A General Framework of Optimal Stochastic Optimization with Dependent Data: Multiple Optimality Guarantee, Sample Complexity, and Tractability

by

Bo PAN

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

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

We study Sample Averaging approximation for data-driven decision-making under uncertainty where the underlying probability measure is partially observable through finitely dependent training samples. Given the data generation procedure in which the datasets are taken along a single trajectory of a stationary Stochastic Process, the central aspect of this paper is to investigate the fundamental results demonstrating that certain SAA procedure retains optimality with dependent samples. Leveraging results from measure concentration, we derive conditions on the considered data-generating process, under which solutions of SAAs retain asymptotically optimality and tractability and, additionally, enjoy finite-sample performance guarantees. While the obtained SAA is tractable, the learning algorithms for the resulting surrogate optimization could be computationally excruciating. To address the numerical difficulty, we further propose a stochastic operator-splitting scheme, referring to a Stochastic algorithm that is easily implementable and highly parallelizable in solving complicated optimization problems. We discuss convergence rates, stability, and finite-sample error bounds for the iterates. The theoretical results we investigate are self-contained and accommodate parametric, non-parametric, and semi-parametric machine learning tasks. Numerical experiments validate our theoretical results and demonstrate empirically that our approach outperforms baseline approaches.

Preface

This thesis is about stochastic convex optimization, a special class of convex optimization that aims to find an optimal estimation under uncertainty. It has arisen in a variety of application domains and research areas. While rich theoretical results have been established, the majority of them critically depend on the IID assumption of the training samples. The basic point of this work is that we explore the foundational result of a data-driven approach of stochastic optimization with Non-IID. Specially, we consider the Sample Average Approximation (SAA) with training samples generated from a single trajectory of stochastic process, which is assumed to be mixing.

Convex optimization problems are prevalent in practice, thus how to solve them in an efficient manner is as equivalently important as a theoretical guarantee. The second part is about the tractability of SAA. We specifically consider operator-splitting schemes, powerful algorithms that are adaptive and scalable enough for a variety of statistical settings. We study the convergence of the proposed algorithms and perform numerical experiments to verify their efficiency.

Parts work of this thesis have been presented in the paper, " Sample average approximation for stochastic optimization with dependent data: Performance guarantees and traceability " [65], which was led by Wang, Yafei (University of Essex) and me with the corporation with Tu, Wei (Queen's University Canada), Liu, Peng (University of Kent), Jiang, Bei (University of Alberta), Gao, Chao (Huawei Canada), Lu, Wei (Huawei Canada) and Jui, Shangling (Huawei Canada) and Kong, Linglong (University of Alberta). Specifically, the paper covers the content of the statistical performance guarantee in Chapter 3, the results of the algorithmic analysis in part II, and the numerical experiment of the lasso-type quadratic minimization problem. My main contribution to the published paper is mainly about the algorithmic analysis of operator-splitting schemes and numerical analysis.

The entire work of this thesis is under development and planned to be submitted to Journal (Ongoing).

To my parents Rongmei Qi, and Linling Pan and my wife Yafei Wang
For raising, accompanying, understanding and encouragement

True optimization is the revolutionary contribution of modern research to decision processes.

– George Bernard Dantzig, 1914 - 2005.

Acknowledgements

I would like to thank all of those who have supported and helped me throughout my MSc studies. My acknowledgment must begin with my supervisors Prof. Linglong Kong and Prof. Bei Jiang. They guide me to learn in a multitude of research areas from statistics to optimization, encouraging and exemplifying patience for research and showing me how to think, write, and present the work. I am also extremely grateful to Prof. Ross Tsuyuki, and Prof. Yazid Al Hamarneh for being my statistical consulting mentor. They encourage me to develop a taste for research problems at the border of theory and practice which maintains the balance in developing methodology and real application.

Thanks should also go to Peng Liu and Peijun Sang for sharing their experience in doing research and showing me the routine goes from junior research to being a senior. I am grateful to Lily Yushko and Henry Lee having all been wonderfully encouraging, giving feedback on database management and multi-task management and Imran Hassan, who lead me to the world of real data application when I just graduated with an undergraduate. My sincere thanks are also extended to group members, Mei Li, Matthew Pietrosanu, and other members of my colleagues for their thorough review, insightful comments, and helpful suggestions.

Finally, I must thank my love wife, Yafei Wang, for her companionship, understanding, and encouragement. My appreciation also goes out to my parents-Mr. Lingling Pan and Mrs. Rongmei Qi for their support all through my studies.

Contents

I Multi-Optimality Guarantees with Non-i.i.d Data 1

1 Introduction and Overview 2
1.1 Motivation and Related Work 3
1.2 Main Contributions 6
1.3 Organization 7

2 Data-Driven Stochastic Programming 8
2.1 Multiple Optimality Guarantees 8
2.2 Data-generating Process 9
2.2.1 Linear Dynamical System 9
2.2.2 Peer-to-peer Decentralized Optimization 10
2.3 Computational Tractability 10
2.3.1 Problem Statement 10
2.3.2 Geometry of Operator Splitting Schemes: Feasibility Problem 11

3 Statistical Performance Guarantees 13
3.1 Preliminary and Assumption 13
3.2 Asymptotic Consistency and Out-of-sample Performance 14

4 Statistical Inference and Sample Complexity of SAA with Non-i.i.d Data 17
4.1 Statistical Inference for Stochastic Optimization 17
4.1.1 Functional Covariance Inequality and Assumptions 17
4.1.2 Functional Central Limit Theorem 18
4.2 Sample Complexity of SAA with Non-IID Data 19
4.2.1 Measurement of Sample Complexity 19

II Algorithmic Analysis via Modern Monotone Operator Theory 24

5 Computational Tractability 25
5.1 Convergence Rate Analysis of S-KM with IID Data 28
5.2 Convergence Rate Analysis of S-KM with Non-IID Data 29

6 Convergence Analysis of Several Splitting Algorithms 32
6.1 Stochastic Generalized FBS 32
6.2 S-rPRS and S-DRS 33
6.3 Numerical Experiments 36
6.3.1 Lasso-type quadratic minimization problem 36
6.3.2 Robust regression problem via P2P network 37

7 Conclusion 38

Reference 39

Appendix A Supplemental Material 43
A.1 Peer-to-peer Sampling Procedure as a Markov Chain 43
A.2 Proof of Theorem 1 43
A.3 Proof of Theorem 4 44
A.4 Proof of Lemma 1 45
A.5 Proof of Lemma 2 47
A.6 Proof of Theorem 5 50
A.7 Proof of Theorem 11 52

A.8 Proof of Theorem 12 53
A.9 Proof of Corollary 3 53
A.10 Proof of Fact 6.2 54
A.11 Proof of Corollary 4 54
A.12 Proof of Theorems 8–10 54
A.12.1 Proof of Theorem 8 55
A.12.2 Proof of Theorem 9 58
A.12.3 Proof of Theorem 10 59
A.13 Proof of Theorem 13 59
A.14 Proof of Theorem 14 61
A.15 Proof of Theorem 15 63

List of Tables

5.1 Overview of several first-order algorithms 25

List of Figures

6.1 Performance of the SP, SP-m, and MR-s methods for various values of m and s as measured by (left) regret, (middle) FPR, and (right) distance between the estimated and true value of x in the lasso-type example. 36
6.2 Robust regression problem via P2P network 37

Part I

Multi-Optimality Guarantees with Non-i.i.d Data

Chapter 1

Introduction and Overview

We study the statistical properties of Sample Averaging Approximation for data-driven decision-making under uncertainty, focusing on optimality and tractability via finitely correlated samples generated by a single trajectory of a stationary Stochastic Process (S.P.), $P^k, \forall k > 0$. Such problems are ubiquitous and, traditionally, modelled as stochastic programs. In this paper, we consider stochastic optimization,

$$J^\star = \min_{x \in \mathbb{X}} \Big\{ \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)] = \int_{\Xi} \ell(x;\xi) \mathbb{P}(\mathrm{d}\xi) \Big\} \tag{1}$$

$$x^\star = \arg\min_{x \in \mathbb{X}} \Big\{ \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)] = \int_{\Xi} \ell(x;\xi) \mathbb{P}(\mathrm{d}\xi) \Big\}. \tag{2}$$

In the formulation 1–2, the feasible region is a closed, convex subset $\mathbb{X}$ of $\mathbb{R}^d$, let $\mathbb{L} = \{\ell(\cdot, \xi) \mid \xi \in \Xi\}$, we assume that every element $\ell(x;\xi) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ of $\mathbb{L}$ is a closed, convex, and proper (c.c.p) function, but not necessarily differentiable, and ξ be the random samples received. ξ is a random vector on the probability space $(\Xi, \mathcal{A}, \mathbb{P})$, where the domain Ξ is a separable metric space. Problems 1–2 widely encompasses research areas, such as convex optimization, operations research, and first- and second-order methods, as well as linearly and sublinearly convergent algorithms. This setting is a powerful modelling paradigm widely used to model large-scale optimization problems in machine learning, signal processing, and other computational sciences [10], [27], [34], [60], [62].

The fundamental task of a generic stochastic convex optimization is to seek an optimal decision variable $x \in \mathbb{X}$ that minimizes a given expected loss $\mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]$ taken with respect to an underlying population distribution $\mathbb{P}$ of a random variable $\xi \in \mathbb{R}^m$. A wide variety of stochastic optimization methods for solving the problems 1–2 have been explored in extensive literature. However, constructing optimal estimators for $x^\star$ and $J^\star$ remains a major challenge in large-scale, data-driven optimization problems. As a first specific point, the true underlying population distribution $\mathbb{P}$ is never known but must be inferred through finitely-many observations. Even if the distribution $\mathbb{P}$ is known, the learning procedure could be computationally excruciating since evaluating the corresponding expectation for a fixed x involves computing a multivariate integral, which could be high-dimensional and intractable, *e.g.,* given the loss function $\ell(\cdot)$ is an affine function, evaluating integral is still NP-hard. Second, while classical data-driven techniques are computationally tractable and enjoy strong asymptotic performance guarantees, similar guarantees do not typically hold in finitely sample settings. By calibrating a stochastic program to a given dataset and estimating optimal decisions through these methods, the generalization performance could consequently be disappointing. Beyond that, most existing performance guarantees critically depend on the assumption that training samples are independent and identically distributed (IID). This assumption, however, can be difficult to justify or outright invalid in practice. Besides that, it is often unrealistic to assume that one has access to a source of independent randomness, so studying the effect of dependence among samples is natural and essential. Third, $J^\star$ and $x^\star$ generally do not have an analytic expression given the objectives that are composed of data fitting and several competing structures, such as sparsity, low rank, and smoothness, which enforce prior knowledge of the form of the solution. Thus, when establishing the practical utility of an underlying optimization scheme, tractability is equally as important as theoretical performance guarantees.

1.1 Motivation and Related Work

As the true distribution $\mathbb{P}$ of random variable ξ is unknown in most cases of practical interest, it could be extremely hard to evaluate the original problem 1–2 exactly—We miss essential information to solve problems directly and it is unclear which problem instance should be solved. Thus, it is convenient to embed a given stochastic optimization problem into a surrogate problem by replacing the unknown (true) probability measure with some approximation $\widehat{\mathbb{P}}_K$, which is independent of the distribution $\mathbb{P}$, *i.e.,* we reformulate the original stochastic optimization problem into some surrogate problem constructed completely from K training samples that can be solved efficiently. One important Monte Carlo simulation-based approach, Sample Average Approximation, is a powerful modelling paradigm that takes $\widehat{\mathbb{P}}_K$ as the (discrete) empirical distribution—the distribution that places a mass of $1/K$ at each of the sample points. SAA approximates $J^\star$ and $x^\star$ in 1–2 as

$$\widehat{J}_K^\star = \min_{x\in\mathbb{X}}\Big\{\mathbb{E}^{\widehat{\mathbb{P}}_K}[\ell(x;\xi)] = \int_\Xi \ell(x;\xi)\widehat{\mathbb{P}}_K(\mathrm{d}\xi) = \frac{1}{K}\sum_{k=1}^{K}\ell(x,\widehat{\xi}_k)\Big\} \tag{3}$$

$$\widehat{x}_K^\star = \arg\min_{x\in\mathbb{X}}\Big\{\mathbb{E}^{\widehat{\mathbb{P}}_K}[\ell(x;\xi)] = \int_\Xi \ell(x;\xi)\widehat{\mathbb{P}}_K(\mathrm{d}\xi) = \frac{1}{K}\sum_{k=1}^{K}\ell(x,\widehat{\xi}_k)\Big\}, \tag{4}$$

both of which are functions of the K samples. Throughout this paper, we reserve the superscript$\widehat{}$for objects that depend on the training data and thus constitute random objects governed by the product distribution $\mathbb{P}^K$.

Variants of the SAA-based methods in optimizations and other contexts are ubiquitous, from operations research to statistical learning problems, and often used tacitly without necessarily being referred to by this name. SAA-based methods are differentiated into three streams based on how data is generated: random samples can be streamed, available as historical data, or simulated by sampling techniques (e.g., Monte Carlo). More explicitly, SAA can be used to estimate $J^\star$ and $x^\star$ when IID data is drawn from an unknown $\mathbb{P}$ is available or when random covariates can be simulated from a known $\mathbb{P}$. In a Monte Carlo setting, it is easy to generate additional data, so the key is understanding the number of samples needed to balance the optimality and computational expense to compute SAA solutions. This number can be large, especially when sampling from a high-dimensional space. In contrast, additional data is unavailable when working with historical data, while streaming data cannot be stored in memory for a long period and should be processed as quickly as possible. Performance guarantees are thus a vital issue—Under mild structural assumptions on the cost function ℓ and the sampling process, SAA enjoys strong convergence properties (as $K \to \infty$) and is known to be computationally tractable for many cost functions $\ell(\cdot)$ and sets $\mathbb{X}$ ([40], [41]):

1. **Asymptotic convergence:** As the number of training samples K tends to infinity, both the estimated optimal value $\widehat{J}_K^\star$ and the estimated optimal solution $\widehat{x}_K^\star$ converge strongly.

2. **Tractability:** For many cost functions $\ell(x;\xi)$ and sets $\mathbb{X}$, finding the optimal value $\widehat{J}_K^\star$ and an optimal solution $\widehat{x}_K^\star$ are computationally tractable.

These strengths underlie SAA's popularity and practical success in data-driven settings. But such performance guarantees critically rely on the assumptions of sufficient large IID samples. However, similar guarantees do not typically hold for SAA for finite or dependent samples. For a review of SAA models where ξ has finitely many realizations, see [8], [26], [40], [41], [45].

A problematic criticism in practice is that in finite-dependent-sample settings, solutions of SAAs and out-of-sample performance can be volatile: perturbations in the data (e.g., in sample size or the level of dependence between samples) can produce significant changes in the solutions, which could lead unstable and untrustworthy decision or false inferences; Applying the resulting decisions on different datasets could be risky, even if the tested dataset is generated from the same distribution. Thus, methods should have guarantees that remain valid when the training samples are finite and exhibit serial dependence.

This paper examines SAA's performance for finite and dependent training samples and expands the method's practical applications. In particular, we derive finite-sample properties and asymptotic guarantees for SAA without the need for IID training samples from the distribution $\mathbb{P}$ over which we optimize. Instead, we assume that training samples are drawn from a single trajectory of ergodic S.P., which converge to the stationary distribution $\mathbb{P}$ ($P^k \to \mathbb{P}$, as $k \to \infty$). This setting accommodates serial dependence among the training samples and is a natural extension because, in many circumstances, the distribution $\mathbb{P}$ is never known or cannot be formulated by the decision maker. In other scenarios, it might not be possible to draw samples efficiently from $\mathbb{P}$. Such S.P.es include random processes on a finite-state space, finite-state Markov chains, and different vector mixing processes, among many others. Even, for IID cases, the dataset can be viewed as generated from IID S.P. where $P^k = \mathbb{P}$, $\forall k$.

To illustrate the importance of sampling efficiency and the need to accommodate intersample dependence, consider the following two examples. First, define $\Xi = \{\xi \in \{0,1\}^d \mid \langle a, \xi \rangle \leq b\}$, where $a \in \mathbb{R}^d, b \in \mathbb{R}$, and $\mathbb{P}$ is the uniform distribution over Ξ. A straightforward way to obtain a sample from $\mathbb{P}$ is by iterative random sampling from $\{0,1\}^d$ until the constraint on Ξ is satisfied: this approach takes $O(2^d)$ draws to obtain a feasible sample. Alternatively, it is possible to design a Markov chain that generates a sample that is ε-close to $\mathbb{P}$ and requires only $\log(\sqrt{d}/\varepsilon)\exp\big(O(\sqrt{d}(\log d)^{5/2})\big)$ draws—a greatly reduced sampling cost. Second, autoregressive processes generate non-IID data, here sequential entries of a time series.

Multiple replications (MR) [30] is an attractive method for generating samples from multiple independent trajectories of serially correlated states. Specifically, MR attempts to remove intersample dependence by using multiple trajectories of a stochastic process. After specifying initial conditions for the stochastic process and generating a sequence $\xi_1, \ldots, \xi_s$ for some s, MR keeps only the last sample ξ_s: it is assumed that the marginal distribution of ξ_s is close to $\mathbb{P}$. While this procedure can be applied to simulate K IID samples, MR may be expensive and wasteful in practice (especially if s is large) or might not apply at all if only one trajectory is available. Although MR is approximately two decades old, to the best of our knowledge, the optimality of SAA under MR has not yet been established. Theorem 1 gives asymptotic optimality results for this setting.

Theorem 1 (Asymptotic optimality of SAA under MR) *Given s initial samples for each of Q trajectories, let $\widehat{J}^\star_Q$ and $\widehat{x}^\star_Q$ represent the optimal value and optimizer obtained via SAA under MR. Assume that $\sup_{q\in[1,Q]} \ell(\widehat{x}^\star_q, \xi_q)$ is bounded above by some finite constant L_q. It follows that $\mathbb{P}\{\lim_{s\to\infty} \lim_{Q\to\infty} \widehat{J}^\star_Q = J^\star\} = 1$ and $\mathbb{P}\{\lim_{s\to\infty} \lim_{Q\to\infty} \widehat{x}^\star_Q = x^\star\} = 1$.*

Theorem 1 suggests that the “wasted” samples are necessary and that a small value of s may introduce bias in an approximation of $\mathbb{P}$ and hinder the convergence and accuracy of optimization algorithms.

In this paper, we instead focus on a single trajectory of a stochastic process $(P^k)_k$ that generates training samples and statistical information about $\mathbb{P}$. We additionally assume that the stochastic process is suitably ergodic, loosely, and that it converges relatively quickly to a stationary distribution $\mathbb{P}$. The existing literature that considers this setting does address algorithmic convergence guarantees with dependent samples but pays little attention to the statistical properties of SAA estimators. As examples, see [24] and [63] and the references therein. To fill this gap, we derive an upper confidence bound on the achievable out-of-sample performance and establish the asymptotic convergence of the optimal value and optimal solution via modern measure concentration results. Beyond asymptotic properties, we investigate the sample complexity of SAA with dependent data in order to obtain an ε-optimal solution. Our results generalize SAA from several recent works on stochastic and nonstochastic optimization, including the randomized incremental subgradient method [36], [37] and the Markov incremental subgradient method [35].

While SAA has demonstrated practical success in data-driven optimization, especially for complex datasets, the computational tractability of SAA suffers when the loss function ℓ itself possesses a complex structure. This problem commonly arises in statistical machine learning methods that enforce prior knowledge of the form of the solution (e.g., sparsity, rank, smoothness) [56]. Stochastic approximation provides a natural way to solve these problems. Indeed, algorithms that are both rich enough to capture the complexity of data and scalable enough to process large volumes of data in a paralleled or fully decentralized fashion have become centrally important. The need for such algorithms continues to motivate new developments, notably those related to stochastic algorithms and operator-splitting schemes. Indeed, the latter is known to significantly reduce computation time in parallel computing environments and accommodate memory requirements.

Operator-splitting techniques were first developed in the 1950s and applied to partial differential equations and inclusion problems [68]. Within the past few decades, certain operator-splitting methods, such as the alternating direction method of multipliers (ADMM) algorithm [12], have been widely applied to large-scale problems in signal/image processing, statistical machine learning, compressed sensing, and matrix completion [7], [16], [17], [32]. The success of operator-splitting methods is largely due to its simplicity and efficiency. These techniques offer several advantages over traditional optimization methods (e.g., Newton-type algorithms and interior-point methods): the former can easily handle nonsmooth terms and abstract linear operators, requires only simple arithmetic operations, and scales well with the dimension of the problem. Underlying these techniques is a reformulation of a given convex optimization problem into one of finding a fixed point of a nonexpansive operator. Operator-splitting methods additionally apply a decomposition procedure in which the original problem is broken into subproblems that can easily be solved. The solutions to these subproblems are combined to find a solution to the original problem. Several sophisticated algorithms and their stochastic versions such as the proximal point algorithm (PPA), forward-backward splitting (FBS), Douglas–Rachford (DRS), and Peaceman–Rachford splitting (PRS) [64], to name a few, have been widely applied to complex problems and have been extended to accommodate distributed and parallel optimization [12]. These techniques can all be embedded into an operator-splitting framework. Refer to [5] and the references therein for a comprehensive overview of operator-splitting schemes.

1.2 Main Contributions

As the main contribution, we demonstrate that SAA with dependent samples retains asymptotic properties and can be implemented efficiently through stochastic operator-splitting schemes for numerous popular loss functions. That is, the main information-theoretic results characterize the attainable results in a variety of applications and statistical settings. This also allows us to explore/develop procedures in the context of data-driven, inferential, and statistical learning tasks whose optimality we can verify. Such results are critical for a myriad of reasons; we can avoid making risky decisions that lacks of generalization or false inferences, we may realize a learning task is limited or even infeasible, and we can explicitly calculate the amount of data necessary for obtain content level of optimality of different statistical problems. We propose an efficient data-driven procedure for constructing a sequence of discrete empirical distributions that asymptotically estimates the true population distribution. This is a natural relaxation, because in many circumstances it may be impossible to draw samples from population distribution directly, such as when Ξ is a combinatorial space but it is possible to design a stochastic process that converges to the distribution. Furthermore, in many practical applications, it is unrealistic to assume that one has access to a source of data distributed separately such distributed medical data is restricted due to either ownership or other regulatory constraints. We find that many stochastic optimization problems can be approximated and solved efficiently by SAA. We further investigate the out-of-sample performance of the estimated optimal decisions, both theoretically and experimentally, and demonstrate the advantage of the purposed sampling procedure over commonly used data-generating processes.

1. We generalize SAA for a class of loss ℓ with IID data to settings with correlated training samples. We show that, when the underlying stochastic process is ϕ-mixing, the estimated optimal solution $\widehat{x}_K^\star$ and value $\widehat{J}_K^\star$ are asymptotically consistent. We further prove that the optimal value obtained through

SAA provides an upper confidence bound for the true optimal out-of-sample cost.

2. We establish the sample complexity of SAA with dependent data under a structural assumption of Lipschitz continuity.

3. We demonstrate that SAA can be implemented efficiently via stochastic operator-splitting schemes and show that the corresponding approximation error is bounded and concentrated around zero. We further establish deviation bounds for the algorithm's iterates.

4. We present several stochastic versions of popular algorithms such as stochastic proximal gradient descent (S-PGD) and the stochastic relaxed Peaceman–Rachford splitting algorithm (S-rPRS) and illustrate that the method enjoys strong convergence guarantees through numerical experiments with various data-generating process, including vector-auto regressive process and finite-state Markov Chain, to demonstrate that our approach outperforms other data-driven methods.

1.3 Organization

In Chapter 2, we introduce a general framework for stochastic optimization and a broad range of data-generating processes. We establish out-of-sample performance guarantees and asymptotic properties for our method in Chapter 3. In Chapter 4, we investigate sample complexity under a variety of structural assumptions with dependent data. In Chapter 5, for numerous objectives, we demonstrate that SAA can be split into composed of several competing structures, such as minimizing the sum of two functions, and, more generally, finding a zero of the sum of two monotone operators. We also derive deviation bounds when the underlying stochastic process is ϕ-mixing together with stability and global convergence results for a family of operator-splitting schemes. We apply established theoretical results to study the convergence properties of several stochastic operator-splitting algorithms, including S-GFBS, S-PRS, and S-DRS with dependent data in Chapter 6. We present numerical experiments to examine the performance of our proposed data-driven procedure that relies on a single trajectory of a given stochastic process.

Chapter 2

Data-Driven Stochastic Programming

From here on, we let $\widehat{\Xi}_K = \{\widehat{\xi}_k\}_{k=1}^K$ denote the training dataset. We emphasize that $\widehat{\Xi}_K$ can be viewed as a realization of a random object governed by the distribution $\mathbb{P}$ supported on Ξ. A data-driven version of the problem in 2 is to find a feasible decision $\widehat{x}_K^\star \in \mathbb{X}$ based on the training set $\widehat{\Xi}_K$. We hereafter suppress the dependence of $\widehat{x}_K^\star$ on the training sample in our notation for simplicity.

2.1 Multiple Optimality Guarantees

The statistical guarantees provided by most existing approaches to data-driven optimization critically rely on large-sample approximations and the often-unreasonable assumption that training samples are IID. We argue that any meaningful approach to data-driven optimization should retain asymptotic optimality and finite sample performance guarantees when the training samples are finite in number and display serial dependencies. We define $\mathbb{E}^{\mathbb{P}^K}[\ell(\widehat{x}_K^\star; \xi)]$ as the in-sample risk of $\widehat{x}_K^\star$, which is a function of the training samples alone and therefore accessible to the decision-maker. An ideal optimal solution would minimize the out-of-sample risk $\mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star, \xi)]$, where ξ is independent of the training set, but since $\mathbb{P}$ is unknown, the best we can hope to do is establish tight bounds on the out-of-sample performance.

