[172] G. Marsily, F. Delay, J. Goncalves, P. Renard, T. Vanessa, and S. Violette., "Dealing with spatial heterogeneity," *Hydrogeology Journal*, vol. 13, no. 1, pp. 161–183, 2005.

[173] J. Delhomme., "Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach," *Water Resources Research*, vol. 15, no. 2, pp. 269–280, 1979.

[174] M. Ostoja-Starzewski., *Microstructural randomness and scaling in mechanics of materials*. CRC Press, 2007.

[175] K. Sobczyk and D. Kirkner., *Stochastic modeling of microstructures*. Springer Science & Business Media, 2012.

[176] M. Hofmann., "L_p estimation of the diffusion coefficient," *Bernoulli*, vol. 5, no. 3, pp. 447–481, 1999.

[177] S. Pastorello., "Diffusion coefficient estimation and asset pricing when risk premia and sensitivities are time varying," *Mathematical Finance*, vol. 6, no. 1, pp. 111–117, 1996.

[178] D. Guignard, "Partial differential equations with random input data: A perturbation approach," *Archives of Computational Methods in Engineering*, vol. 26, no. 5, pp. 1313–1377, 2019.

[179] K. Cliffe, M. Giles, R. Scheichl, and A. Teckentrup., "Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients," *Computing and Visualization in Science*, vol. 14, no. 1, p. 3, 2011.

[180] D. Crevillén-García and H. Power, "Multilevel and quasi-Monte Carlo methods for uncertainty quantification in particle travel times through random heterogeneous porous media," *Royal Society open science*, vol. 4, no. 8, p. 170 203, 2017.

[181] M. Giles and L. Szpruch, "Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation," *The Annals of Applied Probability*, vol. 24, no. 4, pp. 1585–1620, 2014.

[182] H. G. Matthies and A. Keese, "Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations," *Computer methods in applied mechanics and engineering*, vol. 194, no. 12-16, pp. 1295–1331, 2005.

[183] A. Teckentrup, P. Jantsch, C. Webster, and M. Gunzburger., "A multilevel stochastic collocation method for partial differential equations with random input data," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 3, no. 1, pp. 1046–1074, 2015.

[184] J. Charrier, R. Scheichl, and A. Teckentrup., "Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods," *SIAM Journal on Numerical Analysis*, vol. 51, no. 1, pp. 322–352, 2013.

[185] P. Kloeden and E. Platen., *Numerical Solution of Stochastic Differential Equations*, ser. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011, ISBN: 9783540540625.

[186] S. Mishra, C. Schwab, and J. Šukys., "Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions," *Journal of Computational Physics*, vol. 231, no. 8, pp. 3365–3388, 2012.

[187] X. Li and J. Liu., "A multilevel approach towards unbiased sampling of random elliptic partial differential equations," *arXiv preprint arXiv:1605.06349*, 2016.

[188] M. Giles., "Multilevel Monte Carlo path simulation," *Operations Research*, vol. 56, no. 3, pp. 607–617, 2008.

[189] ——, "Multilevel Monte Carlo methods," in *Monte Carlo and Quasi-Monte Carlo Methods 2012*, Springer, 2013, pp. 83–103.

[190] M. B. Giles and F. Bernal, "Multilevel estimation of expected exit times and other functionals of stopped diffusions," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 4, pp. 1454–1474, 2018.

[191] C. Rhee and P. Glynn., "Unbiased estimation with square root convergence for SDE models," *Operations Research*, vol. 63, no. 5, pp. 1026–1043, 2015. eprint: `http://dx.doi.org/10.1287/opre.2015.1404`.

[192] J. Blanchet and P. Glynn, "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization," in *Winter Simulation Conference (WSC), 2015*, IEEE, 2015, pp. 3656–3667.

[193] R. Howard., "The Gronwall inequality," *Lecture notes*, 1998.

[194] T. Lyons., "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.

[195] A. Davie., "Differential equations driven by rough paths: An approach via discrete approximation," *Applied Mathematics Research eXpress*, vol. 2008, 2008.

[196] P. Friz and N. Victoir., *Multidimensional stochastic processes as rough paths: theory and applications*. Cambridge University Press, 2010, vol. 120.

[197] P. Friz and M. Hairer., *A course on rough paths: with an introduction to regularity structures*. Springer, 2014.

[198] M. Hairer, "A theory of regularity structures," *Inventiones mathematicae*, vol. 198, no. 2, pp. 269–504, 2014.

[199] T. Lyons and N. Victoir., "Cubature on Wiener space," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, The Royal Society, vol. 460, 2004, pp. 169–198.

[200] J. Blanchet, X. Chen, and J. Dong, "ϵ-strong simulation for multidimensional stochastic differential equations via rough path analysis," *The Annals of Applied Probability*, vol. 27, no. 1, pp. 275–336, Feb. 2017.

[201] I. Karatzas and S. Shreve., *Brownian motion and stochastic calculus*. Springer Science & Business Media, 2012, vol. 113.

[202] J. Steele., *Stochastic calculus and financial applications*. Springer Science & Business Media, 2012, vol. 45.

[203] D. Burkholder, B. Davis, and R. Gundy., "Integral inequalities for convex functions of operators on martingales," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, University of California Press, 1972, pp. 223–240.