The feasibility of $\widehat{x}_K^\star$ in 1 implies that $J^\star \leq \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star, \xi)]$, but this lower bound is not particularly useful since $J^\star$ is unknown. Our primary goal is then to bound the risk from above. In other words, we seek data-driven solutions $\widehat{x}_K^\star$ with performance guarantees of the form

$$\mathbb{P}^K\big\{\widehat{\Xi}_K \mid \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star, \xi)] \leq \widehat{J}_K^\star + \varepsilon^\star\big\} \geq 1 - \beta \tag{2.1}$$

for some $\varepsilon^\star > 0$ that represents approximation error caused by deviation in P^K from $\mathbb{P}$, where $\beta \in (0,1)$ is a significance parameter with respect to the distribution P^K that governs both $\widehat{x}_K^\star$ and $\widehat{J}_K^\star$.

We approximate $\mathbb{P}$ with the discrete empirical probability distribution $\widehat{\mathbb{P}}_K = \frac{1}{K}\sum_{k=1}^K \delta_{\widehat{\xi}_k}$, that is, the uniform distribution on $\widehat{\Xi}_K$. In terms of stochastic optimization, this amounts to approximating the optimal solution of the original problem 1–2 with SAA, as in 3–**??**. If the feasible set $\mathbb{X}$ is compact and the loss function is uniformly continuous in x for all $\xi \in \Xi$, then the optimal value $\widehat{J}_K^\star$ and every optimal solution $\widehat{x}_K^\star$ of the SAA problem converge almost surely to their counterparts in the original problem as K tends to infinity [26], [40], [41].

We will demonstrate that the optimal value $\widehat{J}_K^\star$, as well as any optimal solution $\widehat{x}_K^\star$ of the SAA problem in 3–**??**, enjoy the following properties.

1. **Asymptotic consistency:** As the number of data points K tends to infinity, the optimal value $\widehat{J}_K^\star$ and any optimal solution $\widehat{x}_K^\star$ converge, both in expectation and with high probability, to the optimal value $J^\star$ and an optimal solution $x^\star$ of the problem in (1)–(2).
2. **Finite sample guarantee:** Given the finite K training samples, by introducing a weakened version of ϕ-mixing, we establish $1-\beta$ confidence bounds on the out-of-sample performance based on the optimal solution obtained by minimizing an SAA problem.
3. **Tractability:** For many convex cost functions ℓ and sets $\mathbb{X}$, the optimal and an optimal solution can be obtained efficiently through operator-splitting schemes.

The stochastic optimization literature has identified the above properties as desirable for optimal values and solutions [8]

2.2 Data-generating Process

We now introduce our proposed sampling algorithm, which is similar to other approaches in existing literature [24], [37]. In short, we consider a single trajectory of an ergodic stochastic process and design a data-driven procedure that provides attractive performance when the sampling from $\mathbb{P}$ is impossible or expensive. Let $\xi_1, \ldots, \xi_K$ denote a time-indexed sequence observed from a stochastic process P. Let P^k denote the marginal distribution of P at time k: P^k thus converges to $\mathbb{P}$.

Our proposed data-driven procedure for 1–2 is related to a family of stochastic optimizations with exogenous correlated noise, where the goal is to minimize objectives as 1–2, but we have access only to samples ξ that are not independent over time. In general, this stochastic process enables us to generate training samples more efficiently and governs convergence guarantees [24]. There are a number of applications for this work: in control problems, data is often coupled over time or may come from an autoregressive process [38]. in distributed sensor networks [37], a set of wireless sensors attempt to minimize an objective corresponding to a sequence of correlated measurements; and in statistical problems, data comes from an unknown distribution and may be dependent [1]. See for other motivating applications.

We now describe several data-generating processes for which the surrogate optimization problem in 3-**??** or the original optimization problem in 1-2 admit the proposed framework.

2.2.1 Linear Dynamical System

We consider the problem of learning a stable linear dynamical system, where the training samples are generated from a single trajectory of correlated state observations. The problem is of fundamental importance in various operation control, such as adaptive control [31], system identification [44], and learning of stable dynamic programming [47]. Specially, the observable data $\{\xi_k\}_{k=1}^K$ follows the vector-auto regressive process

$$\xi_{k+1} = \theta + A\xi_k + \varepsilon_{k+1}$$

with the state space $\mathbb{S} = \mathbb{R}^h$. Here, the drift term $\theta \in \mathbb{R}^h$ is deterministic but unknown and $(\varepsilon_k)_{k=1}^K$ is normally distributed with a mean of zero and a known positive-definite covariance matrix of $\Sigma \in \mathbb{R}^{h\times h}$. The initial state ξ_0 and all disturbances are mutually independent under $\mathbb{P}$. Such data generating process is also a common model in time series analysis and statistical process control [25], [46]. If we additionally assume that $A \in \mathbb{R}^{h\times h}$ is asymptotically stable (in the sense that all of its eigenvalues reside strictly inside the complex unit circle), then the process $\{\xi_k\}_{k\in\mathbb{N}}$ is ergodic and admits a unique stationary distribution [49].

2.2.2 Peer-to-peer Decentralized Optimization

The well-known Markov incremental gradient descent procedure derives from a distributed optimization algorithm that uses a simple peer-to-peer scheme for optimization and communication [37]. The basic idea is we have n processors or computers, each with a convex function $\ell_i : \mathbb{X} \times \mathbb{R}^d \to \mathbb{R}$. The goal is to minimize the objective function

$$L(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}^{\mathbb{P}_i}[\ell_i(x, \xi_i)]$$

over $x \in \mathbb{X}$, where each expectation is taken with respect to a local distribution $\mathbb{P}_i$. In this procedure, a current set of parameters $x^k \in \mathbb{X}$ is passed between the processors in the network: the token $i(k) \in [n]$ indicates the processor holding x^k at iteration k. At iteration k, a sample $\xi_{k,i(k)}$ is drawn from the local distribution $\mathbb{P}_{i(k)}$ and the algorithm computes the update. We can more generally view the token $i(k)$ as evolving according to a Markov chain, See Appendix A.1 for more detail. The observations $\{\xi_k\}_{k\in\mathbb{N}}$ are serially independent and uniformly distributed under $\mathbb{P}$: the corresponding transition probability matrix P satisfies $(P)_{ij} = 1/h$ for all $i, j \in \Xi$. The process has a uniform stationary distribution. We emphasize that the Markov chain converges to the true (uniform) distribution as K grows. The total variation distance $d_{\mathrm{TV}}(P^k e_i, \mathbf{1}/n)$ between the stochastic process initialized at $i(0) = i$ and the true distribution satisfies

$$d_{\mathrm{TV}}(P^k e_i, \mathbf{1}/n) \le \sqrt{n}\|P^k e_i - 1_h/n\|_2 = \sqrt{n}\|P^k(e_i - 1_h)\|_2 \le \sqrt{n}[\rho_2(P)]^k\|e_i - 1_h/n\|_2 \le \sqrt{n}[\rho_2(P)]^k,$$

where e_i denotes the ith standard basis vector, 1_h an h-vector of ones, and $\rho_2(P)$ the second singular value of P. If $k \ge \frac{\frac{1}{2}\log(Kn)}{\log\rho_2(P)^{-1}}$, then $\|P^k e_i - 1_h/n\|_1 \le 1/\sqrt{K}$, and so $\|P^k e_i - 1_h/n\|_1 \to 0$ as $K \to \infty$.

2.3 Computational Tractability

Although SAA offers powerful statistical guarantees, the method would not be useful if the underlying optimization problem could not be solved efficiently. We now demonstrate that the SAA is computationally tractable for several numerically inconvenient loss functions that are common in practice: we develop a numerical procedure to solve 3 when data comes from an ergodic stochastic process that converges to $\mathbb{P}$.

2.3.1 Problem Statement

We consider two types of stochastic problems. The first is the unconstrained problem

$$\min_{x\in\mathbb{X}}\Big\{\mathbb{E}^{\mathbb{P}}[\ell(x;\xi)] = \int_{\Xi}\ell(x;\xi)\mathbb{P}(\mathrm{d}\xi) = \int_{\Xi}[f(x;\xi) + g(x;\xi)]\mathbb{P}(\mathrm{d}\xi)\Big\}, \tag{2.2}$$

while the second is the linearly constrained problem

$$\min_{x\in\mathbb{X},y\in\mathbb{Y}}\Big\{\mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]=\int_{\Xi}\ell(x;\xi)\mathbb{P}(\mathrm{d}\xi)=\int_{\Xi}f(x;\xi)+g(y;\xi)\mathbb{P}(\mathrm{d}\xi)\mid Ax+By=b\Big\}, \tag{2.3}$$

where $\mathbb{X},\mathbb{Y}\subseteq\mathbb{R}^d$ and ξ is a random vector defined on the probability space $(\Xi,\mathcal{A},\mathbb{P})$. Here, $\Xi\subseteq\mathbb{R}^m$ and $\mathcal{A}$ is a σ-algebra that contains the information we are interested in. We assume that $A:\mathbb{X}\to\mathbb{R}$ and $B:\mathbb{Y}\to\mathbb{R}$ are bounded, linear operators.

Many optimization problems arising in image processing, machine learning, statistics, and other areas [3], [9], [15], [29], [59], [70] can be cast as either 2.2 or 2.3 [11]. In practice, the dimensionality of the data can be extremely large, so traditional methods may fail to efficiently (in terms of time) generate solutions. Penalty terms, including those used to restrict the sparsity, rank, or smoothness of a solution, make problems 2.2 or 2.3 even more difficult to optimize jointly. Even if both terms can be handled jointly, even a single iteration of classical algorithms can be infeasible when the data is high-dimensional or contains millions or billions of samples.

The methodology we present differs from that used in classical convex analysis [11], [12], [57]. Classical statistical methods attempt to estimate unknown parameters through an optimization problem. Our approach uses operator-splitting algorithms, which are driven by fixed-point iterations of a given operator, so convergence rates are with respect to the fixed-point residual (FPR) $\|Tx^k-x^k\|^2$ rather than the goal of minimizing a loss function: convergence is due to the contractive property of the given fixed-point operator rather than "descent" in the loss function. Most fixed-point schemes do not decrease the objective function monotonically, so the objective function's convergence is a consequence rather than a cause of fixed-point convergence.

2.3.2 Geometry of Operator Splitting Schemes: Feasibility Problem

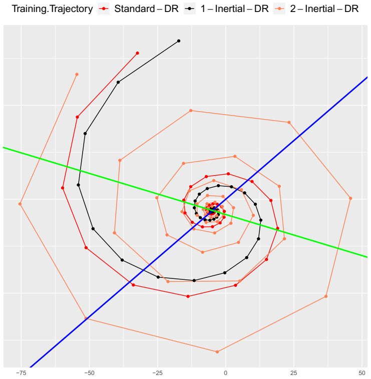
To illustrate fixed-point iteration schemes more clearly, We consider a feasibility problem of two subspaces to (1) provide a geometric explanation of the convergence trajectory of the fixed-point sequence generated by operator-splitting schemes; (2) show that locally, the fixed-point sequence settles onto a regular trajectory such as a logarithmic spiral. We consider the problem in $\mathbb{R}^2$: let $T_1,T_2\subset\mathbb{R}^2$ be two intersecting lines. The problem of finding the common point of T_1,T_2 can be written as

$$\min_{x\in\mathbb{R}^2} l_{T_1}(x)+l_{T_2}(x).$$

As the proximal mapping of indicator functions is projection, the above problem can be easily handled by the Douglas-Rachford splitting method, see Algorithm 2.3.2. The convergence trajectory of sequence $\{x_k\}_{k\in\mathbb{N}}$ is provided as below.

Algorithm 1 Douglas-Rachford splitting

```
**input:** $x^0 \in \mathbb{R}^2$
**while** $j = 0, 1, \ldots, n$ **do**
    $u_g^k \leftarrow \mathrm{Proj}(x^{k-1})$; $u_f^k \leftarrow \mathrm{Proj}(2u_g^k - x^{k-1})$
    $x_1^k \leftarrow x^{k-1} + (u_f^k - u_g^k)$
    $x^k \rightarrow (1 - \alpha_k)x^{k-1} + \alpha_k x_1^k$
**end while**
**return** $x$
```



The above trajectory results are quite different from the "descent" in the loss function, which is evidence that the contractive property of the given fixed-point operator ensures convergence and affects the trajectory of the sequence.

Chapter 3

Statistical Performance Guarantees

3.1 Preliminary and Assumption

Before deriving statistical performance guarantees for the proposed method, we first recall an essential definition from probability theory.

Definition 1 (Total variation distance [19]) *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures defined on a set Ξ with respective densities p and q relative to some underlying measure μ. The total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ is*

$$d_{TV}(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int_{\Xi} |p(\xi) - q(\xi)| \mathrm{d}\mu(\xi) = \sup_{A} |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the supremum is taken over measurable subsets A of Ξ.

We can now describe our notion of a mixing stochastic process. Let $P^k_{[s]}$ denote the distribution of ξ_k conditional on $\mathcal{F}_s = \sigma(\xi_1, \ldots, \xi_s)$. That is, for any measurable $A \subseteq \Xi$, $P^k_{[s]}(A) = P(\xi_k \in A \mid \mathcal{F}_s)$. We use the notion of a mixing coefficient to measure convergence.

Definition 2 (Mixing coefficient) *Define $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and let $(\mathcal{F}_k)_{k=1}^K$ be an increasing sequence of σ-algebras. The ϕ-mixing coefficient of the sample distribution P under the total variation distance is*

$$\phi(v) = \sup_{k\in\mathbb{N}_+, \mathcal{A}\in\mathcal{F}_k} d_{\mathrm{TV}}(P^{v+k}(\cdot \mid \mathcal{A}), \mathbb{P}) = \sup\{A \in \mathcal{F}_k \mid |P^k_{[s]}(A) - \mathbb{P}(A)| < \sigma\}, \tag{3.1}$$

for an arbitrary $\sigma > 0$. Intuitively, a process is ϕ-mixing if $\phi(v) \to 0$ as $v \to \infty$. If the samples are IID, then $\phi(1) = 0$.

Our main assumption is that the stochastic process is suitably mixing, i.e., there is a stationary distribution $\mathbb{P}$ to which the distribution of ξ_k converges as k increases. Below are our main probabilistic assumptions: these guarantee that a stochastic process converges.

Assumption 1 (Ergodicity) *The ϕ-mixing coefficients for the sample distribution are summable, i.e., $\sum_{k=1}^{\infty} \phi(k) < \infty$.*

Assumption 2 (Light-tailed distribution) *There exists $a > 1$ such that,*

$$\mathbb{E}^{\mathbb{P}}[\exp(\|\xi\|^a)] = \int_{\Xi} \exp(\|\xi\|^a)\mathbb{P}(\mathrm{d}\xi) < \infty.$$

Assumption 1 is met by stochastic processes that mix geometrically: such processes, which include autoregressive process and periodic Harris-recurrent Markov processes [49], [50], satisfy $\phi(k) \leq \phi_0 \exp(-\phi_1 k^{\alpha})$ for some $\phi_0 > 0$, $\phi_1 > 0$, and $\alpha > 0$. Of particular note, Definition 2 and Assumption 1 do not require the distribution $\mathbb{P}^k$ to be time-homogeneous and allows randomness in the probability distribution $\mathbb{P}^k$ given the initial s samples. That is, conditional on $\mathcal{F}_s$, the mixing time $\phi(v)$ is an $\mathcal{F}_{[s]}$-measurable random variable. Assumption 2, which ultimately requires that the tails of $\mathbb{P}$ decay at an exponential rate, is a common requirement and is essential for SAA [28]. There are no convergence guarantees for SAA with heavy-tailed distributions that fail to meet Assumption 2 [13], [14]. This assumption holds trivially if Ξ is compact.

3.2 Asymptotic Consistency and Out-of-sample Performance

We now present some modern measure concentration results that provide a basis for more-powerful finite sample guarantees.

Definition 3 (Dependence Coefficient [21]) *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, X a real-valued random variable, and $\mathcal{F}$ a σ-algebra on $\mathcal{A}$. Let $\|\cdot\|_p$ denote the $\mathbb{L}^p$-norm with respect to $\mathbb{P}$ and $\|\cdot\|_{p,\mu}$ the $\mathbb{L}^p$-norm with respect to μ. Let $\mathbb{P}_X$ denote the distribution of X and $\mathbb{P}_{X|\mathcal{F}}$ a regular distribution of X given $\mathcal{F}$. Let $F_X(t) = \mathbb{P}_X((-\infty, t])$ and $F_{X|\mathcal{M}}(t) = \mathbb{P}_{X|\mathcal{F}}((-\infty, t])$. For $p, q \in [1, \infty]$, define*

$$\tau_{\mu,p,q}(\mathcal{F}, X) = \Big\| \Big(\int |F_{X|\mathcal{M}}(t) - F_X(t)|^p \mu(\mathrm{d}t) \Big)^{1/p} \Big\|_q = \Big\| \|F_{X|\mathcal{F}} - F_X\|_{p,\mu} \Big\|_q .$$

Theorem 2 (Measure Concentration [21]) *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(X_i)_{1 \leq k \leq K}$ a sequence of identically distributed, real-valued random variables with a common distribution function F. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and let $(\mathcal{F}_l)_{1 \leq l \leq K}$ be an increasing sequence of σ-algebra such that $\sigma(X_i, 1 \leq k \leq l) \subseteq \mathcal{F}_k$. Define*

$$D_{p,K}(\mu) = \Big(\int |\widehat{F}_K(t) - F(t)|^p \mu(\mathrm{d}t) \Big)^{1/p} = \|\widehat{F}_K - F\|_{p,\mu}.$$

For any $p \in [2, \infty]$, any finite measure μ, and any positive x,

$$\mathbb{P}(\sqrt{K} D_{p,K}(\mu) \geq x) \leq 2 \exp\left(- \frac{K x^2}{2(p-1) \sum_{k=1}^{K} \big(\sum_{l=i}^{N} \big\| \|F_{X_l|\mathcal{F}_i} - F_{X_l|\mathcal{F}_{i-1}}\|_{p,\mu} \big\|_\infty \big)^2} \right).$$

Remark 1 ([21]) *The bound in Theorem 2 is valid for $p \in [2, \infty]$. If $p \in [1, 2)$, then the space $\mathbb{L}^p(\mu)$ is no longer smooth and the method of martingale differences in Banach spaces does not work. However, since $D_{p,K}(\mu) \leq D_{2,n}(\mu)$ for any probability measure μ and any $p \in [1, 2]$, Theorem 2 provides an upper bound for the deviation of $D_{p,K}(\mu)$ in terms of $\tau_{\mu,2,\infty}$ (and hence, also in terms of $\tau_{\mu,1,\infty}$ since $\tau_{\mu,2,\infty}(\mathcal{M}, X) \leq (\tau_{\mu,1,\infty}(\mathcal{M}, X))^{1/2}$).*

Theorem 2 provides a prior estimate of the distribution $\mathbb{P}$ that resides outside of the ε-ball $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_K) = \{\mathbb{Q} \mid \|\widehat{\mathbb{P}}_K - \mathbb{Q}\| \leq \varepsilon\}$. In particular, where $C(p, K, \mu) = \sum_{k=1}^{K} (\sum_{l=k}^{K} (\tau_{\mu,p,\infty}(\mathcal{F}_k, X_l) + \tau_{\mu,p,\infty}(\mathcal{F}_{k-1}, X_l)))^2$, we have the upper bound $\mathbb{P}(\sqrt{K} D_{p,K}(\mu) \geq x) \leq 2 \exp(-K x^2 / [2(p-1) C(p, K, \mu)])$. We can thus use Theorem

2 to estimate the radius of the smallest ε-ball that contains $\mathbb{P}$ with confidence $1-\beta$ for some prescribed $\beta \in (0,1)$:

$$\varepsilon_K(\beta) = \sqrt{\frac{2\log(2\beta^{-1})C(\sum_{k=1}^K \phi(k))}{K^2}}.$$

Let $\|\widehat{\mathbb{P}}_K - \mathbb{P}\| = \int_0^1 |\widehat{\mathbb{P}}_K(t) - \mathbb{P}(t)| \mathrm{d}t$ with $\widehat{\mathbb{P}}_K = \frac{1}{K}\sum_{k=1}^K \delta_{\xi_k}$. If Assumption 1 holds, then by Theorem 2 in [21],

$$\mathbb{P}(\|\widehat{\mathbb{P}}_K - \mathbb{P}\| \geq \varepsilon) \leq 2\exp\Big(-\frac{K^2\varepsilon^2}{2C(\sum_{k=1}^K \phi(k))}\Big)$$

for all $K \geq 1$ and $\varepsilon > 0$, where $C(\sum_{k=1}^K \phi(k))$ is a function of $\sum_{k=1}^K \phi(k)$ and satisfies $C(\sum_{k=1}^K \phi(k)) < \infty$. Therefore, $\varepsilon_K(\beta)$ yields the smallest ε-ball that contains $\mathbb{P}$ with confidence $1-\beta$ for some prescribed $\beta \in (0,1)$.

Theorem 3 (Out-of-sample performance) *Suppose that Assumption 2 holds. Assume also that $\ell(x;\xi)$ is bounded by a constant L for $x \in \mathbb{X}$ and $\xi \in \Xi$. Let $\varepsilon^\star = L\varepsilon_K(\beta) > 0$ and $\beta \in (0,1)$. It follows that*

$$\mathbb{P}(\mathbb{E}^{\mathbb{P}}[\widehat{\Xi}_K \mid \ell(\widehat{x}_K^\star;\xi_{test})] \leq \widehat{J}_K^\star + \varepsilon^\star) \geq 1-\beta. \tag{3.2}$$

Proof: By the definition of $\varepsilon_K(\beta)$, we have that $\mathbb{P}\{\mathbb{P} \in \mathbb{B}_{\varepsilon_K(\beta)}(\widehat{\mathbb{P}}_K)\} \geq 1-\beta$. Thus, by boundedness of ℓ, it follows that $\mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star;\xi)] \leq \mathbb{E}^{\widehat{\mathbb{P}}_K}[\ell(\widehat{x}_K^\star;\xi)] + L\|\widehat{\mathbb{P}}_K - \mathbb{P}\|$. Thus, $\mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star;\xi)] \leq \mathbb{E}^{\widehat{\mathbb{P}}_K}[\ell(\widehat{x}_K^\star;\xi)] + L\varepsilon_K(\beta)$ with probability $1-\beta$. The result of the theorem follows by taking $\varepsilon^\star = L\varepsilon_K(\beta)$. □

We can conclude that the out-of-sample performance of $\widehat{x}_K^\star$ is bounded within a $\varepsilon^\star$-ball about $\widehat{J}_K^\star$ with probability $1-\beta$. It is clear from 3.2 that the approximation error is due to dependence among the samples and the finite sample size. In addition, one can show that if β_K converges to zero at a particular rate, then the solution to the problem in 3 converges to that of 1 as K tends to infinity. The following theorem formalizes this statement.

Theorem 4 (Asymptotic Consistency) *Suppose that Assumption 2 holds. Let $\beta_K \in (0,1)$ with $\lim_{K\to\infty} \varepsilon_K(\beta_K) = 0$ and $\sum_{k=1}^\infty \beta_K < \infty$. Assume also that the loss function ℓ is a convex, closed, and proper function that is bounded by some finite L. Let $\widehat{J}_K^\star$ and $\widehat{x}_K^\star$ represent the optimal value and optimal solution obtained by SAA where ξ_k is drawn from an ergodic stochastic process. It follows that $\mathbb{P}(\lim_{K\to\infty} \widehat{J}_K^\star = J^\star) = 1$ and $\mathbb{P}(\lim_{K\to\infty} \widehat{x}_K^\star = x^\star) = 1$, where $J^\star$ and $x^\star$ are optimal value and optimal solution i 1 and 2.*

A convergence result akin to Theorems 3–4 for distributionally robust optimization is established in [28]. Robust SAA is discussed in [8]. These results are complementary to Theorems 3–4. Indeed, the above results accommodate dependence among training samples—a setting that is not considered in previous works. While we require here that ℓ is a closed, convex, and proper function, our results still hold under the weaker assumption that ℓ is lower semicontinuous in x. We argue that the former assumption is more realistic in many real-world applications, relative to other regularity conditions such as strong convexity, smoothness, and lower semicontinuity.

So far, Theorems 3–4 show that SAA accommodates data-driven procedures with dependent training samples and retains favourable asymptotic and finite-sample guarantees. In practice, however, there are several numerical difficulties.

1. The formula for $\varepsilon^\star$ in 3.2 depends on the mixing coefficient $\phi(v)$ and, thus, on the training samples. Even though $\phi(v)$ can be computed based on Definition 2, since the true distribution $\mathbb{P}$ is unknown, $\phi(v)$ can only be approximated.

2. Even if $\mathbb{P} \notin \mathbb{B}_{\varepsilon_K(\beta)}(\widehat{\mathbb{P}}_K)$, the optimal value $\widehat{J}_K^\star$ may still provide an upper bound on $J^\star$.

3. Unlike the IID setting where asymptotic optimality depends only on the sample size K, in our setting, this optimality also relies on the mixing properties of stochastic progress. Convergence may be faster for certain processes. For example, processes that satisfy $\phi(v) \leq \phi_0 \exp(-\phi_1 v^\alpha)$ for some $\phi_0 > 0, \phi_1 > 0$ and $\alpha > 0$ can converge faster than those where $\phi(v) \leq M v^{-\alpha}$, for some M and $\alpha > 0$.

Chapter 4

Statistical Inference and Sample Complexity of SAA with Non-i.i.d Data

4.1 Statistical Inference for Stochastic Optimization

With the asymptotic consistency in places, we now develop an asymptotic expansion that essentially gives the distributional convergence results. Our approach is fully operational with online/offline data and is rigorously underpinned by a functional covariance inequality and measure concentration theory. We provide more precise inferential guarantees on $\min_{x\in\mathbb{X}} \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]$, which has a couple of key advantages over the existing result. First, we give an extension of asymptotic normality of SAA to the dependent sequence, thereby allowing for hypothesis tests that can be used to construct uncertainty sets for the optimization problems. Second, the inference procedure can be computed in an online fashion that is efficient implementation suitable for massive online data.

4.1.1 Functional Covariance Inequality and Assumptions

Our results on inferential guarantees are built on asymptotic expansions, which we now present. Without additional conditions. It is challenging to provide more precise inferential convergence on $\min_{x\in\mathbb{X}} \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]$. We then prove a crucial functional covariance and exponential inequality for the loss functions and deduce an ϕ-mixing property for the stochastic process, P. These inequalities are the main tools for proving the functional central limit theorem and sample complexity.

Lemma 1 (Functional Covariance Inequality) *Let $\{\xi_k\}_{k=1}^{K}$ be a ϕ-mixing sequence and $\xi_k, \xi_{n+k} \in \mathbb{R}^m$ be measurable w.r.t. to $\mathcal{F}_1^k$ and $\mathcal{F}_{n+k}^{\infty}$ respectively. Suppose a function $f : \mathbb{R}^m \to \mathbb{R}$ is Borel measurable. Then, we have that*

$$|\mathrm{Cov}(f(\xi_k), f(\xi_{k+n}))| \leq 4\|f\|_M^2 \phi(n),$$

where $\|f\|_M = \sup_{(x,\xi)}|f(x,\xi)|$ is the bound of f, which we assume it is finite.

In order to study the statistical properties of $\ell(x;\xi)$, for positive integers $1 \leq p = p(K) \leq K$ and $p \to \infty$, we decompose the set $\{1,\ldots,K\}$ into successive blocks each containing p elements. Considering that $r = r(K)$ be the largest integer with $0 < r < K$, $r \to \infty$, and $2pr \leq K$ which implies that $\frac{K}{2pr} \to 1$, we denote that $Y_k = \ell(x,\xi_k) - \mathbb{E}\ell(x;\xi)$ and $\bar{S}_K = \frac{1}{K}\sum_{k=1}^{K}(\ell(x,\xi_k) - \mathbb{E}\ell(x;\xi)) := \frac{1}{K}S_K$. Let

$$U_i = Y_{2(i-1)p+1} + \cdots + Y_{(2i-1)p}, \quad V_i = Y_{(2i-1)p+1} + \cdots + Y_{2ip}, \quad i = 1,\cdots,r; \tag{1}$$

$$W_K = Y_{2pr+1} + \cdots + Y_K,$$

with

$$\bar{U}_K = \frac{1}{K}\sum_{i=1}^{r} U_i, \quad \bar{V}_K = \frac{1}{K}\sum_{i=1}^{r} V_i, \quad \bar{W}_K = \frac{1}{K}W_K, \tag{2}$$

so that

$$\bar{S}_K = \bar{U}_K + \bar{V}_K + \bar{W}_K. \tag{3}$$

We now show the assumptions necessary to formulate the exponential tail bound for $\bar{S}_K$.

Assumption 3 *For all x, the loss function $\ell(x;\xi)$ is Borel measurable and bounded by a constant M, i.e. $\sup_{x,\xi}|\ell(x;\xi)| \leq M$.*

Assumption 4 *$\{\xi_k\}_{k\geq 1}$ is strictly stationary, $\phi(n) \leq \phi_0 \exp(-\phi_1 n^a)$ for some constant $\phi_0 > 0, \phi_1 > 0, a > 0$. That is, $\{\xi_k\}$ is a geometric mixing sequence.*

Assumption 5 *The group size, number of groups satisfy that $2pr \leq K$, $p/K \to 0$, $K/(2pr) \to 1$ as $K \to \infty$.*

Lemma 2 (Exponential Inequality) *Let $\bar{S}_K$ be defined by (3). For $\varepsilon > 0$, suppose Assumptions (3)-(5) hold. Then, there exists a constant $C_0 > 0$ such that for any fixed $x \in \mathbb{R}^d$,*

$$\mathbb{P}(|\bar{S}_K| \geq \varepsilon) \leq C_0 \exp(-\frac{K\varepsilon^2}{36p\|f\|_M^2}).$$

4.1.2 Functional Central Limit Theorem

We show a uniform variant of the asymptotic normality. While our results apply significantly more generality, the following results cover many practical instances of stochastic optimization problems. The exists a well-developed statistical inference for the optimal objective value obtained from the SAA approach. [61] develops a number of normal approximations and asymptotic normality theory for the stochastic optimization problems, which we now present

$$\sqrt{n}\big(\min_{x\in\mathbb{X}} \mathbb{E}^{\widehat{\mathbb{P}}}[\ell(x;\xi)] - \min_{x\in\mathbb{X}} \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]\big) \sim \mathcal{N}(0, \mathrm{Var}_{\mathbb{P}}(\ell(x^\star;\xi))).$$

Now, we extend the asymptotic behaviour of the SAA optimal objective value for the present dependence structure.

Theorem 5 (Functional Central Limit Theorem) *We suppose that the assumptions* (3)-(5) *hold and, additionally, we assume that,* $\frac{1}{K}\, Var(\sum_{k=1}^{K} \ell(x^\star, \xi_k)) \to Var_{\mathbb{P}}(\ell(x^\star;\xi))$. *Then,*

$$\frac{1}{\sqrt{K}}\left(\sum_{k=1}^{K}\left(\ell(\widehat{x}_K^\star, \xi_k) - \min_{x\in\mathbb{X}} \mathbb{E}^{\mathbb{P}}[\ell(x;\xi)]\right)\right) \to \mathcal{N}(0,\, Var_{\mathbb{P}}(\ell(x^\star;\xi))).$$

With the uniqueness of the optimal solution, the above theorem implies that the optimal objective value of SAA is asymptotically normally distributed, which also indicates that, under the uniform integrability conditions, we have the vanishing rate of bias of $O(K^{-1/2})$. While the SAA optimal value enjoys asymptotic properties, the asymptotic distribution of the optimal solution critically depends on the functional properties of objectives, see [2], [40].

4.2 Sample Complexity of SAA with Non-IID Data

While a study of asymptotic properties is a crucial aspect of understanding the behaviour of SAA, non-asymptotic analysis, such as analysis of finite-sample error bounds, is more valuable as they provide a better understanding of the factors that influence the learning performance with limited sample size. Indeed, these results are useful for a myriad of reasons. For example, we can avoid making bad decisions or false inferences, we may realize that a learning task is unrealistic, and we can explicitly the amount of data necessary for solving different statistical problems. These results typically characterize the error bound of the difference between the true and empirical form of a function in the aforementioned class and especially for evaluating their suprema, which can be thought of as a measure of the worst-case approximated performance of functions.

With the notability of [18], [23], [51], learning complexity has been analyzed in various statistical settings. These error bounds typically provide the theoretical guarantee, with high probability, that the error in an estimator is bounded by an empirical estimated of bias with a penalty term depending on some notation of complexity of the class of functions, as well as the learning algorithms. However, many notations of complexity might depend on the underlying probability measure that is rarely known, or be independent of any characteristics of data. Thus, it is desirable to obtain data-dependent estimates which can readily be computed from the sample. The goal is then to obtain the sharpest possible estimates against the complexity of function classes empirically. In a decision-theoretic setting, we establish general risk bounds in terms of complexities for SAA. We show the deviation of SAA against the ground truth can be bounded by these data-dependent complexities in terms of empirical Rademacher complexity (typically a function in a certain class) and precise sample complexities of given objects, and the established results can be extended to function classes that can be expressed as combinations of functions from basic classes.

4.2.1 Measurement of Sample Complexity

Beyond asymptotic properties, we prove general risk bounds in terms of these complexities that provide a unified perspective enabling us to characterize the finite sample properties. We consider function classes that can be expressed as combinations of functions from basic classes and show how the Rademacher and local complexities of such a function class can be bounded in terms of the complexity of the basic classes. The notion of complexity of a function class might depend on the (unknown) underlying probability measure from which the data are produced. However, estimating these distribution-dependent quantities can be complicated in practice, especially, when $\mathbb{P}$ is unknown. Thus, data-dependent estimates that can be readily computed from a sample are desirable. We now present bounds on the error in terms of an empirical notion of complexity, namely Empirical Rademacher complexity, which are defined *w.r.t* the distribution generating the data. Empirical Rademacher complexity can be estimated reliably from data generated, and they yield generalization error bounds in terms of data-dependent quantities that can often be estimated or bounded empirically [4], [42]. As we consider a general form of ℓ, we attempt to find the sufficient conditions for learning that led us to prove results as uniform convergence bounds in terms of a function class $\mathbb{L}$ and decision set. Typically, The error bound over the function class state that empirical errors of objects from a given class converge to their true errors in terms of Empirical Rademacher complexity. This shows the difference between the training error and generalization error for all functions belonging to $\mathbb{L}$.

In statistical learning theory, the generalization error will be based on the measure of Rademacher complexity of a class of functions $\mathbb{L}$ with respect to sample set $\{\xi_k\}_{k=1}^K$ with high probability. The Empirical Rademacher complexity is given as

Definition 4 (Empirical Rademacher Complexity) *Given $\widehat{\Xi} = \{\xi_k\}_{k=1}^K$, the empirical Rademacher complexity of a class of real-valued functions $\mathbb{L}$ defined over a set Ξ is defined as follows*

$$\widehat{R}_K(\mathbb{L}) = \frac{1}{K}\mathbb{E}_\epsilon \left[\sup_{\ell\in\mathbb{L}} \sum_{k=1}^K \epsilon_k \ell(x,\xi_k) \mid \widehat{\Xi} = (\xi_1,\cdots,\xi_K)\right].$$

The expectation is taken over $\{\epsilon_k\}_{k=1}^K = (\epsilon_1,\ldots,\epsilon_K)$ where ϵ_k s are random variables taking values in $\{-1,+1\}$ called Rademacher random variables. The Rademacher complexity of a hypothesis set $\mathbb{L}$ is defined as the expectation of $R_K(\mathbb{L})$ over all samples of size K

$$R(\mathbb{L}) = \mathbb{E}_\xi \left[\mathbb{E}_\epsilon \left[\sup_{\ell\in\mathbb{L}} \epsilon_k \ell(x;\xi)\right]\right].$$

The Rademacher complexity measures the ability of a class of functions to fit noise. The empirical Rademacher complexity has the added advantage that it is data-dependent and can be measured from .finite samples

Theorem 6 (Data-dependent Complexity of SAA with non-IID Data) *Assume that $\ell : \mathbb{X}\times\Xi \to \mathbb{R}$ is Borel measurable for all x, and uniformly bounded by a constant M, i.e. $\sup_{x,\xi} |\ell(x;\xi)| \leq M$, $\forall \ell(\cdot) \in \mathbb{L}$. Let $\mathbb{L}$ be a class of functions that contains ℓ. Then for a fixed sample size K, with probability at least $1-\beta$, we have that*

$$\sup_{\ell\in\mathbb{L}} \Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^K \ell(x,\xi_k)\Big) \leq 4\widehat{R}_K(\mathbb{L}) + 3M\sqrt{\frac{2\log\frac{2}{\beta}}{K}}.$$

Proof: Consider we have that

$$\mathbb{P}\Big(\text{there exists } \ell \in \mathbb{L} \text{ such that } \Big|\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)\Big| > \varepsilon\Big) = \mathbb{P}\Big(\sup_{\ell\in\mathbb{L}}|\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)| > \varepsilon\Big)$$

$$= \mathbb{P}\Big(\sup_{\ell\in\mathbb{L}}\Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)\Big) > \varepsilon\Big) + \mathbb{P}\Big(\sup_{\ell\in\mathbb{L}}\Big(\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) - \mathbb{E}[\ell(x;\xi)]\Big) > \varepsilon\Big).$$

Let $\Xi = \{\xi_1, \cdots, \xi_k, \cdots, \xi_K\}$, $\Xi' = \{\xi_1, \cdots, \xi_k', \cdots, \xi_K\}$. Denote

$$\Phi(\Xi) = \sup_{\ell\in\mathbb{L}}\Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)\Big),$$

we have

$$\Phi(\Xi) - \Phi(\Xi') \le \sup_{\ell\in\mathbb{L}}\Big(\frac{1}{K}\ell(x,\xi_k) - \frac{1}{K}\ell(x,\xi_k')\Big) \le 2M/K.$$

By McDiarmid's type inequality [54], we have that with probability $1-\beta/2$,

$$\Phi(\Xi) - \mathbb{E}[\Phi(\Xi)] \le M\sqrt{\frac{2\log\frac{2}{\beta}}{K}}. \tag{1.1}$$

Let $R_K(\mathbb{L}) = \mathbb{E}_\Xi[\widehat{R}_K(\mathbb{L})]$, we next prove $\mathbb{E}[\Phi(\Xi)] \le 2R_K(\mathbb{L})$ and the bound of $R_K(\mathbb{L})$. Let $\{\xi_k'\}_{k=1}^K$ be an i.i.d. copy of $\{\xi_k\}_{k=1}^K$. We have that

$$\begin{aligned}
&\mathbb{E}_\xi\Big[\sup_{\ell\in\mathbb{L}}\Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)\Big)\Big] = \mathbb{E}_\xi\Big[\sup_{\ell\in\mathbb{L}}\mathbb{E}_{\xi'}\Big[\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k') - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) \mid \xi_1,\ldots,\xi_k\Big]\Big]\\
&\le \mathbb{E}_{\xi,\xi'}\Big[\sup_{\ell\in\mathbb{L}}\frac{1}{K}\sum_{k=1}^{K}(\ell(x,\xi_k') - \ell(x,\xi_k))\Big] = \mathbb{E}_{\epsilon,\xi,\xi'}\Big[\sup_{\ell\in\mathbb{L}}\frac{1}{K}\sum_{k=1}^{K}\epsilon_k(\ell(x,\xi_k') - \ell(x,\xi_k))\Big]\\
&\le \mathbb{E}_{\epsilon,\xi'}\Big[\sup_{\ell\in\mathbb{L}}\Big(\frac{1}{K}\sum_{k=1}^{K}\epsilon_k\ell(x,\xi_k')\Big)\Big] + \mathbb{E}_{\epsilon,\xi}\Big[\sup_{\ell\in\mathbb{L}}\Big(\frac{1}{K}\sum_{i=1}^{K}\epsilon_k\ell(x,\xi_k)\Big)\Big]\\
&= 2R_K(\mathbb{L}).
\end{aligned} \tag{1.2}$$

The first inequality holds since sup is a convex function, we can apply Jensen's Inequality to move the sup inside the expectation. It follows from inequalities (1.1), (1.2) that

$$\sup_{\ell\in\mathbb{L}}\Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k)\Big) \le 2R_K(\mathbb{L}) + M\sqrt{\frac{2\log\frac{2}{\beta}}{K}}. \tag{1.3}$$

On the other hand, observe that

$$\begin{aligned}
&\mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}}K^{-1}\sum_{k=1}^{K}\epsilon_k\ell(x,\xi_k)\Big] - \mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}}K^{-1}\sum_{k=1}^{K}\epsilon_k\ell(x,\xi_k')\Big]\\
=&\mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}}K^{-1}\sum_{k=1}^{K}\epsilon_k\ell(x,\xi_k) - \sup_{\ell\in\mathbb{L}}\Big(K^{-1}\sum_{k\neq k'}\epsilon_k\ell(x,\xi_k) + K^{-1}\epsilon_k\ell(x,\xi_k')\Big)\Big]\\
\le&\mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}}K^{-1}\epsilon_k(\ell(x,\xi_k) - \ell(x,\xi_k'))\Big] \le 2M/K.
\end{aligned}$$

Similarly, we can prove that $\mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}} K^{-1}\sum_{k=1}^K \epsilon_k \ell(x,\xi_k')\Big] - \mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}} K^{-1}\sum_{k=1}^K \epsilon_k \ell(x;\xi)\Big] \le 2M/K$, and thus

$$\bigg|\mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}} K^{-1}\sum_{k=1}^K \epsilon_k \ell(x,\xi_k')\Big] - \mathbb{E}_\epsilon\Big[\sup_{\ell\in\mathbb{L}} K^{-1}\sum_{k=1}^K \epsilon_k \ell(x;\xi)\Big]\bigg| \le 2M/K.$$

By applying McDiarmid's type inequality [54], we have that for any $\varepsilon > 0$,

$$\mathbb{P}(\widehat{R}_K(\mathbb{L}) - R_K(\mathbb{L}) \ge \varepsilon) \le e^{-\frac{\varepsilon^2 K}{2M^2}}.$$

This implies that with probability $1-\beta/2$,

$$R_K(\mathbb{L}) \le \widehat{R}_K(\mathbb{L}) + M\sqrt{\frac{2\log\frac{2}{\beta}}{K}}. \tag{1.4}$$

Inequalities (1.3) and (1.4) together implies that

$$\sup_{\ell\in\mathbb{L}}\Big(\mathbb{E}[\ell(x;\xi)] - \frac{1}{K}\sum_{k=1}^K \ell(x,\xi_k)\Big) \le 4\widehat{R}_K(\mathbb{L}) + 3M\sqrt{\frac{2\log\frac{2}{\beta}}{K}}.$$

□

One shortcoming of RAs is that they provide global estimates of the complexity of a class of functions: they do not reflect the fact that an algorithm will likely pick functions that have a small error and, in particular, only a small subset of the class will be used. As a result, the best error rate that can be obtained (with respect to RAs) is suboptimal in some situations. In this section, we establish the local sample complexity of SAA.

We now analyze the number of samples required for the SAA solution to be ε-optimal with respect to the solution of the original problem with high probability.

Theorem 7 (Uniform Convergence) *Assume that $\ell(\cdot,\xi)$ is L-Lipschitz continuous and that the decision set $\mathbb{X} \subseteq \mathbb{R}^d$ has a finite diameter $D(\mathbb{X}) > 0$. Suppose the assumptions (3)-(5) hold, then there exists a constant $C_0 > 0$ such that*

$$\mathbb{P}\Big(\sup_{x\in\mathbb{X}}\Big|\frac{1}{K}\sum_{k=1}^K \ell(x,\xi_k) - \mathbb{E}[\ell(x;\xi)]\Big| > \varepsilon\Big) \le C_0\Big(\frac{4LD(\mathbb{X})}{\varepsilon}\Big)^d \exp\Big(-\frac{K\varepsilon^2}{144pM^2}\Big).$$

Proof: The idea of the proof is that we first construct a v-net to remove the supremum over x and then use the inequality given in Lemma 2 to bound the probability in the statement of the theorem.

Pick a v-net $\{x_l\}_{l=1}^Q$ on the decision set $\mathbb{X}$ such that $Lv = \varepsilon/4$. Thus, $Q \le O(1)(4LD(\mathbb{X})/\varepsilon)^d$. By definition, for all $x \in \mathcal{X}$, there exists $l(x) \in [Q]$ with $[Q] := \{1,\cdots,Q\}$ such that $\|x - x_{l(x)}\|_2 \le v = \varepsilon/(4L)$. By the Lipschitz continuity of ℓ,

$$\Big|\frac{1}{K}\sum_{k=1}^K \ell(x,\xi_k) - \frac{1}{K}\sum_{k=1}^K \ell(x_{l(x)},\xi_k)\Big| \le \frac{\varepsilon}{4}, \text{ and } |\mathbb{E}[\ell(x;\xi)] - \mathbb{E}[\ell(x_{l(x)},\xi)]| \le \frac{\varepsilon}{4}.$$

Hence, for any $x \in \mathbb{X}$,

$$\begin{aligned}\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) - \mathbb{E}[\ell(x;\xi)]\Big| \leq &\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) - \frac{1}{K}\sum_{k=1}^{K}\ell(x_{l(x)},\xi_k)\Big| \\ &+\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x_{l(x)},\xi_k) - \mathbb{E}[\ell(x_{l(x)},\xi)]\Big| + |\mathbb{E}[\ell(x_{l(x)},\xi)] - \mathbb{E}[\ell(x;\xi)]| \\ \leq &\frac{\varepsilon}{2} + \Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x_{l(x)},\xi_k) - \mathbb{E}[\ell(x_{l(x)},\xi)]\Big| \\ \leq &\frac{\varepsilon}{2} + \max_{l\in[Q]}\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x_l,\xi_k) - \mathbb{E}[\ell(x_l,\xi)]\Big|.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}\Big(\sup_{x\in\mathbb{X}}\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) - \mathbb{E}[\ell(x;\xi)]\Big| > \varepsilon\Big) &\leq \mathbb{P}\Big(\max_{l\in[Q]}\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x_l,\xi_k) - \mathbb{E}[\ell(x_l,\xi)]\Big| > \frac{\varepsilon}{2}\Big) \\ &\leq \sum_{l=1}^{Q}\mathbb{P}\Big(\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x_l,\xi_k) - \mathbb{E}[\ell(x_l,\xi)]\Big| > \frac{\varepsilon}{2}\Big).\end{aligned}$$

This inequality together with the result in Lemma 2 reduces to

$$\mathbb{P}\Big(\sup_{x\in\mathbb{X}}\Big|\frac{1}{K}\sum_{k=1}^{K}\ell(x,\xi_k) - \mathbb{E}[\ell(x;\xi)]\Big| > \varepsilon\Big) \leq C_0 Q\exp(-\frac{K\varepsilon^2}{144pM^2}).$$

Taking $Q \leq O(1)(4LD(\mathbb{X})/\varepsilon)^d$ yields the desired result. □

Corollary 1 (Uniform Sample Complexity of SAA with Non-IID Data) *Assume that the assumptions of Theorem 7 hold. With probability at $1-\beta$, the solution to the SAA problem is ε-optimal with respect to the solution of the original problem if the sample size K satisfies*

$$K \geq O(1)\frac{p}{\varepsilon^2}\Big[d\log\Big(\frac{4LD(\mathbb{X})}{\varepsilon}\Big) + \log\Big(\frac{1}{\beta}\Big)\Big]$$

If we ignore the log-factors in Corollary 1 with fixed group size p, we can conclude that the sample complexity of SAA in achieving an ε-optimal solution with dependent samples is $O(d/\varepsilon^2)$.

Part II

Algorithmic Analysis via Modern Monotone Operator Theory

Chapter 5

Computational Tractability

We further investigate the stability, global convergence, and generalizability of a family of operator-splitting schemes based on the (KM) iteration (e.g., FBS, PRS, DRS, and ADMM). We do so in a stochastic setting for the purpose of solving large-scale optimization problems with dependent training samples. We focus on contraction properties, convergence rates, and the quality of approximated solutions under different operator-splitting methods. While convergence results for these algorithms already exist in the literature [52], [58], [66], they hold under specific assumptions on the splitting operator (e.g., that the fixed-point operator is contractive or firmly nonexpansive) or the objective function (e.g., strong convexity or Lipschitz continuity). Our work goes beyond these assumptions to a more general iterative scheme that requires only nonexpansiveness and convexity of the operator and the objective function, respectively.

The KM iteration was first introduced by Krasnosel'skiǐ and Mann [43], [48] and frequently appears in convex optimization. Many existing algorithms can be cast into this framework, including gradient descent [59], the proximal point method [53], and various decomposition methods such as FBS [55]. For a comprehensive survey of these methods and their applications, see [6].

Table 5.1: Overview of several first-order algorithms

Algorithm	Operator identity	Subgradient identity
SGD ($g = 0$)	$I - \gamma \nabla f$	$x^{k+1} = x^k - \gamma_k \nabla f(x^k)$
PPA ($g = 0$)	$(I + \gamma \partial f)^{-1}$	$x^{k+1} = \mathrm{prox}_{\gamma_k f}(x^k)$
PGD	$(I + \gamma \partial g)^{-1}(I - \gamma \nabla f)$	$x^{k+1} = \mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f(x^k))$
DRS	$(I + \gamma \partial f)^{-1}[(I + \gamma \partial g)^{-1}(I - \gamma \partial f) + \gamma \partial f]$	$x^{k+1} = \frac{1}{2} x^k + \frac{1}{2}\, \mathrm{refl}_{\gamma_k \partial f} \circ \mathrm{refl}_{\gamma_k \partial g}(x^k)$
Relaxed PRS	$(I + \gamma \partial f)^{-1}(I - \lambda \partial g)(I + \lambda \partial g)^{-1}(I - \lambda \partial f)$	$x^{k+1} = (1 - \lambda_k) x^k + \lambda_k\, \mathrm{refl}_{\gamma_k \partial f} \circ \mathrm{refl}_{\gamma_k \partial g}(x^k)$

The stochastic setting considered in this paper is motivated by the increased focus on methods using stochastic gradients in statistical machine learning [3], [9], [15], [29], [59], [70]. In this setting, an explicit parameterization of the objective function is unavailable and classical gradient- or Hessian-based methods may be expensive or intractable. It is standard to apply gradient descent (SGD), originally developed by Robbins and Monro [56] for smooth stochastic approximation problems, to approximate the gradient using independent samples. From an initial point x_0, SGD proceeds by drawing $\xi_k \overset{\text{IID}}{\sim} \mathbb{P}$ and updating $x^{k+1} = x^k - \gamma_k L^k$ for some $L^k \in \partial(\ell(x^k; \xi_k))$. Methods based on stochastic (sub)gradients have established convergence guarantees and have seen much empirical success in the literature [33], [39]. However, these methods do present some notable difficulties: they rarely adapt to nuanced aspects of numerical difficulty and, for some objective functions, can diverge due to the choice of step size.

Definition 5 (Nonexpansive Operator) *A mapping $T : \mathbb{X} \to \mathbb{X}$ is nonexpansive if $\|Tx - Ty\| \leq \|x - y\|$ holds for all $x, y \in \mathbb{X}$.*

For a given nonexpansive operator T, define $\mathcal{X}^* = \mathrm{Fix}(T) = \{x \in \mathbb{X} : x = T(x)\}$. We assume that $\mathcal{X}^* \neq \emptyset$.