[204] R. Adler., *Random fields and their geometry*. Birkhäuser, 2003.

[205] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

[206] M. Wang, E. X. Fang, and H. Liu, "Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions," *Mathematical Programming*, vol. 161, no. 1-2, pp. 419–449, 2017.

[207] X. Lian, M. Wang, and J. Liu, "Finite-sum Composition Optimization via Variance Reduced Gradient Descent," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, Fort Lauderdale, FL, USA: PMLR, 2017, pp. 1159–1167.

[208] L. Lei and M. I. Jordan, "Less than a single pass: Stochastically controlled stochastic gradient method," *arXiv preprint arXiv:1609.03261*, 2016.

[209] M. Wang and J. Liu, "Accelerating stochastic composition optimization," in *Advances In Neural Information Processing Systems*, 2016, pp. 1714–1722.

[210] S. Ghadimi, A. Ruszczyński, and M. Wang, "A single time-scale stochastic approximation method for nested stochastic optimization," *arXiv preprint arXiv:1812.01094*, 2018.

[211] M. B. Giles, L. Szpruch, *et al.*, "Antithetic multilevel monte carlo estimation for multi-dimensional sdes without lévy area simulation," *The Annals of Applied Probability*, vol. 24, no. 4, pp. 1585–1620, 2014.

[212] S. Dereich and F. Heidenreich, "A multilevel monte carlo algorithm for lévy-driven stochastic differential equations," *Stochastic Processes and their Applications*, vol. 121, no. 7, pp. 1565–1587, 2011.

[213] M. B. Giles and C. Reisinger, "Stochastic finite differences and multilevel monte carlo for a class of spdes in finance," *SIAM Journal on Financial Mathematics*, vol. 3, no. 1, pp. 572–592, 2012.

[214] D. F. Anderson and D. J. Higham, "Multilevel monte carlo for continuous time markov chains, with applications in biochemical kinetics," *Multiscale Modeling & Simulation*, vol. 10, no. 1, pp. 146–179, 2012.

[215] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 567–599, 2013.

[216] N. L. Roux, M. Schmidt, and F. R. Bach, "A stochastic gradient method with an exponential convergence _rate for finite training sets," in *Advances in Neural Information Processing Systems*, 2012, pp. 2663–2671.

[217] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[218] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *International Conference on Machine Learning*, 2015, pp. 1–9.

[219] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.

[220] Z. Allen-Zhu and Y. Yuan, "Improved svrg for non-strongly-convex or sum-of-non-convex objectives," arXiv preprint, Tech. Rep., 2016.

[221] R. Harikandeh, M. O. Ahmed, A. Virani, M. Schmidt, J. Konečnỳ, and S. Sallinen, "Stop-wasting my gradients: Practical svrg," in *Advances in Neural Information Processing Systems*, 2015, pp. 2251–2259.

[222] L. Lei and M. Jordan, "Less than a single pass: Stochastically controlled stochastic gradient," in *Artificial Intelligence and Statistics*, 2017, pp. 148–156.

[223] P. Gong and J. Ye, "Linear convergence of variance-reduced stochastic gradient without strong convexity," *arXiv preprint arXiv:1406.1102*, 2014.

[224] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Advances in Neural Information Processing Systems*, 2014, pp. 1574–1582.

[225] S. Lacoste-Julien, M. Schmidt, and F. Bach, "A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method," *arXiv preprint arXiv:1212.2002*, 2012.

[226] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford, "Competing with the empirical risk minimizer in a single pass," in *Conference on learning theory*, 2015, pp. 728–763.

[227] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[228] C. Sutton, A. McCallum, and K. Rohanimanesh, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," *Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 693–723, 2007.

[229] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, 2003, pp. 134–141.

[230] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Association for Computational Linguistics, 2003, pp. 188–191.

[231] S. Nowozin, C. H. Lampert, *et al.*, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

[232] F. Barahona, "On the computational complexity of ising spin glass models," *Journal of Physics A: Mathematical and General*, vol. 15, no. 10, p. 3241, 1982.

[233] V. Chandrasekaran, N. Srebro, and P. Harsha, "Complexity of inference in graphical models," *arXiv preprint arXiv:1206.3240*, 2012.

[234] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[235] S. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," ACM, 2006, pp. 969–976.

[236] M. Schmidt, R. Babanezhad, M. Ahmed, A. Defazio, A. Clifton, and A. Sarkar, "Non-uniform stochastic average gradient method for training conditional random fields," in *Artificial Intelligence and Statistics*, 2015, pp. 819–828.

[237] R. T. Rust and A. J. Zahorik, "Customer satisfaction, customer retention, and market share," *Journal of Retailing*, vol. 69, no. 2, pp. 193–215, 1993.

[238] F. Shahrokhi and D. W. Matula, "The maximum concurrent flow problem," *Journal of the ACM (JACM)*, vol. 37, no. 2, pp. 318–334, 1990.

[239] R. D. Cox, "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society*, vol. 34, pp. 187–220, 1972.

[240] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[241] ——, "Regression models and life-tables," in, Springer, 1992, pp. 527–541.

[242] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Advances in Neural Information Processing Systems*, 2004, pp. 25–32.