Definition 6 (Stochastic iteration) *Let T be a nonexpansive operator with a non-empty set of fixed points (i.e., $\mathrm{Fix}(T) \neq \emptyset$). The stochastic iteration is as given in Algorithm* **??***,*

$$x^{k+1} = T_{\lambda_k, \epsilon_k}(x^k, \xi) = x^k + \lambda_k(T(x^k, \xi) - x^k + \epsilon_k) = T_{\lambda_k}(x^k, \xi) + \lambda_k \epsilon_k,$$

where ϵ_k is the error of approximating $T(x^k)$, and ξ is a sequence of random variables taking values in Ξ.

The following fundamental definitions and properties can be found in [6]. Given a convex, closed, and proper function f, let $\partial f(x)$ denote the set of subgradients $\widetilde{\nabla} f(x)$ at x. The convex conjugate of a convex, closed, and proper function f is $f^*(y) = \sup_{x \in \mathbb{X}} \langle y, x \rangle - f(x)$. Let $I_{\mathbb{X}}$ denote the identity map on $\mathbb{X}$. For any point $x \in \mathbb{X}$ and $\gamma \in \mathbb{R}_+$, define the proximal and reflection operators as $\mathrm{prox}_{\gamma,f}(x) = \arg\min_{y \in \mathbb{X}} f(y) + (2\gamma)^{-1}\|y - x\|^2$ and $\mathrm{refl}_{\gamma,f} = 2\,\mathrm{prox}_{\gamma,f} - I_{\mathbb{X}}$, respectively. Define the PRS operator as $T_{\mathrm{PRS}} = \mathrm{refl}_{\gamma,f} \circ \mathrm{refl}_{\gamma,g}$.

We further introduce ergodic stochastic as Algorithm 5, which adopts a stochastic operator-splitting scheme to deal with non-IID data in a fast, online and efficient manner. The results of this paper apply not only to existing algorithms that are based on atomic evaluations of the proximal and gradient operators but also to algorithms that can be viewed as fixed-point iterations via a nonexpansive operator. Let $\lambda_k \in (0, 1)$ be a sequence of relaxation parameter values and choose x^0 arbitrarily from $\mathbb{X}$. The S-KM iteration of T with data ξ_k generated from P^k at time k is

$$\begin{aligned} x^k &= T_{\lambda,\epsilon}(x^{k-1}; \xi_k) = x^{k-1} + \lambda_{k-1}(T(x^{k-1}; \xi_k) - x^{k-1} + \epsilon_{k-1}) \\ &= T_{\lambda_k}(x^{k-1}; \xi_k) + \lambda_{k-1}\epsilon_{k-1}, \end{aligned} \tag{5.1}$$

where the stochastic error ϵ_{k-1} is caused by randomness in sampling. We include ξ_k as an argument of T to explicitly indicate that the kth iteration depends only on the sample drawn most recently.

Algorithm 2 Stochastic iteration (S-KM): point-wised iteration

```
input: initial value $x^0$, $\delta$ optimality parameter

while $\|T(x^{k-1}; \xi^k) - x^{k-1}\|^2 > \delta$ do
    sample $\xi_k \sim P^k_{[s]}$
    $x^k \leftarrow x^{k-1} + \lambda_{k-1}(T(x^{k-1}; \xi^k) - x^{k-1})$
end while
return $x$
```

Algorithm 3 Stochastic iteration (S-KM): ergodic iteration

```
input: initial value $x^0$, $\delta$ optimality parameter

while $\|T(\bar{x}^{k-1}; \xi^k) - \bar{x}^{k-1}\|^2 > \delta$ do
    sample $\xi_k \sim P^k_{[s]}$
    $x^k \leftarrow \bar{x}^{k-1} + \lambda_{k-1}(T(\bar{x}^{k-1}; \xi^k) - \bar{x}^{k-1})$
    $\bar{x}^k \longrightarrow \frac{k-1}{k}\bar{x}^{k-1} + \frac{1}{k}x^k$
end while
return $x$
```

We now provide a comprehensive analysis of the convergence rates of stochastic operator-splitting schemes under the mild assumptions of strong convexity, subdifferentiability, and the existence of a feasible solution.

We show that stochastic operator-splitting schemes automatically adapt to the regularity of a problem with dependent samples and achieve convergence guarantees that also hold for independent samples. Define the fixed-point residual (FPR) as $e_k = \|T(z^k) - z^k\|$ and the residual as $r_k = z^{k+1} - z^k$. From this and the above definition, it follows that $e_k^2 = \|r_k - \lambda_k \epsilon_k\|^2/\lambda_k^2$. The main goal of an operator-splitting algorithm is to construct a monotonic, nonexpansive operator $T : \mathbb{X} \to \mathbb{X}$ that can be used as a building block for complex computations.

In general, algorithms based on a nonexpansive operator fail to converge without additional restrictions. To ensure the convergence of an algorithm based on the contraction property of averaged operator Proposition 5. We can also see the average nonexpansive operator which can be seen as a relaxed version of nonexpansive operators (Proposition 5).

Let $T : \mathbb{X} \to \mathbb{X}$ be a nonexpansive operator. Then for all $\lambda \in (0, 1]$ and $(x, y) \in \mathbb{X} \times \mathbb{X}$, the averaged operator T_λ satisfies

$$\|T_\lambda x - T_\lambda y\|^2 \leq \|x - y\|^2 - \frac{1-\lambda}{\lambda}\|(I_{\mathbb{X}} - T_\lambda)x - (I_{\mathbb{X}} - T_\lambda)y\|^2. \tag{5.2}$$

A mapping $Q : \mathbb{X} \to \mathbb{X}$ is α-averaged if there exists a nonexpansive mapping $R : \mathbb{X} \to \mathbb{X}$ and $\alpha \in (0, 1)$ such that $Q = (1 - \alpha)I_{\mathbb{X}} + \alpha R$.

Remark 2 *An operator $N : \mathbb{X} \to \mathbb{X}$ satisfies 5.2 (with N in place of T_λ) if and only if it is λ-averaged. If $\lambda = 1/2$, then T_λ is called firmly nonexpansive. A rearrangement of 5.2 shows that a nonexpansive operator T is firmly nonexpansive if and only if, for all $x, y \in \mathbb{X}$, $\|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y\rangle$.*

Because splitting algorithms are driven by fixed-point operators, it is natural to perform an analysis in terms of FPRs, which are related to differences between successive KM iterates through $x^{k+1} - x^k = \lambda_k(T(x^k) - x^k)$. In the stochastic setting, we lose some of the usual properties of FPRs such as summability and monotonicity due to approximation error. In first-order algorithms, the FPR is typically related to the norm of the gradient of the (convex) objective function. For example, $x^k = x^{k-1} - \nabla f(x^{k-1})$ in unit-step gradient descent algorithms and so the FPR is $\|\nabla f(x^{k-1})\|^2$. The FPR for the proximal point algorithm is $\|\widetilde{\nabla} f(x^{k+1})\|$, where $\widetilde{\nabla} f(x^{k+1}) = (x^k - x^{k+1}) \in \partial f(x^{k+1})$. When the objective function is a sum of multiple functions, the FPR is a combination of the (sub)gradients of those functions. Thus, the convergence of the FPR naturally implies the convergence of $\|x^{k+1} - x^*\|^2$. This motivates us to establish the properties of the FPR in the following analysis.

We first list a few assumptions that are standard in the analysis of stochastic KM iterations.

Assumption 6 *The objective functions used in optimization are additive and $f, g : \mathcal{H} \to (-\infty, \infty]$ are convex, closed, and proper functions.*

Assumption 7 (Subdifferentiability) *Each function is subdifferentiable. Unless otherwise stated, we do not require the function to be differentiable.*

An operator-splitting algorithm can be applied directly to the closed, convex, and proper functions $f, g : \mathbb{X} \to (-\infty, \infty]$ by applying a splitting method to $\partial f(x)$ and $\partial g(x)$ under the condition that $\partial(f+g)(x) =$

$\partial f(x) + \partial g(x)$. However, this condition does not hold when the objective function includes both a smooth and nonsmooth function, so we impose Assumption 8.

Assumption 8 (Splittability and Solvability) *A convex optimization problem is primal splittable and solvable if* $\operatorname{zer}(\partial f + \partial g) = \operatorname{zer}(\partial(f+g))$ *and* $\operatorname{zer}(\partial(f+g)) \neq \emptyset$*, respectively.*

In some cases [57], the splittability condition makes Assumption 8 slightly stronger than the assumption of the existence of a minimizer. If Assumption 8 is not satisfied, then it is possible for the primal problem to have an optimal solution and $\partial f(x) + \partial g(x)$ to have no zero points. In this case, a splitting algorithm would not be useful. Under Assumption 8, Proposition 5 establishes that the zero points of $\partial f(x) + \partial g(x)$ and $\partial(f+g)$ are equal. As a general note, any fixed-point iteration with a well-defined operator-splitting operator requires certain regularity conditions on the objective function or the direct assumption that the primal problem is solvable and splittable.

Any zero of the operator $\partial f(x) + \partial g(x)$ is an optimal solution to the problems in 2.2–2.3. If Assumption 3 holds, then every optimal solution of the problems in 2.2–2.3 is a zero of $\partial f(x) + \partial g(x)$.

Proof: By Theorem 23.8 in [57], $\partial f(x) + \partial g(x) \subseteq \partial(f+g)$, so any zero of $\partial f + \partial g$ is a zero of $\partial(f+g)$. On the other hand, if $\partial f + \partial g = \partial(f+g)$, then any zero of $\partial(f+g)$ is also a zero of $\partial f + \partial g$. □

5.1 Convergence Rate Analysis of S-KM with IID Data

Although many works focus on special cases of the S-KM iteration to solve the structured optimization problems in 2.2–2.3 via specific splitting operators, analyses of KM iterations in the stochastic setting are limited even when samples are IID. For example, in the IID setting, [67] uses a primal-dual stochastic gradient method to solve convex problems with many functional constraints. [58] studies stochastic proximal gradient algorithms and [69] considers applications of these algorithms and stochastic ADMM to machining learning and deep learning. All of these algorithms can be seen as special cases of stochastic KM iteration.

It is worth emphasizing that, while the KM algorithm itself is not new, no previous work has considered the general assumption of a nonexpansive operator in the stochastic setting. Before proceeding to our main results on S-KM iteration with dependent samples, we mend this gap in the literature and establish convergence results for S-KM when samples are IID following $\mathbb{P}$. These results provide convergence guarantees for new algorithms in the stochastic setting with high-dimensional, IID data.

Theorem 8 (Pointwise Convergence of S-KM with Averaged Nonexpansive Operators) *Let* $T_k : \mathbb{X} \to \mathbb{X}$ *be an averaged nonexpansive operator. Let* $\prod_{k=1}^{K} T_k = T_1 \circ \cdots \circ T_K$ *with* $\bigcap_{k\in\mathbb{N}} \operatorname{Fix}(T_k) \neq \emptyset$*. The sequence* $(x^k)_{k\in\mathbb{N}}$ *obeys the recursion* $x^{k+1} = (1-\lambda_k)z^k + \lambda_k(\prod_{k=1}^{K} T_k x^k + \epsilon_k)$ *for some* $k \in \mathbb{N}_0$*,* $x^0 \in \mathbb{X}$ *with* $\lambda_k \in (0,1]$*. Suppose that* $\sum_{k=1}^{\infty} \lambda_k \mathbb{E}[\|\epsilon_k\|] < \infty$ *and* $\sum_{k=1}^{\infty}(1+k)\lambda_k^2 \mathbb{E}[\|\epsilon_k\|^2] < \infty$*. Define*

$$\|e_k\|^2 = \Big\| \prod_{k=1}^{K} T_k x^k - x^k \Big\|^2 = \Big\| \frac{(x^k - x^{k+1})}{\lambda_k} + \epsilon_k \Big\|^2.$$

It follows that (i) $\|x^k - x^*\|^2$ *converges almost surely with* $x^* \in \bigcap_{n\in\mathbb{K}}$ *and (ii)* $e_k \to 0$ *almost surely and* $e_k = o_p((k+1)^{-1/2})$*.*

Theorem 9 (Pointwise Convergence of S-KM with Nonexpansive Operators) *Let $T : \mathbb{X} \to \mathbb{X}$ be a nonexpansive operator, $(x^k)_{k\in\mathbb{N}_+} \subseteq \mathbb{X}$ generated following (5.1), and $(\lambda_k)_{k\in\mathbb{N}_+} \subseteq (0,1)$. Assume that $\sum_{k=1}^{\infty} \lambda_k \mathbb{E}[\|\epsilon_k\|] < \infty$, $\sum_{k=1}^{\infty}(k+1)\lambda_k^2 \mathbb{E}[\|\epsilon_k\|^2] < \infty$, and $\tau = \inf_{k\in\mathbb{N}_+} \tau_k \in (\varepsilon, \infty)$ for some $\varepsilon > 0$ with $\tau_k = \lambda_k(1-\lambda_k)$. It follows that (i) there exists $x^* \in \mathcal{X}^*$ such that $\|x^k - x^*\| \to 0$ almost surely and (ii) $\|T(x^k) - x^k\|^2 = o_p((k+1)^{-1})$.*

The property that $\|x^k - x^*\|$ converges to zero may fail, but $\|x^k - x^*\|$ may still be bounded by a finite value. We thus study $\|Tx^k - x^k\|^2$ due to the property that $\lim_{k\to\infty}\|Tx^k - x^k\| = 0$ always holds when a fixed point of T exists. We next give a convergence rate for FPR, which we immediately improve by showing that there exists $x^* \in \mathcal{X}^\star$ such that $\|x^k - x^*\| \to 0$ almost surely. Theorem 9 indicates that there exists $x^* \in \mathcal{X}^\star$ such that $\|x^k - x^*\| \to 0$ almost surely that improves the existing result.

Theorem 10 (Ergodic Convergence Rate of S-KM with a Nonexpansive Operator) *Make the same assumptions as Theorem 9. Let $r = \|x^1 - x^*\|$ and $\bar{e}_k = \Lambda_k^{-1} \sum_{l=1}^{k} \lambda_l e^l$ where $e^l = Tx^l - x^l$ and $\Lambda_k = \sum_{l=1}^{k} \lambda_l$. Then $\|\bar{e}_k\| = O_p([(k+1)\lambda]^{-1})$ and $\mathbb{E}[\|\bar{e}^k\|] \leq 2\Lambda_k^{-1}(r + \sum_{k=1}^{k} \mathbb{E}[\|\lambda_k \epsilon_k\|])$.*

Theorems 8–9 establish a strong convergence guarantee for the S-KM iteration. However, convergence relies heavily on the IID assumption. It is natural to next examine these properties again with non-IID data, particularly, where samples are generated from an ergodic stochastic process. In the next few sections, we provide our main tractability results and evaluate the performance of the S-KM iteration with dependent data. Under some mild conditions, S-KM iterates concentrate around the true value. These general results are fundamental and cover many splitting algorithms as special cases.

5.2 Convergence Rate Analysis of S-KM with Non-IID Data

Stability of Fixed Point Iteration with Nonexpansive Operators

The stability and boundness of S-KM iterates are the main focus of this section. We first formalize our notion of stability.

Definition 7 *A sequence of iterates $(x^k)_{k\in\mathbb{N}}$ generated according to Algorithm 5 is stable in probability if, for all $x \in \mathbb{X}$, $\sup_{k\in\mathbb{N}} \operatorname{dist}(x^k, x^{k+1}) < \infty$ with probability one.*

Remark 3 *Definition 7 is similar to the requirement introduced in [2], which aims to find stability guarantees for stochastic proximal point methods, that $\sup_{k\in\mathbb{N}} \operatorname{dist}(x^k, \mathcal{X}^*) < \infty$. Similar results have been established for stochastic subgradient methods and require that $\sup_{k\in\mathbb{N}} \operatorname{dist}(x^k, \mathcal{X}^*) < \infty$ whenever $\mathbb{E}[\|\ell'(x,\xi)\|_2^2] \leq C_0 + C_1 \operatorname{dist}(x, \mathcal{X}^\star)^2$ for all $x \in \mathbb{X}$. Definition 7 differs in that, instead of considering the distance between iterates an optimal solution, which can be bounded above if the solution is nonunique, defines stability in terms of the residual.*

The following assumption is a sufficient condition for the stability (in the sense of Definition 7) of the proposed S-KM iteration. The subsequent theorem gives sufficient conditions for the stability of the proposed stochastic operator-splitting schemes.

Assumption 9 (Boundedness) *For any $x, x^* \in \mathbb{X}, \|x - x^*\| \leq r < \infty$. That is, $\mathbb{X}$ is compact and has a finite radius r.*

Theorem 11 (FPR Bound) *Under Assumption 9 and the mixing conditions in Definition 3.1, let $\bar{e}_K = \Lambda_K^{-1} \sum_{k=1}^K \lambda_k e_k$ with $e_k = Tx^k - x^k$, $\Lambda_K = \sum_{k=1}^K \lambda_K$, and $(\lambda_k)_{k\in\mathbb{N}} \subseteq (0,1)$. We have that*

$$\mathbb{E}[\|\bar{e}_K\|] \leq \frac{2(r + 2\sum_{k=1}^K \mathbb{E}[\|\lambda_k \epsilon_k\|]) + 2\phi(1)}{\Lambda_K}. \tag{5.3}$$

with the assumption that for all approximation error $\{\epsilon_k\}_{k\in\mathbb{N}}$ with $\sum_{k=1}^K \|\epsilon_k\| < \infty$, $\bar{e}_K$ converge almost surely in expectation with convergence rate $O(\Lambda_k^{-1})$.

Remark 4 *A comparison of Theorems 5.3 and 10 suggests that dependence among training samples adds a penalty of $\phi(1)$ to the upper FPR bound. The convergence rate given in Theorem 11 is the optimal convergence rate and was established for exact KM iterations [20].*

A few more-consequential corollaries regarding stability follow.

Corollary 2 *Under the assumption that the error term satisfies $\|\epsilon\| < C$ for some $C \leq \infty$ and the conditions of Theorem 11, $\sup_{k\in\mathbb{N}} \operatorname{dist}(x^k, x^{k+1}) < \infty$ and $\operatorname{dist}(x^k, x^{k+1})$ converges to some finite value with probability one at the rate $O(1/\Lambda_k)$.*

By combining Theorem 11 and Corollary 2, we see that the stochastic operator-splitting schemes are stable according to Definition 7. This is in strong contrast to pointwise stochastic algorithms, which can be unstable even in relatively simple problems.

We next show that the approximation error arising in the iteration is bounded and vanishes as $K \to \infty$. The FPR bounds depend on the (possibly stochastic) error sequence ϵ_k. Here, we develop bounds for the expected FPR in terms of the cumulative error $\mathbb{E}[{\Lambda_K}^{-1} \sum_{k=1}^K \lambda_k e_k]$, where $\Lambda_K = \sum_{k=1}^K \lambda_k$. In practical settings, the error process may stem from various sources such as random noise. Here, we attribute the error to random sampling from $\mathbb{P}^k$ rather than from $\mathbb{P}$. It is therefore required that both path variations and errors diminish with k. The path length may diminish if the target being tracked slows down over time or eventually stops. Likewise, ϵ may diminish if the noisy gradients available from the adversary can be corrected or improved with time. This assumption is commonly employed in convergence analyses, though of course, there is a possibility that the assumption of vanishing errors is not realistic. This motivates us to investigate this property of the approximation error to ensure that this convergence is valid.

Theorem 12 (Approximation Error Bound) *Under Assumption 9, the norm of the difference between the true function $T(x^k; \xi) - x^k$, where ξ is drawn from $\mathbb{P}$, and its approximation $T(x^k; \xi_{k+1}) - x^k$, where ξ_{k+1} is drawn from $\widehat{\mathbb{P}}_K$, is uniformly bounded in expectation. Specifically, $\mathbb{E}[\|\epsilon_k\|] \leq \Delta$, where*

$$\Delta = \Big(\frac{8r^2 C(\sum_{k=1}^\infty \phi(v))}{K^2}\Big)^{1/2} \Gamma\Big(\frac{1}{2}\Big)$$

and $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) \mathrm{d}x$ is the gamma function.

Theorem 12 provides a strong theoretical justification for the boundedness of the approximation error, which suggests that the approximation becomes close to the truth as K increases. However, the amount of noise can be large when K is small. This is because we are considering drifting distributions and, in the worse case, P^k can be quite far away from $\mathbb{P}$. Theorem 12 indicates that underestimating the mixing time can potentially backfire.

Generalization Bounds for Stochastic Operator-splitting Algorithms with Non-IID Data

We proceed with a theorem that provides a high-probability generalization guarantee for stochastic operator-splitting algorithms with a nonexpansive operator.

Assumption 10 (Iteration Boundedness) *There is a nonincreasing sequence $\kappa(k)$ such that, for successive S-KM iterates x^k and x^{k+1}, $\mathbb{E}[\|x^{k+1} - x^k\| \mid \mathcal{F}_k] \leq \kappa(k)$.*

Assumption 11 (Iterate Boundedness) *For a sequence of samples $\xi_1, \ldots, \xi_K$, the S-KM iteration produces a sequence of iterates $x^1, \ldots, x^{T-1}$ such that $\sum_{k=0}^{K-1} \|T_{\lambda_k}(x^k; \xi_{k+1}) - T_{\lambda_k}(x^*; \xi_{k+1})\| \leq R_{K-1}$.*

Assumptions 10–11 stipulate that the iteration trajectory is appropriately stable. As established previously, iterates from the fixed-point iteration of a nonexpansive operator, through averaging, are stable. These mild assumptions do not sacrifice much power. An averaged operator has these contraction properties, which suggests that these assumptions are reasonable: for example, SGD satisfies $R_K = O(\sqrt{K})$ and $\kappa(k) = O(1/\sqrt{k})$. We now establish an upper bound for S-KM iterates around the true value.

Theorem 13 (Iterate Deviation Bound) *Under Assumptions 9–11, for any $\tau > 0$,*

$$\mathbb{E}\Big[\Big\| \sum_{k=1}^{K} (x^k - x^*) \Big\|\Big] \leq \underbrace{\mathbb{E}[R_{K-1}]}_{\textit{average regret}} + \underbrace{\tau\Big(\sum_{k=1}^{K-\tau} \mathbb{E}[\kappa(k-1)] + r \Big)}_{\textit{deviation bound}} + \underbrace{\phi(1)\mathbb{E}[R_{K-1}] + 2\sqrt{2}(K-\tau) r \sqrt{\phi(\tau+1)}}_{\textit{non-IID penalty}}.$$

The proof of Theorem 13 (given in the Appendix) requires that we understand the impact of the ergodic sequence $\xi_1, \xi_2, \ldots, \xi_K$ on the data-driven procedure. The performance of iterates under the operator T is bounded by the sum of three main quantities: the average regret of the algorithm; the deviation bound, which in turn depends on how far P^k deviates from $\mathbb{P}$; and the mixing time of the data source. We observe that setting $\phi(1) = 0$ and $\tau = 0$ recovers an expected version of the results for IID samples. It is clear that the stability assumption plays a key role in our result whenever $\tau > 0$, that is when the samples are dependent. Ultimately, for large enough τ, the sample $\xi_{k+\tau}$ is "nearly independent" of the parameters x^k.

Chapter 6

Convergence Analysis of Several Splitting Algorithms

Many works focus on using stochastic operator-splitting algorithms to solve structured optimization problems [52], [58], [67], [69]. We next present several examples of nonexpansive operators that cover numerous algorithms based on the proximal and gradient operators. We use our results to analyze several special cases: FBS, DRS, PRS, and the ADMM algorithm.

6.1 Stochastic Generalized FBS

Consider the monotone inclusion problem of finding $x \in \mathbb{X}$ such that $0 \in \partial f(x) + \partial g(x)$ where $\partial g, \partial f : \mathbb{X} \to \mathbb{R}^d$ are assumed to be β-cocoercive for some $\beta > 0$. This problem corresponds to the problem in 2.2, where $f : \mathbb{X} \to \mathbb{R}$ is a convex, differentiable function with a β^{-1}-Lipschitz-continuous gradient and $g : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ is a proper, closed, lower semicontinuous, convex function.

Corollary 3 *Suppose that $\gamma_k \in (0, 2\beta)$ and $\lambda_k \in (0, (4\beta - \gamma_k)/(2\beta)]$ for $k \in \mathbb{N}$. Let $\partial g, \partial f : \mathbb{X} \to \mathbb{R}^d$ be β-cocoercive for some $\beta > 0$ and let $(\epsilon_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^d$ be a sequence of random variables adapted to $\mathcal{F}_k = \sigma(\epsilon_k, l \leq k)$. Define $T_{FB}(x) = \mathcal{J}_{\gamma_k \partial g}(I - \gamma_k \partial f)(x)$. Then $\operatorname{zer}(\partial f + \partial g) = \operatorname{Fix}(T_{\mathrm{FB}})$ and the sequence of iterates $(x^k)_{k \in \mathbb{N}}$ generated by the algorithm satisfies*

$$\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}(x^k - x^\star)\Big\|\Big] \leq \mathbb{E}[R_{K-1}] + \tau\Big(\sum_{k=1}^{K-\tau} \mathbb{E}[\kappa(k-1)] + r\Big) + \phi(1)\mathbb{E}[R_{K-1}] + 2\sqrt{2}(K-\tau)r\sqrt{\phi(\tau+1)}.$$

Additionally, let $\bar{e}_K = \Lambda_K^{-1} \sum_{k=1}^K \lambda_k e_k$ with $e_k = T_{FB}(x^k) - x^k, \Lambda_K = \sum_{k=1}^K \lambda_k$, and $(\lambda_k)_{k \in \mathbb{N}} \in (0,1)$. Then

$$\mathbb{E}[\|\bar{e}_K\|] \leq \frac{2(r + 2\sum_{k=1}^K \mathbb{E}[\|\lambda_k \epsilon_k\|]) + 2\phi(1)}{\Lambda_K}.$$

In particular, if $\sum_{k=0}^\infty \mathbb{E}[\lambda_k \|\epsilon_{f,k}\|] < \infty$ and $\sum_{k=0}^\infty \mathbb{E}[\lambda_k \|\epsilon_{g,k}\|] < \infty$, then $\bar{e}_K = O_p(\Lambda_K^{-1})$.

Remark 5 *If $\lambda_k = 1$, then S-GFBS reduces to S-FBS, which is also known as the stochastic proximal gradient method. It has the subgradient representation*

$$x^{k+1} = \mathcal{J}_{\gamma_k \partial g}(x^k - \gamma_k \nabla f(x^k) + \epsilon_{f,k}) + \epsilon_{g,k} = x^k - \gamma_k \tilde{\nabla} g(x^{k+1}) - \gamma_k \nabla f(x^k) + \epsilon_k,$$

where $\tilde{\nabla} g(x^{k+1}) = \gamma_k^{-1}(x^k - x^{k+1} - \gamma_k \nabla f(x^k)) \in \partial g(x^{k+1})$, and x^{k+1} and $\tilde{\nabla} g(x^{k+1})$ are unique given x^k and $\gamma_k > 0$. The update $x^{k+1} = x^k + \lambda_k(T_{FB}(x^k) - x^k)$ can be viewed as adding momentum based on x^k, which is also known as the inertial proximal gradient method [22].

The following result is used in the proof of Theorem 14 to establish a generalized error bound for regret in S-GFBS.

Theorem 14 (Generalized Error Bound for Regret) *Let $\bar{x} = K^{-1}\sum_{k=1}^{K} x^k$ and $\underline{\tau} = \inf_k \tau_k$, where $\tau_k = \lambda_k(1-\lambda_k)$ for $k \in \mathbb{N}$. Under Assumption 9,*

$$\mathbb{E}[f(\bar{x}) + g(\bar{x}) - (f(x^\star) + g(x^\star))] \le \frac{r^2}{2K\gamma} + \Big(\frac{1}{\beta} - \frac{1}{\gamma}\Big)\Big(4r^2 + \frac{8r^2\pi C(\sum_{k=1}^{\infty}\phi(k))}{K\underline{\tau}}\Big).$$

6.2 S-rPRS and S-DRS

A direct application of the proximal point algorithm to the problem of minimizing $f+g$ would require computing the proximal operator $\operatorname{prox}_{\gamma_k(f+g)}$, which can be difficult to evaluate. The DRS algorithm eliminates this difficulty by separately evaluating the proximal operators of f and g. We first establish the equivalence of this problem to that in 2.2.

Let $\partial f, \partial g : \mathbb{X} \to \mathbb{R}^d$ be two maximal monotone operators. Then $x^\star$ is an optimal solution to the problem in 2.2 if and only if, for any $\gamma_k > 0$ and $\lambda_k \in \mathbb{R}$, $x^\star = \mathcal{J}_{\gamma_k \partial f}(x^\star)$ and

$$x^\star = x^\star + \lambda_k[\mathcal{J}_{\gamma_k \partial g}(2\mathcal{J}_{\gamma_k \partial f}(x^\star) - x^\star) - \mathcal{J}_{\gamma_k \partial f}(x^\star)].$$

Specifically, when $\lambda = 1$,

$$x^\star = 2^{-1}[x^\star + \mathcal{J}_{\gamma_k \partial g} \circ \mathcal{J}_{\gamma_k \partial f}(x^\star)] = T_{\gamma_k, \partial f, \partial g}(x^\star),$$

where the operator $T_{\mathrm{DR}} = 2^{-1}(\operatorname{refl}_{\gamma_k \partial f} \operatorname{refl}_{\gamma_k \partial g} + I)$ is known as the Douglas–Rachford operator. When $\lambda = 2$,

$$x^\star = x^\star + 2\mathcal{J}_{\gamma_k \partial g}(2\mathcal{J}_{\gamma_k \partial f}(x^\star) - x^\star) - (2\mathcal{J}_{\gamma_k \partial f}(x^\star) - x^\star) = x^\star + \mathcal{J}_{\gamma_k \partial g}\mathcal{J}_{\gamma_k \partial f}(x^\star) = T_{\lambda_k, \partial f, \partial g}(x^\star),$$

where the operator $T_{\mathrm{PR}} = \operatorname{refl}_{\lambda_k g} \circ \operatorname{refl}_{\lambda_k f}$ is known as the Peaceman–Rachford operator.

We now establish the asymptotic behavior of the D-DRS algorithm given in Algorithm **??**.

Corollary 4 *Let $\partial f, \partial g : \mathbb{X} \to \mathbb{R}^d$ be two maximal monotone operators. Suppose that $\gamma_k > 0$, $(\lambda_k)_{k\in\mathbb{N}} \subseteq (0, 2)$, and that $(\epsilon_k)_{k\in\mathbb{N}} \subseteq \mathbb{X}$ is a sequence of random variables that are adapted to $\mathcal{F}_k = \sigma(\epsilon_k, l \le k)$. Given any arbitrary $x^0 \in \mathbb{X}$, the iterates generated by the algorithm with $1/2$ can be written as the fixed-point iteration of x^k and the operator T_{DR} as*

$$\begin{aligned} x^{k+1} &= x^k + \lambda_k\Big(\frac{1}{2}(\operatorname{refl}_{\gamma_k \partial f} \operatorname{refl}_{\gamma_k \partial g} + I)x^k - x^k + \epsilon_k\Big) \\ &= x^k + \lambda_k(\mathcal{J}_{\gamma_k \partial f}(2(\mathcal{J}_{\gamma_k \partial g}x^k + \epsilon_{g,k}) - x^k) + \epsilon_{f,k} - (\mathcal{J}_{\gamma_k \partial g}x^k + \epsilon_{g,k})). \end{aligned}$$

The sequence of iterates $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm **??** *satisfies*

$$\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}(x^k - x^*)\Big\|\Big] \leq \mathbb{E}[R_{K-1}] + \tau\Big(\sum_{k=1}^{K-\tau}\mathbb{E}[\kappa(k-1)] + r\Big) + \phi(1)\mathbb{E}[R_{K-1}] + 2\sqrt{2}(K-\tau)r\sqrt{\phi(\tau+1)}.$$

Additionally, let $\bar{e}_K = \Lambda_K^{-1}\sum_{k=1}^{K}\lambda_k e_k$ with $e_k = T_{DR}x^k - x^k, \Lambda_K = \sum_{k=1}^{K}\lambda_k$, and $(\lambda_k)_{k\in\mathbb{N}} \subseteq (0,1)$. Then

$$\mathbb{E}[\|\bar{e}_K\|] \leq \frac{2(r + 2\sum_{k=1}^{K}\mathbb{E}[\|\lambda_k\epsilon_k\|]) + 2\phi(1)}{\Lambda_K}.$$

In particular, if $\sum_{k=0}^{\infty}\mathbb{E}[\lambda_k\|\epsilon_{f,k}\|] < \infty$ and $\sum_{k=0}^{\infty}\mathbb{E}[\lambda_k\|\epsilon_{g,k}\|] < \infty$, then $\bar{e}_K = O_p(\Lambda_K^{-1})$.

Remark 6 *Both DRS and PRS are special cases of relaxed PRS, which follows the iteration*

$$x^{k+1} = (1-\lambda_k)x^k + \lambda_k \operatorname{refl}_{\lambda f} \circ \operatorname{refl}_{\lambda g}(x^k).$$

Taking $\lambda_k = 1$ yields PRS and $\lambda_k = 1/2$ yields DRS.

We now turn to the analysis of the S-rPRS algorithm (given in Algorithm **??**), where we establish a generalized error bound for the regret.

Corollary 5 *Define $T_{PRS} = \operatorname{refl}_{\lambda f} \circ \operatorname{refl}_{\lambda g}$. Let $x^\star$ be a fixed point of T_{PRS} and $(\epsilon_k)_{k\in\mathbb{N}} \subseteq \mathcal{X}$ a sequence of random variables adapted to $\mathcal{F}_k = \sigma(\epsilon_k, l \leq k)$. Additionally, let $(\gamma_k)_{k\in\mathbb{N}} \subseteq (0,\infty)$, and $x^0 \in \mathbb{X}$. If $(\lambda_k)_{k\in\mathbb{N}} \subseteq (0,2)$ then for any $\tau > 0$, the sequence of iterates $(x^k)_{k\in\mathbb{N}}$ generated by Algorithm* **??** *satisfies*

$$\mathbb{E}\Big[\Big\|\sum_{k=1}^{K}(x^k - x^\star)\Big\|\Big] \leq \mathbb{E}[R_{K-1}] + \tau\Big(\sum_{k=1}^{K-\tau}\mathbb{E}[\kappa(k-1)] + r\Big) + \phi(1)\mathbb{E}[R_{K-1}] + 2\sqrt{2}(K-\tau)r\sqrt{\phi(\tau+1)}.$$

Let $\bar{e}_K = \Lambda_K^{-1}\sum_{k=1}^{K}\lambda_k e_k$, with $e_k = T_{PRS}x^k - x^k$ and $\Lambda_K = \sum_{k=1}^{K}\lambda_k$. If, additionally, $\lambda_k \in (0,1)$, then

$$\mathbb{E}[\|\bar{e}_K\|] \leq \frac{2(r + 2\sum_{k=1}^{K}\mathbb{E}[\|\lambda_k\epsilon_k\|]) + 2\phi(1)}{\Lambda_K}.$$

In particular, if $\sum_{k=0}^{\infty}\mathbb{E}[\lambda_k\|\epsilon_{f,k}\|] < \infty$ and $\sum_{k=0}^{\infty}\mathbb{E}[\lambda_k\|\epsilon_{g,k}\|] < \infty$, then $\bar{e}_K = O(\Lambda_K^{-1})$.

Theorem 15 (Generalized Error Bound for Regret) *Under Assumption 9, let $\bar{x}^f = K^{-1}\sum_{k=1}^{K}x_k^f$ and $\bar{x}^g = K^{-1}\sum_{k=1}^{K}x_k^g$, where $x_k^g = \operatorname{prox}_{\gamma,g}(x^k) + \epsilon_g$ and $x_k^f = \operatorname{prox}_{\gamma,f}(\operatorname{refl}_{\gamma\partial g}(x^k)) + \epsilon_f$. Define $\underline{\tau} = \inf_k \tau_k$, with $\tau_k = \lambda_k(1-\lambda_k)$. It follows that*

$$\mathbb{E}[f(\bar{x}^f;\xi) + g(\bar{x}^g;\xi) - (f(x^\star;\xi) + g(x^\star;\xi))] \leq \frac{r^2}{4\gamma\lambda K} + \frac{2(\lambda-1)r^2}{\gamma\lambda^2} + \Big(1 - \frac{1}{\lambda}\Big)\frac{4Cr^2\pi\sum_{k=1}^{\infty}\phi(k)}{\gamma\lambda\underline{\tau}K}.$$

Remark 7 *Suppose that the function f in problem 2.2 is differentiable and that ∇f is $(1/\beta)$-Lipschitz. Under this smoothness assumption, we can apply the S-FBS algorithm to the problem: given $x^0 \in \mathbb{X}$, for all $k \geq 0$, define $x^{k+1} = \operatorname{prox}_{\gamma_k g}(x^k - \gamma_k\nabla f(x^k))$. To ensure convergence, the step size parameter γ_k must be strictly less than 2β. Because the gradient operator is typically easier to evaluate than the proximal operator, it may be preferable to use S-FBS instead of S-rPRS whenever one of the terms in the objective function is differentiable. However, the following empirical and theoretical points should also be considered to determine whether S-rPRS is preferable over S-FBS.*

1. *The gradient and proximal point operators are involved in S-FBS, while S-rPRS uses only the proximal operator. In both theory and practice, proximal point methods generally converge faster than gradient descent methods.*

2. *Gradient descent methods can generate out-of-domain iterates, while proximal methods can ensure iterate feasibility.*

3. *If the Lipschitz constant of ∇f is not known, then a line search procedure can find an appropriate step size to ensure convergence. However, this presents another practical challenge: if this procedure is more expensive than evaluating the proximal operator, then S-rPRS should be used. Even if the Lipschitz constant of ∇f is known, S-rPRS is known to converge faster in practice than S-FBS, which indicates that we can do no worse by using S-rPRS.*

4. *As shown in Theorem 15, S-rPRS iterates converge regardless of the chosen step size, while S-rFBS may fail to converge. S-rPRS thus "demistifies" parameter selection, which may partially explain the perceived practical advantages of relaxed PRS over FBS.*

6.3 Numerical Experiments

We conduct numerical experiments based on two applications, a Lasso-type quadratic minimization problem with data generated from a vector-auto regressive process and a robust regression problem via training samples coming from a peer-to-peer (P2P) network. The goal of experiments is to verify the theoretical results for SAA with dependent data.

6.3.1 Lasso-type quadratic minimization problem

For some insight into the introduce a data-driven procedure, we present some numerical results for a simple lasso-type quadratic minimization problem where samples are generated from a stationary vector-auto regressive process. We wish to compute

$$\widehat{x}_K^\star = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \left\{ \frac{1}{K}\sum_{k=1}^{K} \|\langle x, \xi_k^1\rangle - \xi_k^2\|_2^2 + \lambda\|x\|_1 \right\},$$

where λ is a preset tuning parameter. Our data-generating mechanism resembles that in [24]. Let A be a subdiagonal matrix with entries $A_{i,i-1} \overset{\text{IID}}{\sim} \text{U}[0.8, 0.99]$. We draw a sparse vector $x \in \mathbb{R}^{1000}$ with all but its first 50 elements of x equal to zero. The data $\{(\xi_k^1, \xi_k^2)\}_{k\in\mathbb{N}}$ is generated according to the vector-auto regressive process, $\xi_k^1 = A\xi_{k-1}^1 + e_1 W_k$, $\xi_k^2 = \langle x, \xi_k^1\rangle + E_k$, where $W_k \overset{\text{IID}}{\sim} N(0,1)$ and the E_ks are IID bi-exponential random variables, each with a common variance of one.

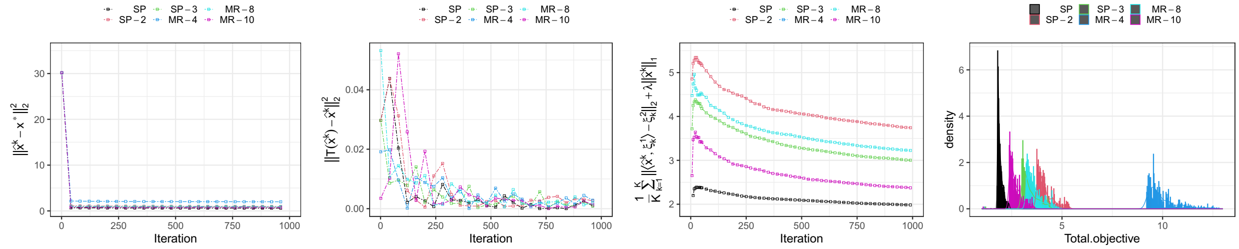


Figure 6.1: Performance of the SP, SP-m, and MR-s methods for various values of m and s as measured by (left) regret, (middle) FPR, and (right) distance between the estimated and true value of x in the lasso-type example.

We use stochastic FBS to compute the objective function. Our goal is to compare the proposed procedure (in two flavors) and MR in terms of performance and the number of samples used. Specifically, we generate samples as (SP) every element of a single trajectory, (SP-m) every mth element of a single trajectory, and (MR-s) the sth element of independent trajectories starting from the same state. In all three methods, K samples are generated from one (in the case of SP and SP-m) or K (for MR-s) trajectories. Intuitively, we expect SP-m to weaken the dependence between samples by not using every sample. SP and MR closely resemble the sampling techniques in [24] and [63]. For a fixed sample size $K = 1000$, we consider $m = 2, 3$ for SP-m and $s = 4, 6, 8, 10$ for MR-s.

Figure 6.1 illustrates our numerical results and the convergence of the three methods as evaluated by three criteria: regret $(L(\bar{x}) - L(x^*))$, FPR, and the difference between the current iterate and the true value of x. As expected, MR shows poor performance when s is small as the true mixing time is underestimated.

In particular, MR-4 has the worst performance. It is clear that using every sample (in SP) is more computationally efficient than taking every mth sample (in SP-m) in an attempt to weaken serial dependence between sequential samples. While the latter approach is indeed able to weaken this dependence, it leads to no notable improvements in performance.

6.3.2 Robust regression problem via P2P network

We consider the robust regression problem with the training samples ξ coming from P2P network

$$\min_x F(x) = \mathbb{E}_{\xi=(a,b)} \left|\eta^\top x - b\right| ,$$

where $\xi = (a, b)$, $a \in \mathbb{R}^d$ is a random feature vector, and $b \in \mathbb{R}$ is the response. $\eta = \mathcal{N}\left(a, \sigma_\eta^2 I_d\right)$ is a perturbed noisy observation of the input feature vector a. Let $d = 10$, and the data is generated as follows, with 10 servers (each with 20000 samples stored), $a_i \sim \mathcal{N}\left(\mu_i, \sigma_\xi^2 I_d\right)$, $b_i = a_i^T x^\star$ with $\sigma_\xi^2 = 1$ and pre-specified $\mu_i, x^\star$. Here, μ_i is allowed to be different among servers but the same for σ_ξ^2. For a given sample budge T, ranging from 10^3 to 10^6, we adopt different sample allocation strategies for $N = \{O(T^{\frac{1}{2}}), O(T^{\frac{1}{3}}), O(T^{\frac{1}{4}})\}$, with $M = O(T^{\frac{1}{2}})$ and repeat 30 runs for each sample allocation to report the average performance. Performance measures include bias, in-sample risk against the true values $x^\star$ and $F(x^\star)$, and probability guarantee defined as $\mathbb{P}(\widehat{x}_{N,M}^\star \in \mathbb{B}_{\varepsilon=0.5}(x^\star))$. In addition, we also consider the special case with independent inner and outer randomness to compare the performance of the two sampling schemes by choosing different inner samples. In independent sampling scheme, $\eta_{ij} = \eta_{1j}$ for all $i > 1$. The results are shown in Figure 6.2.

Figure 6.2 shows that the setting with $N = O(T^{\frac{1}{2}})$ has the best performance, smallest bias and highest probability guarantee, for robust regression, which is consistent with our sample complexity results. We can see from Figure 6.2 (b) that the probabilistic guarantee increases exponentially as the sample size increases. Figure 6.2 (c) visualizes the performance of probability guarantee and in-sample risk simultaneously. For two sampling schemes, Figure 6.2 (d) indicates that the independent sampling scheme has a smaller in-sample risk, and the gap gradually decreases as sample size increases, which is also consistent with our theoretical analysis.

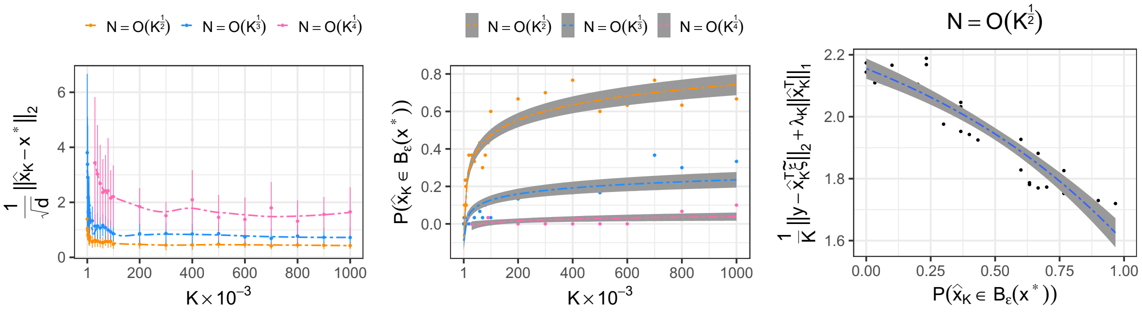


Figure 6.2: Robust regression problem via P2P network

Chapter 7

Conclusion

In this paper, we demonstrate that SAA retains its asymptotic properties in settings with dependent samples and can be implemented efficiently via stochastic operator-splitting schemes for numerous important loss functions. We propose an efficient, data-driven procedure for constructing a sequence of discrete empirical distributions that converges to the true underlying distribution. Our analysis, which includes a derivation of the sample complexity and MSE of SAA, shows that many stochastic optimization frameworks can be approximated by SAA and solved efficiently. We also investigate the out-of-sample performance of the resulting optimal decisions, both theoretically and experimentally, and analyze its advantage over a few common types of data-generating processes. In the future, we will study the general performance of SAA under the framework of conditional stochastic optimization and the application of SAA to dependent data.

References

[1] R. Alfred, J. H. Obit, M. H. Ahmad Hijazi, A. A. Ag Ibrahim, *et al.*, "A performance comparison of statistical and machine learning techniques in learning time series data," *Advanced Science Letters*, vol. 21, no. 10, pp. 3037–3041, 2015.

[2] H. Asi and J. C. Duchi, "Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity," *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 2257–2290, 2019.

[3] L. C. Baird III and A. W. Moore, "Gradient descent for general reinforcement learning," in *Advances in neural information processing systems*, 1999, pp. 968–974.

[4] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Machine Learning*, vol. 48, no. 1, pp. 85–113, 2002.

[5] H. H. Bauschke, R. S. Burachik, and D. R. Luke, *Splitting Algorithms, Modern Operator Theory, and Applications.* Springer, 2019.

[6] H. H. Bauschke, P. L. Combettes, *et al.*, *Convex analysis and monotone operator theory in Hilbert spaces.* Springer, vol. 408.

[7] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, vol. 2010, no. 1-38, p. 3, 2011.

[8] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Mathematical Programming*, vol. 171, no. 1, pp. 217–282, 2018.

[9] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186.

[10] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

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

[12] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers.* Now Publishers Inc, 2011.

[13] C. Brownlees, E. Joly, and G. Lugosi, "Empirical risk minimization for heavy-tailed losses," *The Annals of Statistics*, vol. 43, no. 6, pp. 2507–2536, 2015.

[14] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," in *Annales de l'IHP Probabilités et statistiques*, vol. 48, 2012, pp. 1148–1185.

[15] S. Clémençon, A. Bellet, O. Jelassi, and G. Papa, "Scalability of stochastic gradient descent based on" smart" sampling techniques.," 2015.

[16] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, Springer, 2011, pp. 185–212.

[17] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[18] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local rademacher complexity," *Advances in neural information processing systems*, vol. 26, 2013.

[19] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, no. 1-4, pp. 191–213, 1972.

[20] D. Davis and W. Yin, "Convergence rate analysis of several splitting schemes," in *Splitting methods in communication, imaging, science, and engineering*, Springer, 2016, pp. 115–163.

[21] J. Dedecker and F. Merlevede, "The empirical distribution function for dependent variables: Asymptotic and nonasymptotic results in," *ESAIM: Probability and Statistics*, vol. 11, pp. 102–114, 2007.

[22] P. Duan, Y. Zhang, and Q. Bu, "New inertial proximal gradient methods for unconstrained convex optimization problems," *Journal of Inequalities and Applications*, vol. 2020, no. 1, pp. 1–18, 2020.

[23] Y. Duan, C. Jin, and Z. Li, "Risk bounds and rademacher complexity in batch reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2021, pp. 2892–2902.

[24] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic mirror descent," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578, 2012.

[25] S. Eickmeier, W. Lemke, and M. Marcellino, "Classical time varying factor-augmented vector autoregressive models—estimation, forecasting and structural analysis," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 178, no. 3, pp. 493–533, 2015.

[26] A. Emelogu, S. Chowdhury, M. Marufuzzaman, L. Bian, and B. Eksioglu, "An enhanced sample average approximation method for stochastic optimization," *International Journal of Production Economics*, vol. 182, pp. 230–252, 2016.

[27] Y. M. Ermoliev and R.-B. Wets, *Numerical techniques for stochastic optimization.* Springer-Verlag, 1988.

[28] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[29] R. Fioresi, P. Chaudhari, and S. Soatto, "A geometric interpretation of stochastic gradient descent using diffusion metrics," *Entropy*, vol. 22, no. 1, p. 101, 2020.

[30] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical science*, vol. 7, no. 4, pp. 457–472, 1992.

[31] L. Gerencsér and Z. Vágó, "Adaptive control of multivariable linear stochastic systems. a strong approximation approach," in *1999 European Control Conference (ECC)*, IEEE, 1999, pp. 1643–1647.

[32] R. Glowinski, S. J. Osher, and W. Yin, *Splitting methods in communication, imaging, science, and engineering.* Springer, 2017.

[33] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*, PMLR, 2016, pp. 1225–1234.

[34] D. P. Heyman and M. J. Sobel, *Stochastic models in operations research: stochastic optimization.* Courier Corporation, 2004, vol. 2.

[35] M. Jerrum and A. Sinclair, "The markov chain monte carlo method: An approach to approximate counting and integration," *Approximation Algorithms for NP-hard problems, PWS Publishing*, 1996.

[36] B. Johansson, M. Rabi, and M. Johansson, "A simple peer-to-peer algorithm for distributed optimization in sensor networks," in *2007 46th IEEE Conference on Decision and Control*, IEEE, 2007, pp. 4705–4710.

[37] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2010.

[38] W. Jongeneel, T. Sutter, and D. Kuhn, "Efficient learning of a linear dynamical system with stability guarantees," *IEEE Transactions on Automatic Control*, 2022.

[39] R. K. Kennedy, T. M. Khoshgoftaar, F. Villanustre, and T. Humphrey, "A parallel and distributed stochastic gradient descent implementation using commodity clusters," *Journal of Big Data*, vol. 6, no. 1, p. 16, 2019.

[40] S. Kim, R. Pasupathy, and S. G. Henderson, "A guide to sample average approximation," *Handbook of Simulation Optimization*, pp. 207–243, 2015.

[41] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

[42] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[43] M. A. Krasnosel'skii, "Two comments on the method of successive approximations," *Usp. Math. Nauk*, vol. 10, pp. 123–127, 1955.

[44] P. Kumar and P. Varaiya, "Stochastic systems: Estimation, identification and adaptive control (prentice-hall information & system sciences series)," 1986.

[45] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications.* Springer Science & Business Media, 2003, vol. 35.

[46] J. F. MacGregor and T. Kourti, "Statistical process control of multivariate processes," *Control engineering practice*, vol. 3, no. 3, pp. 403–414, 1995.

[47] G. Mamakoukas, O. Xherija, and T. Murphey, "Memory-efficient learning of stable linear dynamical systems for prediction and control," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 527–13 538, 2020.

[48] W. R. Mann, "Mean value methods in iteration," *Proceedings of the American Mathematical Society*, vol. 4, no. 3, pp. 506–510, 1953.

[49] A. Mokkadem, "Mixing properties of arma processes," *Stochastic processes and their applications*, vol. 29, no. 2, pp. 309–315, 1988.

[50] R. R. Montenegro and P. Tetali, *Mathematical aspects of mixing times in Markov chains.* Now Publishers Inc, 2006.

[51] T. Nguyen-Tang, S. Gupta, H. Tran-The, and S. Venkatesh, "Sample complexity of offline reinforcement learning with deep relu networks," *arXiv preprint arXiv:2103.06671*, 2021.

[52] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, 2013, pp. 80–88.

[53] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[54] D. Paulin, "Concentration inequalities for markov chains by marton couplings and spectral methods," *Electronic Journal of Probability*, vol. 20, pp. 1–32, 2015.

[55] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

[56] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[57] R. T. Rockafellar, *Convex analysis.* Princeton university press, 1970, vol. 36.

[58] L. Rosasco, S. Villa, and B. C. Vũ, "Convergence of stochastic proximal gradient algorithm," *Applied Mathematics & Optimization*, pp. 1–27, 2019.

[59] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[60] J. Schneider and S. Kirkpatrick, *Stochastic optimization.* Springer Science & Business Media, 2007.

[61] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory.* SIAM, 2021.

[62] J. C. Spall, "Stochastic optimization," in *Handbook of computational statistics*, Springer, 2012, pp. 173–201.

[63] T. Sun, Y. Sun, and W. Yin, "On markov chain gradient descent," *arXiv preprint arXiv:1809.04216*, 2018.

[64] A. Themelis and P. Patrinos, "Douglas–rachford splitting and admm for nonconvex optimization: Tight convergence results," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 149–181, 2020.

[65] Y. Wang, B. Pan, W. Tu, *et al.*, "Sample average approximation for stochastic optimization with dependent data: Performance guarantees and tractability," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 3859–3867.

[66] 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.

[67] Y. Xu, "Primal-dual stochastic gradient method for convex programs with many functional constraints," *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1664–1692, 2020.

[68] Y. Yazici, "Operator splitting methods for differential equations," *Izmir Institute of Technology*, 2010.

[69] J. Yun, A. C. Lozano, and E. Yang, "A general family of stochastic proximal gradient methods for deep learning," *arXiv preprint arXiv:2007.07484*, 2020.

[70] J. Zhang, "Gradient descent based optimization algorithms for deep learning models training," *arXiv preprint arXiv:1903.03614*, 2019.

Appendix A

Supplemental Material

A.1 Peer-to-peer Sampling Procedure as a Markov Chain

Let $(\xi_k)_{k\in\mathbb{N}}$ be a time-homogeneous, ergodic Markov chain with the state space $\mathbb{S} = [h]$ and a dummy deterministic initial state $\xi_0 = i_0 \in \Xi$. Define θ as satisfying $\lim_{k\to\infty} \mathbb{P}(\xi_k = i, \xi_{k+1} = j) = \theta_{ij} > 0$ for $i, j \in \mathbb{S}$. Similarly, let $\theta^\star$ encode the stationary probability mass function of (ξ_t, ξ_{t+1}). Then

$$\sum_{j\in\mathbb{S}} \theta^\star_{ij} = \lim_{k\to\infty} \sum_{j\in\mathbb{S}} \mathbb{P}(\xi_k = i, \xi_{k+1} = j) = \lim_{k\to\infty} \mathbb{P}(\xi_k = i) = \lim_{k\to\infty} \sum_{j\in\mathbb{S}} \mathbb{P}(\xi_{k-1} = j, \xi_k = i) = \sum_{j\in\Xi} \theta^\star_{ji}.$$

In other words, the row sums of $\theta^\star$ coincide with its column sums. This property of $\theta^\star$ prompts us to define

$$\Theta = \{\theta \in \mathbb{R}^{h\times h} \mid \theta_{i,j} \geq 0 \text{ for } i, j \in \mathbb{S}; \sum_{i,j\in\mathbb{S}} \theta_{ij} = 1; \sum_{j\in\mathbb{S}} \theta_{ij} = \sum_{j\in\mathbb{S}} \theta_{ji} \text{ for } i \in \mathbb{S}\}.$$

That is, Θ is the set of doubly stochastic probability mass functions with balanced marginals. Every $\theta \in \Theta$ induces a unique row vector $\pi \in \mathbb{R}^{1\times h}$ of stationary state probabilities and a unique transition probability matrix $P \in \mathbb{R}^{h\times h}$ defined via $\pi_i = \sum_{j\in\mathbb{S}} \theta_{ij}$ and $P_{ij} = \theta_{ij}/\pi_i$, respectively. By construction, P is a stochastic matrix whose rows represent strictly positive probability vectors

$$P^K(\xi_1 = i_1, \ldots, \xi_K = i_K) = \prod_{k=1}^{K} P_{i_{k-1}, i_{k+1}}$$

for all $(i_1, \ldots, i_K) \in \mathbb{S}^K$ and $K \in \mathbb{N}$. The stationary distribution π satisfies $\pi P = \pi$. Define the empirical distribution of sequential states as $\widehat{\mathbb{P}}^K \in \mathbb{R}^{h\times h}$ with

$$\widehat{\mathbb{P}}^K_{ij} = \frac{1}{K} \sum_{k=1}^{K} \delta_{(\xi_{k-1}, \xi_k) = (i,j)}$$

for $i, j \in \mathbb{S}$.

A.2 Proof of Theorem 1

Proof: Since $J^\star \leq \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}^\star_Q, \xi)] \leq \widehat{J}^\star_Q + L\|\widehat{\mathbb{P}}^{[s]}_Q - \mathbb{P}\|_p$, then by the assumptions that $\beta_Q \in (0,1)$, $\lim_{Q\to\infty} \varepsilon_Q(\beta_Q) = 0$, and $\sum_{Q=1}^{\infty} \beta_Q < \infty$ and the convergence of the empirical distribution, we have that

$$\mathbb{P}(J^* \leq \widehat{J}^\star_Q + L\varepsilon_Q(\beta_Q)) \geq 1 - \beta_Q \Longrightarrow \mathbb{P}(\liminf_{Q\to\infty} \widehat{J}^\star_Q \geq J^*) = 1.$$

Choose any $\delta > 0$ and fix a δ-optimal solution $x_\delta \in \mathbb{X}$ with

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_Q(\beta)}(\widehat{\mathbb{P}}_Q^{[s]})} \mathbb{E}^{\mathbb{Q}}[\ell(x_\delta, \xi)] \leq \mathbb{E}^{\widehat{\mathbb{P}}_Q^{[s]}}[\ell(x_\delta, \xi)] + \delta.$$

It follows that

$$\begin{aligned}
\limsup_{Q\to\infty} \limsup_{s\to\infty} \widehat{J}_Q^\star &\leq \limsup_{Q\to\infty} \limsup_{s\to\infty} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_Q(\beta)}(\widehat{\mathbb{P}}_Q)} \mathbb{E}^{\mathbb{Q}}[\ell(x_\delta, \xi)] \leq \limsup_{Q\to\infty} \limsup_{s\to\infty} \mathbb{E}^{\widehat{\mathbb{P}}_Q^{[s]}}[\ell(x_\delta, \xi)] + \delta \\
&\leq \limsup_{Q\to\infty} \limsup_{s\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(x_\delta, \xi)] + L\|\mathbb{P} - \widehat{\mathbb{P}}_Q^{[s]}\|_p + \delta \\
&= \mathbb{E}^{\mathbb{P}}[\ell(x_\delta; \xi)] + \delta + \varepsilon \\
&\leq J^\star + 2\delta,
\end{aligned}$$

where the equality holds almost surely under $\mathbb{P}^\infty$ and the mixing conditions since

$$\|\mathbb{P} - \widehat{\mathbb{P}}_Q^{[s]}\| \leq \|\mathbb{P} - \widehat{\mathbb{P}}_Q\| + \|\widehat{\mathbb{P}}_Q - \widehat{\mathbb{P}}_Q^{[s]}\| \leq \varepsilon + \varepsilon_T(\beta_T).$$

By combining this result with the fact that $\mathbb{P}(\|\mathbb{P} - \widehat{\mathbb{P}}_Q\| \leq \varepsilon_Q(\beta_Q)) \geq 1 - \beta_Q$, it follows that $\mathbb{P}(\|\mathbb{P} - \widehat{\mathbb{P}}_Q\| \leq 2\varepsilon_Q(\beta_Q)) \geq 1 - \beta_Q$. Under the assumption that $\sum_{Q=1}^{\infty} \beta_Q \leq \infty$, it follows that $\mathbb{P}(\lim_{Q\to\infty} \lim_{s\to\infty} \|\mathbb{P} - \widehat{\mathbb{P}}_Q^{[s]}\| = 0) = 1$. Since $\delta > 0$ was chosen arbitrarily, we can conclude that $\limsup_{Q\to\infty} \limsup_{s\to\infty} \widehat{J}_Q^\star \leq J^\star$, and so

$$\limsup_{Q\to\infty} \limsup_{s\to\infty} \widehat{J}_Q^\star \leq J^\star \leq \limsup_{s\to\infty} \liminf_{Q\to\infty} \widehat{J}_Q^\star$$

in probability. That is, $\limsup_{Q\to\infty} \lim_{s\to\infty} \widehat{J}_Q^\star = J^\star$.
Now fix an arbitrary realization of the stochastic process $(\xi_Q)_{Q\in\mathbb{N}}$ such that $J^* = \lim_{Q\to\infty} \lim_{s\to\infty} \widehat{J}_Q^\star$ and $J^* \leq \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_Q^\star, \xi)] \leq \widehat{J}_Q^\star + L\|\mathbb{P} - \widehat{\mathbb{P}}_Q^{[s]}\|$ for all sufficiently large Q and s. By the previous result, we know these two conditions are satisfied the probability. Then $\liminf_{Q\to\infty} \liminf_{s\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_Q^\star, \xi)] \leq \lim_{Q\to\infty} \lim_{s\to\infty} \widehat{J}_Q^\star = J^*$. Now let $\lim_{Q\to\infty} \lim_{s\to\infty} \widehat{x}_Q^\star = x^*$. Since $\mathbb{X}$ is closed, $x^* \in \mathbb{X}$. Furthermore,

$$J^* \leq \mathbb{E}^{\mathbb{P}}[\ell(x^*, \xi)] \leq \mathbb{E}^{\mathbb{P}}[\liminf_{Q\to\infty} \liminf_{s\to\infty} \ell(\widehat{x}_Q^\star, \xi)] \leq \liminf_{Q\to\infty} \liminf_{s\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_Q^\star, \xi)] \leq J^*,$$

where the second inequality holds since $\ell(x)$ is lower semicontinuous. Thus, equality holds throughout and the desired result, that $\lim_{Q\to\infty} \lim_{s\to\infty} \widehat{x}_Q^\star = x^*$, follows. □

A.3 Proof of Theorem 4

Proof: We first prove that $\mathbb{P}(\lim_{K\to\infty} \widehat{J}_K^\star = J^\star) = 1$. Note that $J^\star \leq \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star; \xi)] \leq \widehat{J}_K^\star + L\|\widehat{\mathbb{P}}_K - \mathbb{P}\|$. Under the assumption that $\beta_K \in (0,1)$, $\lim_{K\to\infty} \varepsilon_K(\beta_K) = 0$ and $\mathbb{P}(J^\star \leq \widehat{J}_K^\star + L\varepsilon_K(\beta_K)) \geq 1 - \beta_K$. Therefore, $\mathbb{P}(\liminf_{K\to\infty} \widehat{J}_K^\star \geq J^\star) = 1$.

Choose any $\delta > 0$, fix a δ-optimal solution $x_\delta \in \mathbb{X}$ with $\mathbb{E}^{\mathbb{P}}[\ell(x_\delta; \xi)] \leq J^\star + \delta$, and let $\widehat{\mathbb{Q}}_K \in \mathbb{B}_{\varepsilon_K(\beta_K)}(\widehat{\mathbb{P}}_K)$ be a δ-optimal distribution corresponding to x_δ with

$$\sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon_K(\beta_K)}(\widehat{\mathbb{P}}_K)} \mathbb{E}^{\mathbb{Q}}[\ell(x_\delta; \xi)] \leq \mathbb{E}^{\widehat{\mathbb{Q}}_K}[\ell(x_\delta; \xi)] + \delta.$$

It follows that

$$\begin{aligned}
\limsup_{K\to\infty} \widehat{J}_K^\star &\le \limsup_{K\to\infty} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon_K(\beta_K)}(\widehat{\mathbb{P}}_K)} \mathbb{E}^{\mathbb{Q}}[\ell(x_\delta;\xi)] \\
&\le \limsup_{K\to\infty} \mathbb{E}^{\widehat{\mathbb{Q}}_K}[\ell(x_\delta;\xi)] + \delta \\
&\le \limsup_{K\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(x_\delta;\xi)] + L\|\mathbb{P} - \widehat{\mathbb{Q}}_K\| + \delta. \\
&= \limsup_{K\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(x_\delta;\xi)] + \delta \\
&\le J^\star + 2\delta.
\end{aligned}$$

The equality holds $\mathbb{P}$-almost surely since $\mathbb{P}\{\|\mathbb{P} - \widehat{\mathbb{Q}}_K\| \le 2\varepsilon_K(\beta_K)\} \ge 1 - \beta_K$ because

$$\|\mathbb{P} - \widehat{\mathbb{Q}}_K\| \le \|\mathbb{P} - \widehat{\mathbb{P}}_K\| + \|\widehat{\mathbb{P}}_K - \widehat{\mathbb{Q}}_K\| \le \|\mathbb{P} - \widehat{\mathbb{P}}_K\| + \varepsilon_K(\beta_K)$$

and $\mathbb{P}(\|\mathbb{P} - \widehat{\mathbb{P}}_K\| \le \varepsilon_K(\beta_K)) \ge 1 - \beta_K$.

Under the assumption that $\sum_{K=1}^{\infty} \beta_K \le \infty$, the Borel–Cantelli lemma implies that $\mathbb{P}(\lim_{K\to\infty} \|\mathbb{P} - \widehat{\mathbb{Q}}_K\| = 0) = 1$. As $\delta > 0$ was chosen arbitrarily, we conclude that, $\mathbb{P}$-almost surely, $\limsup_{K\to\infty} \widehat{J}_K^\star \le J^\star$. Therefore, $\mathbb{P}(\lim_{K\to\infty} \widehat{J}_K^\star = J^\star) = 1$.

We next prove that $\mathbb{P}(\lim_{K\to\infty} \widehat{x}_K^\star = x^\star) = 1$. Fix an arbitrary realization of the stochastic process $(\xi_k)_{k\in\mathbb{N}}$ such that $J^\star = \lim_{K\to\infty} \widehat{J}_K^\star$ and $J^\star \le \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star;\xi)] \le \widehat{J}_K^\star + L\|\mathbb{P} - \widehat{\mathbb{P}}_K\|$ for all sufficiently large K. From the previous result, we know these two conditions are satisfied in probability. Then

$$\liminf_{K\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star;\xi)] \le \lim_{K\to\infty} \widehat{J}_K^\star = J^\star$$

almost surely. Let $\lim_{K\to\infty} \widehat{x}_K^\star = x^\star$, where $x^\star \in \mathbb{X}$ since $\mathbb{X}$ is closed. Now,

$$J^\star \le \mathbb{E}^{\mathbb{P}}[\ell(x^\star;\xi)] \le \mathbb{E}^{\mathbb{P}}[\liminf_{K\to\infty} \ell(\widehat{x}_K^\star;\xi)] \le \liminf_{K\to\infty} \mathbb{E}^{\mathbb{P}}[\ell(\widehat{x}_K^\star;\xi)] \le J^\star,$$

where the second inequality holds because ℓ is lower semicontinuous as ℓ is convex, closed and proper. Thus, $\mathbb{E}^{\mathbb{P}}[\ell(x^\star;\xi)] = J^\star$ and so $\mathbb{P}(\lim_{K\to\infty} \widehat{x}_K^\star = x^\star) = 1$.

□

A.4 Proof of Lemma 1

Proof:

$$|\mathrm{Cov}\,(f(\xi_i), f(\xi_j))| = |\mathbb{E}[f(\xi_i)f(\xi_j)] - (\mathbb{E}[f(\xi_i)])\,(\mathbb{E}[f(\xi_j)])$$

Let $\xi = \xi_i$ and $\eta = \xi_j$, where ξ_i is $\mathcal{F}_1^k$ measurable, ξ_j is $\mathcal{F}_{n+k}^\infty$ measurable. Thus, we have that

$$\begin{aligned}
|\mathbb{E}[f(\xi)f(\eta)] - (\mathbb{E}[f(\xi)])\,(\mathbb{E}[f(\eta)])| &= |\mathbb{E}\left[\mathbb{E}[f(\xi)f(\eta) \mid \mathcal{F}_1^k]\right] - (\mathbb{E}[f(\xi)])\,(\mathbb{E}[f(\eta)])| \\
&= |\mathbb{E}\left[f(\xi) \cdot \mathbb{E}[f(\eta) \mid \mathcal{F}_1^k]\right] - (\mathbb{E}[f(\xi)])\,(\mathbb{E}[f(\eta)])|
\end{aligned}$$

(With the assumption $\sup_{\xi,x}|f(x,\xi)| \leq M$.)

$$\leq M \cdot \mathbb{E}\left|\mathbb{E}[f(\eta) \mid \mathcal{F}_1^k] - (\mathbb{E}[f(\eta)])\right|$$

(Let $\tilde{\xi} = \operatorname{sign}\left(\mathbb{E}[f(\eta) \mid \mathcal{F}_1^k] - (\mathbb{E}[f(\eta)])\right)$ is $\mathcal{F}_1^k$ measurable.)

$$\begin{aligned}
&= M \cdot \mathbb{E}\left[\tilde{\xi}\left(\mathbb{E}[f(\eta) \mid \mathcal{F}_1^k] - (\mathbb{E}[f(\eta)])\right)\right] \\
&= M \cdot \mathbb{E}\left[\tilde{\xi}\mathbb{E}[f(\eta) \mid \mathcal{F}_1^k] - \tilde{\xi}\,(\mathbb{E}[f(\eta)])\right] \\
&= M \cdot \left[\mathbb{E}\left[\tilde{\xi}f(\eta)\right] - \left(\mathbb{E}[\tilde{\xi}]\right)(\mathbb{E}[f(\eta)])\right]
\end{aligned}$$

$$\begin{aligned}
|\mathbb{E}\left[\tilde{\xi}f(\eta)\right] - \left(\mathbb{E}[\tilde{\xi}]\right)(\mathbb{E}[f(\eta)])| &= \left|\mathbb{E}\left[\mathbb{E}\left[\tilde{\xi}f(\eta) \mid \mathcal{F}_{n+k}^\infty\right]\right] - \left(\mathbb{E}[\tilde{\xi}]\right)(\mathbb{E}[f(\eta)])\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}f(\eta)\left[\tilde{\xi} \mid \mathcal{F}_{n+k}^\infty\right]\right] - \left(\mathbb{E}[\tilde{\xi}]\right)(\mathbb{E}[f(\eta)])\right| \\
&\leq M \cdot \left|\mathbb{E}\left[\mathbb{E}\left[\tilde{\xi} \mid \mathcal{F}_{n+k}^\infty\right]\right] - \left(\mathbb{E}[\tilde{\xi}]\right)\right|
\end{aligned}$$

(Let $\tilde{\eta} = \operatorname{sign}\left(\mathbb{E}\left[\tilde{\xi} \mid \mathcal{F}_{n+k}^\infty\right] - \left(\mathbb{E}[\tilde{\xi}]\right)\right)$ is $\mathcal{F}_{n+k}^\infty$ measurable.)

$$\begin{aligned}
&= M \cdot \mathbb{E}\left[\tilde{\eta}\left(\mathbb{E}\left[\tilde{\xi} \mid \mathcal{F}_{n+k}^\infty\right] - \left(\mathbb{E}[\tilde{\xi}]\right)\right)\right] \\
&= M \cdot \left[\mathbb{E}[\tilde{\eta}\tilde{\xi}] - (\mathbb{E}[\tilde{\eta}])\left(\mathbb{E}[\tilde{\xi}]\right)\right]
\end{aligned}$$

It follows that

$$|\mathbb{E}[f(\xi)f(\eta)] - (\mathbb{E}[f(\eta)]\mathbb{E}[f(\xi)])| \leq M^2 \cdot \left|\mathbb{E}[\tilde{\eta}\tilde{\xi}] - (\mathbb{E}[\tilde{\eta}])\left(\mathbb{E}[\tilde{\xi}]\right)\right|$$

Let $B = \{\tilde{\eta} = 1\} \in \mathcal{F}_{n+k}^\infty$, $A = \{\tilde{\xi} = 1\} \in \mathcal{F}_1^k$. Therefore,

$$\begin{aligned}
&|\mathbb{E}[f(\xi)f(\eta)] - (\mathbb{E}[f(\eta)]\mathbb{E}[f(\xi)])| \\
&\leq M^2 \cdot \left|\mathbb{P}(AB) + \mathbb{P}(\bar{A}\bar{B}) - \mathbb{P}(A\bar{B}) - \mathbb{P}(\bar{A}B) - \left[\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(\bar{B}) - \mathbb{P}(\bar{A})\mathbb{P}(B) + \mathbb{P}(\bar{A})\mathbb{P}(\bar{B})\right]\right| \\
&= M^2 \cdot \left|[\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)] + \left[\mathbb{P}(\bar{A}\bar{B}) - \mathbb{P}(\bar{A})\mathbb{P}(\bar{B})\right] - \left[\mathbb{P}(\bar{A}B) - \mathbb{P}(\bar{A})\mathbb{P}(B)\right] - \left[\mathbb{P}(A\bar{B}) - \mathbb{P}(A)\mathbb{P}(\bar{B})\right]\right| \\
&\leq M^2 \cdot 4\alpha(n) \leq 4M^2 \cdot \phi(n).
\end{aligned}$$

Then, we have that

$$|\mathrm{Cov}(f(\xi_k), f(\xi_{k+n}))| \leq 4\|f\|_M^2\phi(n),$$

□

A.5 Proof of Lemma 2

Proof: With the notation stated, we are now presenting the bound of the term $\bar{U}_k$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\exp(t\bar{U}_k)\right] = \mathbb{E}\left[\exp\left(\frac{t}{k}\sum_{i=1}^{r}U_i\right)\right] \\
&= \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right) + \left(\mathbb{E}[\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right)]\right)\left(\mathbb{E}[\exp\left(\frac{t}{k}U_r\right)]\right) \\
&= \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right) + \left(\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right), \exp\left(\frac{t}{k}U_{r-1}\right)\right)\right) \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&= \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right) + \left(\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&= \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right) + \left(\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&\quad + \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-3}U_i\right), \exp\left(\frac{t}{k}U_{r-2}\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}\sum_{i=1}^{r-3}U_i\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-2}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&= \\
&\ \vdots \\
&= \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-1}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right) + \left(\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-2}U_i\right), \exp\left(\frac{t}{k}U_r\right)\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) \\
&\quad + \mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-3}U_i\right), \exp\left(\frac{t}{k}U_{r-2}\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_r\right)\right]\right) + \cdots \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_1\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_2\right)\right]\right)\cdots\left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_{r-1}\right)\right]\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_r\right)\right]\right)
\end{aligned}$$

For jth term,

$$\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-j}U_i\right), \exp\left(\frac{t}{k}U_{r-j+1}\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_{r-j+2}\right)\right]\right)\cdots\left(\mathbb{E}\left[\exp\left(\frac{t}{K}U_r\right)\right]\right)$$

$$\mathrm{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-j}U_i\right), \exp\left(\frac{t}{k}U_{r-j+1}\right)\right) \leq \frac{t^2}{k^2}\exp\left(\frac{t}{k}(r-j+1)2PM\right)\sum_{j\in A_1}\sum_{k\in A_2}\mathrm{Cov}\left(U_j, U_k\right)$$

where

$$
\begin{aligned}
A_1 &= \left\{ \underbrace{1, \cdots, p}_{A_{1,1}}, \underbrace{2p+1, \cdots, 3p}_{A_{1,2}}, \cdots, \underbrace{2(r-j-1)p+1, \cdots, (2(r-j)-1)p}_{A_{1,r-j}} \right\} \\
A_2 &= \{2(r-j)p+1, \cdots, (2(r-j)+1)p\} \\
A_{1i} &= \{2(i-1)p+1, \cdots, (2i-1)p\}
\end{aligned}
$$

$$
\begin{aligned}
\sum_{j \in A_{11}} \sum_{l \in A_2} \operatorname{Cov}(U_j, U_l) &\leq p \cdot \operatorname{Cov}\left(Y_1, Y_{2(r-j)p+1}\right) + p \cdot \operatorname{Cov}\left(Y_2, Y_{2(r-j)+1}\right) \\
&\quad + \cdots + p \cdot \operatorname{Cov}\left(Y_p, Y_{2(r-j)p+1}\right) \\
\sum_{j \in A_{1i}} \sum_{l \in A_2} \operatorname{Cov}(U_j, U_l) &\leq p \cdot \operatorname{Cov}\left(Y_{2(i-1)p+1}, Y_{2(i-j)p+1}\right) + p \cdot \operatorname{Cov}\left(Y_{2(i-1)p+2}, Y_{2(i-j)p+1}\right) \\
&\quad + \cdots + p \cdot \operatorname{Cov}\left(Y_{2(i-1)p+p}, Y_{2(r-j)p+1}\right) \\
&= p \cdot \sum_{(2y-2j-2i+2)p-p+1}^{(2r-2j-2i+2)p} \operatorname{Cov}(Y_1, Y_l)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_{i=1}^{r-j} \operatorname{Cov}(U_i, U_{r-j+1}) &\leq \sum_{i=1}^{r-j} p \cdot \sum_{(2y-2j-2i+2)p-p+1}^{(2r-2j-2i+2)p} \operatorname{Cov}(Y_1, Y_l) \\
&\leq \sum_{i=1}^{r-j} p \cdot 4M^2 \sum_{l} \phi(l), (\text{ define } \phi(l) = \operatorname{Cov}(Y_1, Y_l)) \\
&= 4pM^2 \sum_{i=1}^{r-j} \sum_{(2y-2j-2i+2)p-p+1}^{(2r-2j-2i+2)p} \phi(l) \leq 4pM^2 \sum_{l=p}^{\infty} \phi(l)
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(t\bar{U}_k\right)\right] &\leq \sum_{j=1}^{r-1} \operatorname{Cov}\left(\exp\left(\frac{t}{k}\sum_{i=1}^{r-j} U_i\right), \exp\left(\frac{t}{k}U_{r-j+1}\right)\right)\left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{j-1} \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{r} \\
&\leq \sum_{j=1}^{r-1} \frac{t^2}{k^2} \exp\left(\frac{t}{k}(r-j+1)2pM\right)\left(4pM^2\right) \sum_{l=p}^{\infty} \phi(p) \cdot \left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{j-1} \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{r} \\
&= \frac{4pM^2 t^2 V(p)}{k^2} \exp\left(\frac{t}{k}2pMr\right) \sum_{j=0}^{r-2} \exp\left(-\frac{2pMt}{k}j\right) \cdot \left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{j} \\
&\quad + \left(\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right]\right)^{r}
\end{aligned}
$$

Note that $|U_1| \leq 2pM$, $\mathbb{E}\left[\exp\left(\frac{t}{k}U_1\right)\right] \leq \exp\left(\frac{t^2}{k^2} \cdot \frac{1}{8}\left(4pM\right)^2\right) = \exp\left(\frac{2p^2M^2t^2}{k^2}\right)$. Therefore

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\bar{U}_k\right)\right] \leq& \frac{4pM^2t^2V(p)}{k^2}\exp\left(\frac{t}{k}2pMr\right)\sum_{j=0}^{r-2}\exp\left(-\frac{2pMt}{k}j\right)\exp\left(\frac{2p^2M^2t^2j}{k^2}\right)\\ &+\exp\left(\frac{2p^2M^2t^2r}{k^2}\right)\\ \leq& \frac{4pM^2t^2V(p)}{k^2}\exp\left(\frac{t}{k}2pMr\right)\sum_{j=0}^{r-2}\exp\left(\frac{\left(2p^2M^2t^2-2pMkt\right)j}{k^2}\right)\\ &+\exp\left(\frac{2p^2M^2t^2r}{k^2}\right)\end{aligned}$$

Based on Markov's Inequality,

$$\begin{aligned}\mathbb{P}\left(\bar{U}_k \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[\exp\left(t\bar{U}_k\right)\right]}{\exp\left(t\varepsilon\right)} \leq& \frac{4pM^2t^2V(p)}{k^2}\exp\left(-\frac{t}{\varepsilon}+\frac{2pMrt}{k}\right)\sum_{j=0}^{r-2}\exp\left(\frac{2pMt(pMt-k)}{k^2}j\right)\\ &+\exp\left(\frac{2p^2M^2t^2r}{k^2}-t\varepsilon\right)\end{aligned}$$

Let $t = \frac{k^2\varepsilon}{4p^2M^2r}$,

$$\begin{aligned}\mathbb{P}\left(\bar{U}_k \geq \varepsilon\right) \leq& \frac{4pM^2t^2V(p)}{k^2}\exp\left(-t\varepsilon+\frac{2pMrt}{k}\right)\sum_{j=0}^{r-2}\exp\left(\frac{2pMt(pMt-k)}{k^2}j\right)\\ &+\exp\left(-\frac{k^2\varepsilon^2}{8p^2M^2r}\right)\end{aligned}$$

If $\varepsilon < M$, under the fixed value $t = \frac{k^2\varepsilon}{4p^2M^2r}$, $\sum_{j=0}^{r-2}\exp\left(\frac{2pMt(pMt-k)}{k^2}j\right)$ is convergent. It follows that there exists a constant c such that

$$\begin{aligned}\mathbb{P}\left(\bar{U}_k \geq \varepsilon\right) &\leq \left(c \cdot \frac{\varepsilon^2k^2V(p)}{4p^3r^2M^2}\exp\left(\frac{k\varepsilon}{2pM}\right)+1\right)\exp\left(-\frac{k^2\varepsilon^2}{8p^2M^2r}\right)\\ &= \left(c \cdot \frac{k^2\varepsilon^2V(p)}{4p^3r^2M^2}\exp\left(\frac{k\varepsilon}{2pM}\right)+1\right)\exp\left(-\frac{k^2\varepsilon^2}{8p^2M^2r}\right)\\ &\leq \left(c \cdot \frac{k^2\varepsilon^2}{4p^3r^2M^2}\exp\left(\frac{k\varepsilon}{2pM}\right)V(p)+1\right)\exp\left(-\frac{k\varepsilon}{4pM^2}\right)\end{aligned}$$

Assumption $\phi(p) \leq \phi_0\exp\left(\phi_1k^a\right)$, for some $\phi_0 > 0, \phi_1 > 0, a > 0$, S.P. is geometric mixing sequence. Then,

$$\mathbb{P}\left(\bar{U}_k \geq \varepsilon\right) \leq c \cdot \exp\left(-\frac{k\varepsilon^2}{4pM^2}\right)$$

It is obvious that $\bar{V}_k$ satisfies the same inequalities as $\bar{U}_k$, that is,

$$\mathbb{P}\left(\bar{V}_k \geq \varepsilon\right) \leq c \cdot \exp\left(-\frac{k\varepsilon^2}{4pM^2}\right)$$

We next show that the term $\bar{W}_k$ can be dispensed. $\bar{W}_k$ contains $n-2pr$ terms and $n-2pr < p$. Then, $|\bar{W}_k| < p \cdot \frac{2M}{K}$, so that,

$$\mathbb{P}\left(|\bar{W}_k| \geq \varepsilon\right) \leq \mathbb{P}\left(M \geq \frac{k\varepsilon}{2p}\right) = 0,$$

for sufficiently large k, because $\frac{k}{p} \to \infty$. As a result, $\mathbb{P}\left(|\bar{W}_k| \geq \varepsilon\right) = 0$, for $k \to \infty$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(|\bar{S}_k| \geq \varepsilon\right) &\leq \mathbb{P}\left(|\bar{U}_k| \geq \frac{\varepsilon}{3}\right) + \mathbb{P}\left(|\bar{V}_k| \geq \frac{\varepsilon}{3}\right) + \mathbb{P}\left(|\bar{W}_k| \geq \frac{\varepsilon}{3}\right) \\
&\leq \mathbb{P}\left(|\bar{U}_k| \geq \frac{\varepsilon}{3}\right) + \mathbb{P}\left(|\bar{V}_k| \geq \frac{\varepsilon}{3}\right), (\text{ for } k \geq k_0, \text{ say })\\
&\leq 4c \cdot \exp\left(-\frac{k\varepsilon^2}{36pM^2}\right)
\end{aligned}$$

□

A.6 Proof of Theorem 5

Proof: Under the assumptions, we have that

$$\mathbb{P}\left(|\bar{S}_k| \geq \varepsilon\right) \leq c \cdot \exp\left(-\frac{k\varepsilon^2}{36pM^2}\right)$$

for some constant c. To proof

$$\frac{1}{\sqrt{K}}\left[\sum_{k=1}^{K}\left(\ell(x,\xi_k) - \mathbb{E}\left[\ell(x;\xi)\right]\right)\right] \to \mathcal{N}(0,\sigma^2)$$

i.e.

$$\frac{1}{\sqrt{K]}}\left(\sum_{k=1}^{K} Y_k\right) \to \mathcal{N}(0,\sigma^2), \text{ where } \frac{1}{K}\operatorname{Var}\left(\sum_{k=1}^{K}\ell(x,\xi_k)\right) \to \sigma^2 \in (0,\infty)$$

for any $x \in \mathbb{R}^d$. Given $p \in \mathbb{N}$, let $m = \lfloor \frac{k}{p} \rfloor$, and redefine the blocks,

$$U_j = \sum_{i=(j-1)p+1}^{jp} Y_i, \quad j = 1, \cdots, m, \quad U_{m+1} = \sum_{i=mp+1}^{K} Y_i$$

Let $\varphi_k(t)$ represnet the characteristic function of $\frac{1}{\sqrt{K}}S_K$, where $S_K = \sum_{k=1}^{K} Y_k$. We only need to establish that $\left|\varphi_k(t) - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right| \to 0$. Decompose $\left|\varphi_k(t) - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right|$ into the form

$$\begin{aligned}
&\left|\varphi_k(t) - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right| \\
&= \left|\varphi_k(t) - \varphi_{mp}(t) + \varphi_{mp}(t) - \varphi_p^m(t) + \varphi_p^m(t) - \exp\left(-\frac{t^2\sigma_p^2}{2} +\right) + \exp\left(-\frac{t^2\sigma_p^2}{2} +\right) - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right| \\
&\leq \left|\varphi_k(t) - \varphi_{mp}(t)\right| + \left|\varphi_{mp}(t) - \varphi_p^m(t)\right| + \left|\varphi_p^m(t) - \exp\left(-\frac{t^2\sigma_p^2}{2} +\right)\right| + \left|\exp\left(-\frac{t^2\sigma_p^2}{2} +\right) - \exp\left(-\frac{\sigma^2 t^2}{2}\right)\right| \\
&= D_1 + D_2 + D_3 + D_4
\end{aligned}$$

$$\begin{aligned}
D_1 &\leq \mathbb{E}\left|\exp\left(\frac{itS_k}{\sqrt{K}}\right) - \exp\left(\frac{itS_{mp}}{\sqrt{mp}}\right)\right| \leq \mathbb{E}\left|t\left(\frac{S_K}{\sqrt{K}} - \frac{S_{mp}}{\sqrt{mp}}\right)\right| \leq |t|\left(\mathbb{E}\left[\left(\frac{S_K}{\sqrt{K}} - \frac{S_{mp}}{\sqrt{mp}}\right)^2\right]\right)^{\frac{1}{2}} \\
&= |t|\left(\mathbb{E}\left[\left(\frac{1}{\sqrt{K}}(S_{mp} + U_{m+1}) - \frac{S_{mp}}{\sqrt{mp}}\right)^2\right]\right)^{\frac{1}{2}} = |t|\left\{\mathbb{E}\left[\left(\frac{1}{\sqrt{K}} - \frac{1}{\sqrt{mp}}\right)S_{mp} + \frac{1}{\sqrt{K}}U_{m+1}\right]^2\right\}^{\frac{1}{2}} \\
&= |t|\left\{\left(\frac{1}{\sqrt{K}} - \frac{1}{\sqrt{mp}}\right)^2\left(\mathbb{E}[S_{mp}^2]\right) + \frac{1}{K}\left(\mathbb{E}[U_{m+1}^2]\right)\right\}^{\frac{1}{2}}
\end{aligned}$$

It follows from the stationary of the sequence of random variables ξ_k that for m large enough, $\mathbb{E}[S_{mp}^2] \le 2\sigma^2 \cdot mp$ and $\mathbb{E}[U_{m+1}^2] \le 2\sigma^2(k-mp) < 2\sigma^2 p$. Therefore, as $k \to \infty$, which implies that $m \to \infty$, it follows that

$$
\begin{aligned}
|\varphi_k(t) - \varphi_{mp}(t)| &\le |t| \left\{ \left(\frac{1}{\sqrt{K}} - \frac{1}{\sqrt{mp}} \right)^2 \cdot 2\sigma^2 mp + \frac{1}{K} \cdot 2\sigma^2 p \right\}^{\frac{1}{2}} \\
&= |t| \left\{ \left(\sqrt{\frac{mp}{K}} - 1 \right)^2 s\sigma^2 + \frac{2\sigma^2 p}{K} \right\}^{\frac{1}{2}} \to 0.
\end{aligned}
$$

Denote $W_j = \frac{1}{\sqrt{P}} U_j$ with characteristic function $\varphi_p(t)$, $\sigma_p^2 = \frac{1}{\sqrt{p}} \operatorname{Var}(S_p)$.

$$
\begin{aligned}
D_2 &= \left| \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{m}} \sum_{j=1}^{m} W_j \right) \right] - \prod_{j=1}^{m} \mathbb{E}\left[\exp\left(\frac{it}{\sqrt{m}} W_j \right) \right] \right| \le \sum_{j=2}^{m} \frac{t^2}{m} \sum_{k=1}^{j-1} |\operatorname{Cov}(W_j, W_k)| \\
&= \sum_{l=2}^{m} \frac{t^2}{mp} \sum_{j=1}^{l-1} |\operatorname{Cov}(U_l, U_j)| = \frac{t^2}{mp} \sum_{l=2}^{m} \sum_{k \in A} \sum_{j \in B} |\operatorname{Cov}(Y_k, Y_j)|,
\end{aligned}
$$

where

$$
\begin{aligned}
A &= \left\{ \underbrace{1, 2, \cdots, p}_{A_1}, \underbrace{p+1, p+2, \cdots, 2p}_{A_2}, \cdots, \underbrace{(l-2)p+1, (l-2)p+2, \cdots, (l-1)p}_{A_{l-1}} \right\} \\
B &= \{(l-1)p+1, (l-1)p+2, \cdots, lp\}
\end{aligned}
$$

$$
\begin{aligned}
D_2 &\le \frac{t^2}{mp} \sum_{l=2}^{m} \sum_{j=1}^{l-1} \left[\sum_{u=0}^{p-2} (p-1-u)\phi((l-j)p-u) + \sum_{v=1}^{p} \phi((l-j)p+v) \right] \\
&= \frac{t^2}{mp} \sum_{l=2}^{m} \sum_{j=1}^{l-1} \sum_{u=0}^{p-2} (p-u-1)\phi((l-j)p-u) + \frac{t^2}{mp} \sum_{l=2}^{m} \sum_{j=1}^{l-1} \sum_{v=1}^{p} (p-v-1)\phi((l-j)p+v) \\
&= \frac{t^2}{mp} \sum_{l=1}^{m-1} (m-l) \sum_{k=1}^{p} k\phi((l-1)p+k) + \frac{t^2}{mp} \sum_{l=1}^{m-1} (m-l) \sum_{k=1}^{p-1} (p-k)\phi(k+lp) \\
&= \frac{t^2}{mp} \sum_{l=1}^{m-1} (m-l) \sum_{k=1}^{p} k\phi(k) + \frac{t^2}{mp} \sum_{l=2}^{m-1} (m-l)p \sum_{k=1}^{p} \phi(k+lp) \\
&= \frac{t^2}{p} \sum_{l=1}^{m-1} (1 - \frac{l}{m}) \sum_{l=1}^{p} k\phi(k) + \frac{pt^2}{p} \sum_{l=2}^{m-1} (1 - \frac{l}{m}) \sum_{k=1}^{p} \phi(k+lp) \\
&\le \frac{t^2}{p} \sum_{l=1}^{m-1} (1 - \frac{l}{m}) \sum_{l=1}^{p} k\phi(k) + t^2 \sum_{l=2}^{m-1} (1 - \frac{l}{m}) \sum_{k=lp}^{\infty} \phi(k) \\
&(\sum_{l=1}^{p} k\phi(k) < \infty, \sum_{k=p}^{\infty} \phi(k) \to 0) \\
&\to 0, \text{ as } p \to \infty.
\end{aligned}
$$

$$
\begin{aligned}
D_2 &\le \frac{t^2}{m} \sum_{l=2}^{m} \sum_{j=1}^{l-1} \sum_{u=(l-j)p-(p-2)}^{(l-j)p} \operatorname{Cov}(Y_1, Y_u) = \frac{t^2}{m} \sum_{l=2}^{m} \sum_{j=1}^{l-1} \sum_{u=(l-j)p-(p-2)}^{(l-j)p} \phi(u) \\
&\le \frac{t^2}{m}(m-1) \sum_{u=1}^{\infty} \phi(u) \le c \cdot t^2
\end{aligned}
$$

For D_3 and D_4,

$$D_3; \lim_{m\to\infty} \left| \varphi_p^m(t) - \exp\left(-\frac{t^2\sigma_p^2}{2}\right) \right| \to 0, \quad D_4 \leq \frac{t^2}{2} \left|\sigma_p^2 - \sigma^2\right|$$

it follows that

$$\lim_{k\to\infty} \sup \left| \varphi_k(t) - \exp\left(-\frac{\sigma^2 t^2}{2}\right) \right| \leq \frac{t^2}{2} \left|\sigma_p^2 - \sigma^2\right| + c \cdot t^2.$$

Based on the fact that $\lim_{p\to+\infty} \sigma_p^2 = \sigma^2$, we have

$$\lim_{k\to\infty} \sup \left| \varphi_k(t) - \exp\left(-\frac{\sigma^2 t^2}{2}\right) \right| = 0$$

The next proof is based on the functional CLT and delta method. Consider the Banach space $C(\mathbb{X})$ of continuous functions $h_i : \mathbb{X} \to \mathbb{R}$ equipped with the sup-norm $|h| := \sup_{x\in\mathbb{X}} |h(x)|$. Define $H(h) = \inf_{x\in\mathbb{X}} h(x)$. Since $\mathbb{X}$ is a closed, convex subset of $\mathbb{R}^d$, the function $H : C(\mathbb{X}) \to \mathbb{R}$ is real value and measurable. Moreover, it can be shown that

$$|H(h1) - H(h_2)| \leq \inf_{x\in\mathbb{X}} |h_1(x) - h_2(x)|| \leq |h_1 - h_2|$$

for any $h_1, h_2 \in C(\mathbb{X})$. That is, $H(\cdot)$ is Lipschitz continuous with Lipschitz constant 1. It follows from Danskin theorem that $H(\cdot)$ is directionally differentiable at any $F \in C(\mathbb{X})$, and $H'_F(\sigma) = \inf_{x\in F(\bar{\mathbb{X}})} \sigma(x), \forall \sigma \in C(\mathbb{X}), F(\bar{\mathbb{X}}]) = \arg\min_{x\in\mathbb{X}} F(x)$. Since $H(\cdot)$ is Lipschitz continuous and directionally differentiable, we have that $H(\cdot)$ is Hadamard directionally differentiable at any $F \in C(\mathbb{X})$. Note that $\widehat{J}_K^\star = H(\widehat{J}_K), J^\star = H(J)$, where $\widehat{J}_K = \frac{1}{K}\sum_{k=1}^K \ell(x, \xi_k), J = \mathbb{E}[\ell(x;\xi)]$. By applying Delta method and using the factor $\sqrt{K}(\widehat{J}_K - J) \to \mathcal{N}(0, \sigma_x^2)$ for any $x \in \mathbb{X}$, we have that

$$\sqrt{K}(\widehat{J}_K^\star - J) = H'^J\left(\sqrt{K}(\widehat{J}_K) - J\right) + O_p(1) \to \mathcal{N}(0, \sigma_{x^\star}^2)$$

□

A.7 Proof of Theorem 11

Proof: Note that

$$\begin{aligned}\mathbb{E}\Big[\Big\| \frac{1}{\Lambda_K} \sum_{k=1}^K \lambda_k (T(x^k) - x^k) \Big\| \mid \mathcal{F}_k\Big] &= \frac{1}{\Lambda_K} \mathbb{E}\Big[\Big\| \sum_{k=1}^K (x^{k+1} - x^k - \lambda_k \epsilon_k) \Big\| \mid \mathcal{F}_k\Big] \\ &= \frac{1}{\Lambda_K} \mathbb{E}[\|x^{k+1} - x^1 - \lambda_k \epsilon_k\|].\end{aligned}$$

Additionally,

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - x^*\| \mid \mathcal{F}_k] &= \mathbb{E}[\|T_\lambda(x^k; \xi_{k+1}) - x^* + \lambda_k \epsilon_k\| \mid \mathcal{F}_k] \\ &\leq \int_\Xi \|T_\lambda(x^k;\xi) - x^*\| (p_{[k]}^{k+1} - p)(\xi) \mathrm{d}\xi + \int_\Xi \|T_\lambda(x^k;\xi) - x^*\| p(\xi) \mathrm{d}\xi + \mathbb{E}[\|\lambda_k \epsilon_k\|] \\ &\leq 2r d_{\mathrm{TV}}(P_k^{k+1}, \mathbb{P}) + \mathbb{E}^{\mathbb{P}}[\|T(x^k;\xi) - x^*\| \mid \mathcal{F}_k] + \mathbb{E}[\|\lambda_k \epsilon_k\|] \\ &\leq 2r d_{\mathrm{TV}}(P_k^{k+1}, \mathbb{P}) + \|x^k - x^*\| + \mathbb{E}[\|\lambda_k \epsilon_k\|] \\ &\leq r(1 + \phi(1)) + \mathbb{E}[\|\lambda_k \epsilon_k\|].\end{aligned}$$

Therefore,

$$\mathbb{E}\Big[\Big\|\frac{1}{\Lambda_K}\sum_{k=1}^{K}\lambda_k(T(x^k)-x^k)\Big\|\Big] \le \frac{2r(1+\phi(1))+2\sum_{k=1}^{K}\mathbb{E}[\|\lambda_k\epsilon_k\|]}{\Lambda_K}.$$

□

A.8 Proof of Theorem 12

Proof: Under Assumption 9 and the requirement that T is a nonexpansive operator,

$$\|\mathbb{E}^{\widehat{\mathbb{P}}_K}[T(x;\xi)-x]-\mathbb{E}^{\mathbb{P}}[T(x;\xi)-x]\| \le 2r\|\widehat{\mathbb{P}}_K-\mathbb{P}\|.$$

This implies that

$$\begin{aligned}
&\mathbb{E}\{\|\mathbb{E}^{\widehat{\mathbb{P}}_K}[T(x;\xi)-x]-\mathbb{E}^{\mathbb{P}}[T(x;\xi)-x]\|\}\\
&\quad=\int_0^\infty \mathbb{P}\{\|\mathbb{E}^{\widehat{\mathbb{P}}_K}[T(x;\xi)-x]-\mathbb{E}^{\mathbb{P}}[T(x;\xi)-x]\|\ge t\}\mathrm{d}t\\
&\quad\le\int_0^\infty \mathbb{P}\{2r\|\widehat{\mathbb{P}}_K-\mathbb{P}\|\ge t\}\mathrm{d}t\\
&\quad\le\int_0^\infty 2\exp\Big(-\frac{K^2t^2}{8r^2C\sum_{k=1}^{\infty}\phi(k)}\Big)\mathrm{d}t\\
&\quad=\int_0^\infty t^{-1/2}\exp\Big(-\frac{K^2t}{8r^2C\sum_{k=1}^{\infty}\phi(k)}\Big)\mathrm{d}t\\
&\quad=\Big(\frac{8r^2C\sum_{k=1}^{\infty}\phi(k)}{K^2}\Big)^{1/2}\Gamma\Big(\frac{1}{2}\Big).
\end{aligned}$$

□

A.9 Proof of Corollary 3

Proof: Corollary 3 is a consequence of Theorems 12–13. We show that S-GFBS is a special case of the S-KM algorithm. Set $T_1(x)=\mathcal{J}_{\gamma_k\partial g}(x)$ and $T_2=(I-\gamma_k\nabla f)(x)$. Because $\mathcal{J}_{\gamma_k\partial g}(x)$ is (1/2)-averaged and $I-\gamma_k\partial f(x)$ is $(\gamma_k/(2\beta))$-averaged, it follows that T_{FB} is $2\beta/(4\beta-\gamma_k)$-averaged when $\gamma<2\beta$. Furthermore, since $\partial f(x)$ is scalar-valued, then for any $k\in\mathbb{N}$ and $x\in\mathbb{X}$,

$$x\in(\partial f+\partial g)^{-1}(0) \iff x-\gamma_k\nabla f(x)\in x+\gamma_k\partial g(x) \iff x\in\mathrm{Fix}(T_1T_2).$$

For the error term ϵ_k, we have that

$$\begin{aligned}
\|\epsilon_k\| =&\|(\mathcal{J}_{\gamma_k\partial g}(x^k-\gamma_k(\nabla f(x^k)+\epsilon_{f,k}))+\epsilon_{g,k}-\mathcal{J}_{\gamma_k\partial g}(x^k-\gamma_k\nabla f(x^k)))\|\\
\le&\|\mathcal{J}_{\gamma_k\partial g}(x^k-\gamma_k(\nabla f(x^k)+\epsilon_{f,k}))-\mathcal{J}_{\gamma_k\partial g}(x^k-\gamma_k\nabla f(x^k))\|+\|\epsilon_{g,k}\|\\
\le&\|x^k-\gamma_k(\nabla f(x^k)+\epsilon_{f,k})-(I-\gamma_k\nabla f)x^k\|+\|\epsilon_{g,k}\|\\
\le&\|\epsilon_{f,k}\|+\|\epsilon_{g,k}\|.
\end{aligned}$$

It is easy to verify that the conditions of Theorems 12–13 are satisfied. The desired result then follows.

□

A.10 Proof of Fact 6.2

Proof:

$$\begin{aligned}\|\mathrm{refl}_{\lambda f}(x)-\mathrm{refl}_{\lambda f}(y)\|_2^2 &= \|2\mathcal{J}_{\lambda f}(x)-2\mathcal{J}_{\lambda f}(y)-(x-y)\|_2^2\\ &= 4\|\mathcal{J}_{\lambda\partial f}(x)-\mathcal{J}_{\lambda\partial f}(y)\|_2^2-4\langle\mathcal{J}_{\lambda\partial f}(x)-\mathcal{J}_{\lambda f}(y),x-y\rangle+\|x-y\|_2^2\\ &\leq 4\|\mathcal{J}_{\lambda\partial f}(x)-\mathcal{J}_{\lambda\partial f}(y)\|_2^2-4\|\mathcal{J}_{\lambda\partial f}(x)-\mathcal{J}_{\lambda\partial f}(y)\|_2^2+\|x-y\|_2^2\\ &= \|x-y\|_2^2.\end{aligned}$$

□

A.11 Proof of Corollary 4

Proof: Suppose that $\lambda_k' = \lambda_k/2$ and $\epsilon_k = 2\epsilon_{f,k} + \mathrm{refl}_{\gamma_k\partial f}(\mathrm{refl}_{\gamma_k\partial g}\, z^k + 2\epsilon_{g,k})\, \mathrm{refl}_{\gamma_k\partial f}(\mathrm{refl}_{\gamma_k\partial g}\, z^k)$. It follows from Fact 6.2, Lemma **??**, and straightforward manipulation that we can rewrite the iteration as a fixed-point iteration with the reflection operator, which is nonexpansive. For the error term, we have that

$$\begin{aligned}\sum_{k\in\mathbb{N}}\lambda_k\|\epsilon_k\| &\leq \sum_{k\in\mathbb{N}}\lambda_k\|\epsilon_{f,k}\|+\sum_{k\in\mathbb{N}}\lambda_k\|\mathrm{refl}_{\gamma_k\partial f}(\mathrm{refl}_{\gamma_k\partial g}\, z^k+2\epsilon_{g,k})-\mathrm{refl}_{\gamma_k\partial f}(\mathrm{refl}_{\gamma_k\partial f}\, z^k)\|/2\\ &\leq \sum_{k\in\mathbb{N}}\lambda_k(\|\epsilon_{f,k}\|+\|\epsilon_{g,k}\|)<\infty\end{aligned}$$

The result then follows from Theorems 12–13.

□

A.12 Proof of Theorems 8–10

Lemma 3 (Asi & Duchi, 2019 [2]) *Let A_k, B_k, C_k, and D_k be non-negative random variables adapted to the filtration $\mathcal{F}_k$ and satisfying $\mathrm{E}[A_{k+1} \mid \mathcal{F}_k] \leq (1+B_k)A_k + C_k - D_k$. Then for the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, there is a random variable $A_\infty < \infty$ such that $A_k \to A_\infty$ almost surely and $\sum_k D_k < \infty$.*

Lemma 4 (David & Yin, 2016 [20]) *Suppose that the nonnegative scalar sequences $(\lambda_j)_{j\geq 0}$ and $(a_j)_{j\geq 0}$ satisfy $\sum_{i=0}^{\infty}\lambda_i a_i < \infty$. Let $\Lambda_k = \sum_{i=0}^{k}\lambda_i$ for $k \geq 0$ and let $(e_j)_{j\geq 0}$ be a sequence of scalars. Suppose that $a_{k+1} \leq a_k + e_k$ for all k and that $\sum_{i=0}^{\infty}\Lambda_i e_i < \infty$. Then*

$$a_k \leq \frac{1}{\Lambda_k}\Big(\sum_{i=0}^{\infty}\lambda_i a_i + \sum_{i=0}^{\infty}\Lambda_i e_i\Big)$$

and $a_k = o((\Lambda_k - \Lambda_{\lceil k/2\rceil})^{-1})$.

Lemma 5 (David & Yin, 2016 [20]) *Let $T : \mathcal{H} \to \mathcal{H}$ be a nonexpansive operator. Then for all $\lambda \in (0,1]$ and $(x,y) \in \mathcal{H}\times\mathcal{H}$, the averaged operator T_λ satisfies*

$$\|T_\lambda x - T_\lambda y\|^2 \leq \|x-y\|^2 - \frac{1-\lambda}{\lambda}\big\|(I_{\mathcal{H}} - T_\lambda)x - (I_{\mathcal{H}} - T_\lambda)y\big\|^2.$$

Lemma 6 *For any operator $T : \mathcal{H} \to \mathcal{H}$, the reflection operator $\mathrm{refl}_{\gamma T}$ of T is defined as $\mathrm{refl}_{\gamma T} = (2\mathcal{J}_{\gamma T} - I_{\mathcal{H}})$. In addition, $\mathrm{refl}_{\gamma T}$ is a nonexpansive operator for any $\gamma > 0$.*

Proof: Since $\mathcal{J}_{\gamma T}$ is a (1/2)-averaged operator, $\|\mathcal{J}_{\gamma T}(x) - \mathcal{J}_{\gamma T}(y)\| < \|x - y\|$ for any $x, y \in \mathcal{H}$. Thus,

$$\begin{aligned}
\|\mathrm{refl}_{\gamma T}(x) - \mathrm{refl}_{\gamma T}(y)\|^2 &= \|2\mathcal{J}_{\gamma T}(x) - 2\mathcal{J}_{\gamma T}(y) - (x-y)\|^2 \\
&= 4\|\mathcal{J}_{\gamma T}(x) - \mathcal{J}_{\gamma T}(y)\|^2 - 4\langle \mathcal{J}_{\gamma T}(x) - \mathcal{J}_{\gamma T}(y), x - y\rangle + \|x-y\|^2 \\
&\leq 4\|\mathcal{J}_{\gamma T}(x) - \mathcal{J}_{\lambda T}(y)\|^2 - 4\|\mathcal{J}_{\lambda T}(x) - \mathcal{J}_{\lambda T}(y)\|^2 + \|x-y\|^2 \\
&= \|x-y\|^2.
\end{aligned}$$

□

A.12.1 Proof of Theorem 8

Proof: (i) Since $(T_n)_{n=1}^N$ is a sequence of averaged nonexpansive operators, by Lemma 5,

$$\Big\| \prod_{n=1}^N T_n(x) - \prod_{n=1}^N T_n(y) \Big\| \leq \Big\| \prod_{n=1}^{N-1} T_n(x) - \prod_{n=1}^{N-1} T_n(y) \Big\| \leq \cdots \leq \|x-y\|.$$

Thus,

$$\begin{aligned}
&\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] = \Big\|(1-\lambda_k)(x^k - x^*) + \lambda_k\big(\prod_{n=1}^N T_n x^k - x^* + \epsilon_k\big)\Big\|^2 \\
&=(1-\lambda_k)\|x^k - x^*\|^2 + \lambda_k\Big\|\prod_{n=1}^N T_n x^k - x^* + \epsilon_k\Big\|^2 - \lambda_k(1-\lambda_k)\Big\|\prod_{n=1}^N T_n x^k - x^k + \epsilon_k\Big\|^2 \\
&=(1-\lambda_k)\|x^k - x^*\|^2 + \lambda_k\Big(\Big\|\prod_{n=1}^N T_n x^k - x^*\Big\|^2 + 2\big\langle \prod_{n=1}^N T_n x^k - x^*, \epsilon_k\big\rangle + \|\epsilon_k\|^2\Big) \\
&\quad - \lambda_k(1-\lambda_k)\Big(\Big\|\prod_{n=1}^N T_n x^k - x^k\Big\|^2 + 2\big\langle \prod_{n=1}^N T_n x^k - x^k, \epsilon_k\big\rangle + \|\epsilon_k\|^2\Big) \\
&\leq\|x^k - x^*\|^2 - \tau_k\Big\|\prod_{n=1}^N T_n x^k - x^k\Big\|^2 + \Big(\lambda_k^2\|\epsilon_k\|^2 + 2\lambda_k^2\Big\|\prod_{n=1}^N T_n x^k - x^k\Big\|\|\epsilon_k\| + 2\lambda_k\|x^k - x^*\|\|\epsilon_k\|\Big) \\
&=\|x^k - x^*\|^2 - \tau_k\Big\|\prod_{n=1}^N T_n x^k - x^k\Big\|^2 + \tilde{\xi}_k, \qquad \text{(A.1)}
\end{aligned}$$

where the last equality serves to define $\tilde{\xi}_k$.

Next, we show that $\|\prod_{n=1}^N T_n x^k - x^k\|^2$ is uniformly bounded over k. Since every averaged nonexpansive operator can be written as a linear form of a nonexpansive operator, there exists a nonexpansive operator T' and $\lambda' \in (0,1)$ such that $\prod_{n=1}^N T_n = (1-\lambda')I + \lambda' T'$. Therefore,

$$\begin{aligned}
\Big\|\prod_{n=1}^N T_n x^{k+1} - x^{k+1}\Big\|^2 &= \|x^{k+1} + \lambda'(T'x^{k+1} - x^{k+1}) - x^{k+1}\|^2 = (\lambda')^2\|T'x^{k+1} - x^{k+1}\|^2 \\
&\leq \|T'x^{k+1} - x^{k+1}\|^2.
\end{aligned}$$

Moreover,

$$\begin{aligned}\|T'x^{k+1} - x^{k+1}\|^2 =& \|T'x^k - x^k\|^2 + \|T'x^{k+1} - x^{k+1} - T'x^k + x^k\|^2 \\ &+ 2\langle T'x^{k+1} - x^{k+1} - T'x^k + x^k, T'x^k - x^k\rangle \\ =& \|T'x^k - x^k\|^2 + \|(T'x^{k+1} - T'x^k) - (x^{k+1} - x^k)\|^2 \\ &+ \frac{2}{\lambda'\lambda_k}\langle (T'x^{k+1} - T'x^k) - (x^{k+1} - x^k), x^{k+1} - x^k - \lambda_k\epsilon_k\rangle. \end{aligned} \tag{A.2}$$

Since T' is a nonexpansive operator, $\|T'x^{k+1} - T'x^k\|^2 \leq \|x^{k+1} - x^k\|^2$, and so

$$\begin{aligned}2\langle T'x^{k+1} - T'x^k - (x^{k+1} - x^k), x^{k+1} - x^k\rangle =& \|T'x^{k+1} - T'x^k\|^2 - \|x^{k+1} - x^k\|^2 \\ &- 2\|x^{k+1} - x^k\|^2 - \|T'x^{k+1} - T'x^k - (x^{k+1} - x^k)\|^2 \\ \leq& - \|T'x^{k+1} - T'x^k - (x^{k+1} - x^k)\|^2. \end{aligned} \tag{A.3}$$

Then by (A.2)–(A.3),

$$\begin{aligned}\|T'x^{k+1} - x^{k+1}\|^2 \leq& \|T'x^k - x^k\|^2 - \frac{1 - \lambda_k\lambda'}{\lambda_k\lambda'}\|(T'x^{k+1} - T'x^k) - (x^{k+1} - x^k)\|^2 \\ &- \frac{2}{\lambda'}\langle (T'x^{k+1} - T'x^k) - (x^{k+1} - x^k), \epsilon_k\rangle \\ =& \|T'x^k - x^k\|^2 - \frac{1 - \lambda_k\lambda'}{\lambda_k\lambda'}\left\|(T'x^{k+1} - T'x^k) - (x^{k+1} - x^k) - \frac{\lambda_k}{1 - \lambda_k\lambda'}\epsilon_k\right\|^2 \\ &+ \frac{\lambda_k}{\lambda'(1 - \lambda'\lambda_k)}\|\epsilon_k\|^2 \\ \leq& \|T'x^k - x^k\|^2 + \frac{\lambda_k^2}{\lambda_k\lambda'(1 - \lambda'\lambda_k)}\|\epsilon_k\|^2 \\ =& \|T'x^k - x^k\|^2 + \tilde{\epsilon}_k, \end{aligned} \tag{A.4}$$

where the last equality serves to define $\tilde{\epsilon}_k$.

Under the condition that $\sum_{k=1}^{\infty} \lambda_k^2 \mathbb{E}[\|\epsilon_k\|^2] < \infty$, we have that $\sup_k \|\prod_{n=1}^{N} T_n x^k - x^k\|^2 < \infty$ almost surely. On the other hand,

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - x^*\| \mid \mathcal{F}_k] &= \|(1 - \lambda_k)(x^k - x^*) + \lambda_k(Tx^k - x^*) + \lambda_k\epsilon_k\| \\ &\leq \|x^k - x^*\| + \|\lambda_k\epsilon_k\| \\ &\leq \|x^0 - x^*\| + \sum_{l=0}^{k}\|\lambda_l\epsilon_l\|. \end{aligned} \tag{A.5}$$

Therefore, under the condition that $\sum_{l=1}^{\infty} \mathbb{E}[\|\lambda_l\epsilon_l\|] < \infty$, it follows (with probability one) that

$$\sup_k \|x^k - x^*\| < \infty. \tag{A.6}$$

Additionally, $\sum_{k=1}^{\infty} \mathbb{E}[\|\tilde{\xi}_k\|] < \infty$ under the condition that $\sum_{k=1}^{\infty} \mathbb{E}[\|\lambda_k\epsilon_k\|]^2 < \infty$. Let $\tilde{\mathcal{X}}^* = \{x \in \mathcal{H} \mid x = \prod_{n=1}^{N} T_n\}$. By applying Lemma 3 with $A_k = \|x^k - x^*\|^2$, $B_k = 0$, $C_k = \tilde{\xi}_k$, and $D_k = \tau_k\|\prod_{n=1}^{N} T_n x^k - x^k\|^2$ in inequality (A.1), it follows that, for any $x^* \in \mathcal{X}^*$, there is a random variable $V(x^*) < \infty$ such that $\|x^k - x^*\| \to V(x^*)$ almost surely and $\sum_k D_k < \infty$. This implies that $\lim_{k\to\infty}\|Tx^k - x^k\|^2 = 0$ almost surely. Therefore, $x^k \in \mathcal{X}^*$ as $k \to \infty$ and $\mathrm{dist}(x^k, \mathcal{X}^*) \to 0$ almost surely.

We next show that $V(x^*) = 0$ for some $x^* \in \mathcal{X}^*$. Let $\mathbb{B} = \{x \in \mathcal{H} \mid \|x\| \leq 1\}$. Since $V(x^*) < \infty$, there exists a $c < \infty$, such that $x^k \in c\mathbb{B}$ for all k. As $\operatorname{dist}(x^k, \mathcal{X}^*) \to 0$ almost surely, $c\mathbb{B} \cap \mathcal{X}^* \neq \emptyset$ by the compactness of $\mathbb{B}$. Let $x^* \in c\mathbb{B} \cap \mathcal{X}^*$. The projection P_x of any $x \in c\mathbb{B}$ onto $\mathcal{X}^*$ satisfies $\|P_x - x\| \leq \|x^* - x\| \leq 2\operatorname{dist}(x, \mathcal{X}^*)$ and $\|P_x\| \leq 3\operatorname{dist}(x, \mathcal{X}^*)$. Define the set $S = 3\operatorname{dist}(x, \mathcal{X}^*)\mathbb{B}$. It follows that $\operatorname{dist}(x^k, \mathcal{X}^*) = \operatorname{dist}(x^k, \mathcal{X}^* \cap S)$ for all k. Now fix $\omega > 0$ and let $\{x_i^*\}_{i=1}^N$ be a ω-net of $\mathcal{X}^* \cap S$ with $N < \infty$. Since $x^k \in S$ for all k,

$$\min_{i \in [N]} \|x^k - x_i^*\| - \omega \leq \operatorname{dist}(x^k, \mathcal{X}^* \cap S) = \operatorname{dist}(x^k, \mathcal{X}^*) \to 0$$

almost surely, and $\min_{i \in [N]} \|x^k - x_i^*\| \to \min_{i \in [N]} V(x_i^*)$ almost surely. So for any $\omega > 0$, there exists $x_\omega \in \mathcal{X}^* \cap S$ such that $V(x_\omega) \leq \omega$ and so $\inf_{x \in \mathcal{X}^* \cap S} V(x) = 0$. Thus, x^k converges to this $x^* \in \mathcal{X}^*$.

(ii) Since $\prod_{n=1}^N T_n$ is an averaged nonexpansive operator,

$$\begin{aligned}
\Big\| \prod_{n=1}^N T_n x^k - x^* \Big\|^2 =& \Big\| \prod_{n=1}^N T_n x^k - \prod_{n=1}^N T_n x^* \Big\|^2 \\
\leq& \Big\| \prod_{n=2}^N T_n x^k - \prod_{n=2}^N T_n x^* \Big\|^2 - \frac{1 - \lambda_{1,k}}{\lambda_{1,k}} \Big\| (\operatorname{Id} - T_1) \prod_{n=2}^N T_n x^k - (\operatorname{Id} - T_1) \prod_{n=2}^N T_n x^* \Big\|^2 \\
\leq& \Big\| \prod_{n=3}^N T_n x^k - \prod_{n=3}^N T_n x^* \Big\|^2 - \frac{1 - \lambda_{2,k}}{\lambda_{2,k}} \Big\| (\operatorname{Id} - T_2) \prod_{n=3}^N T_n x^k - (\operatorname{Id} - T_2) \prod_{n=3}^N T_n x^* \Big\|^2 \\
&- \frac{1 - \lambda_{1,k}}{\lambda_{1,k}} \Big\| (\operatorname{Id} - T_1) \prod_{n=2}^N T_n x^k - (\operatorname{Id} - T_1) \prod_{n=2}^N T_n x^* \Big\|^2 \\
\leq& \|x^k - x^*\|^2 - \sum_{i=1}^N \frac{1 - \lambda_{i,k}}{\lambda_{i,k}} \Big\| (\operatorname{Id} - T_i) \prod_{n=i+1}^N T_n x^k - (\operatorname{Id} - T_i) \prod_{n=i+1}^N T_n x^* \Big\|^2. \qquad \text{(A.7)}
\end{aligned}$$

By (A.1),

$$\begin{aligned}
\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq& \|x^k - x^*\|^2 + \tilde{\xi}_k \\
\leq& \Big\| (1 - \lambda_{k-1})(x^{k-1} - x^*) + \lambda_{k-1}\big(\prod_{n=1}^N T_n x^{k-1} - x^* \big) \Big\|^2 + \|\lambda_{k-1}\epsilon_{k-1}\|^2 + \tilde{\xi}_k \\
&+ 2\langle T_\lambda x^{k-1} - x^*, \lambda_{k-1}\epsilon_{k-1} \rangle \\
=& (1 - \lambda_{k-1})\|x^{k-1} - x^*\|^2 + \lambda_{k-1} \Big\| \prod_{n=1}^N T_n x^{k-1} - x^* \Big\|^2 \\
&- \lambda_{k-1}(1 - \lambda_{k-1}) \Big\| \prod_{n=1}^N T_n x^{k-1} - x x^{k-1} \Big\|^2 \\
&+ 2\langle T_\lambda x^{k-1} - x^*, \lambda_{k-1}\epsilon_{k-1} \rangle + \|\lambda_{k-1}\epsilon_{k-1}\|^2 + \tilde{\xi}_k \\
\leq& (1 - \lambda_{k-1})\|x^{k-1} - x^*\|^2 + \lambda_{k-1} \Big\| \prod_{n=1}^N T_n x^{k-1} - x^* \Big\|^2 \\
&- \lambda_{k-1}(1 - \lambda_{k-1}) \Big\| \prod_{n=1}^N T_n x^{k-1} - x^{k-1} \Big\|^2 \\
&+ 2\|x^0 - x^*\| \|\lambda_{k-1}\epsilon_{k-1}\| + \|\lambda_{k-1}\epsilon_{k-1}\|^2 + \tilde{\xi}_k, \qquad \text{(A.8)}
\end{aligned}$$

where $T_\lambda x^{k-1} = (1 - \lambda_{k-1})(x^{k-1} - x^*) + \lambda_{k-1}(\prod_{n=1}^N T_n x^{k-1} - x^*) + x^*$. Together, (A.7) and (A.8) give that

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-x^*\|^2 \mid \mathcal{F}_k] \le &\|x^{k-1}-x^*\|^2\\ &-\lambda_{k-1}\sum_{i=1}^{N}\frac{1-\lambda_{i,k-1}}{\lambda_{i,k1}}\Big\|(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^{k-1}-(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^*\Big\|^2\\ &-\lambda_{k-1}(1-\lambda_{k-1})\Big\|\prod_{n=1}^{N}T_n x^{k-1}-x^{k-1}\Big\|^2\\ &+2\|x^0-x^*\|\|\lambda_{k-1}\epsilon_{k-1}\|+\|\lambda_{k-1}\epsilon_{k-1}\|^2+\tilde{\xi}_k.\end{aligned}$$

Summing the above over $k=1,2,\dots$ yields that

$$\begin{aligned}&\sum_{k=1}^{\infty}\Big\{\tau_{k-1}\mathbb{E}\Big[\Big\|\prod_{n=1}^{N}T_n x^{k-1}-x^{k-1}\Big\|^2\Big]\\ &+\lambda_{k-1}\sum_{i=1}^{N}\frac{1-\lambda_{i,k-1}}{\lambda_{i,k-1}}\mathbb{E}\Big[\Big\|(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^{k-1}-(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^*\Big\|^2\Big]\Big\}\\ &\le \|x^0-x^*\|^2+\mathbb{E}[\|x^1-x^*\|^2]+2\|x^0-x^*\|\sum_{k=1}^{\infty}\lambda_{k-1}\mathbb{E}[\|\epsilon_{k-1}\|]+\sum_{k=1}^{\infty}\mathbb{E}[\|\lambda_{k-1}\epsilon_{k-1}\|^2]+\sum_{k=1}^{\infty}\mathbb{E}[\tilde{\xi}_k]\\ &<\infty\end{aligned}$$

and so $\sum_{k=1}^{\infty}\tau_k\mathbb{E}[\|\prod_{n=1}^{N}T_n x^k-x^k\|^2]<\infty$ and

$$\sum_{k=1}^{\infty}\lambda_k\sum_{i=1}^{N}\frac{1-\lambda_{i,k}}{\lambda_{i,k}}\mathbb{E}\Big[\Big\|(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^k-(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^*\Big\|^2\Big]<\infty.$$

This indicates that, for all k,

$$\sum_{i=1}^{N}\mathbb{E}\Big[\Big\|(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^k-(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^*\Big\|^2\Big]<\infty$$

which in turn implies that

$$\lim_{k\to\infty}\mathbb{E}\Big[\Big\|(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^k-(\operatorname{Id}-T_i)\prod_{n=i+1}^{N}T_n x^*\Big\|^2\Big]=0,$$

$$\Big(\prod_{n=i}^{N}T_n x^k-\prod_{n=i+1}^{N}T_n x^k\Big)\to\Big(\prod_{n=i}^{N}T_n x^*-\prod_{n=i+1}^{N}T_n x^*\Big)$$

almost surely, and that $\prod_{n=1}^{N}T_n x^k-x^k\to 0$ almost surely. In other words, e_k converges almost surely to 0.

Since $\Lambda_k=\sum_{i=0}^{k}\lambda_i\le(k+1)$, then under the condition that $\sum_{i=1}^{\infty}(i+1)\mathbb{E}[\tilde{\epsilon}_k]<\infty$, the previous result, and (A.10),

$$\mathbb{E}\Big[\Big\|\prod_{n=1}^{N}T_n x^k-x^k\Big\|^2\Big]=o((k+1)^{-1})$$

by Lemma 4. □

A.12.2 Proof of Theorem 9

The proof is similar to that of Theorem 8.

A.12.3 Proof of Theorem 10

Proof: (i) First, we have that

$$\begin{aligned}\mathbb{E}[\|x^{k+1}-x^*\| \mid \mathcal{F}_k] &= \|T_{\lambda_k}(x^k)-T_{\lambda_k}(x^*)+\lambda_k\epsilon_k\| \\ &\le \|x^k-x^*\|+\|\lambda_k\epsilon_k\| \\ &\le \|x^0-x^*\|+\sum_{l=0}^{k}\|\lambda_l\epsilon_l\|.\end{aligned}$$

Therefore, $\mathbb{E}[\|x^{k+1}-x^*\|] \le D_0+\sum_{l=0}^{k}\mathbb{E}[\|\lambda_l\epsilon_l\|]$, from which follows that

$$\begin{aligned}\mathbb{E}[\|\bar{e}^k\|] &= \mathbb{E}\Big[\Big\|\frac{1}{\Lambda_k}\sum_{l=1}^{k}\lambda_l e^l\Big\|\Big] \\ &= \frac{1}{\Lambda_k}\mathbb{E}\Big[\Big\|\sum_{l=1}^{k}(x^{l+1}-x^l)-\lambda_l\epsilon_l\Big\|\Big] \\ &= \frac{1}{\Lambda_k}\mathbb{E}\Big[\Big\|x^{k+1}-x^1-\sum_{l=1}^{k}\lambda_l\epsilon_l\Big\|\Big] \\ &\le \frac{2(D_0+\sum_{l=1}^{k}\mathbb{E}[\|\lambda_l\epsilon_l\|])}{\Lambda_k}.\end{aligned}$$

(ii) Under the assumption that $\sum_{k=1}^{\infty}\lambda_k\mathbb{E}[\|\epsilon_k\|]<\infty$ and $\underline{\lambda}>0$, both $\lim_{k\to\infty}\mathbb{E}[\bar{e}^k]=0$ and $\|\bar{e}^k\| = O_p([\underline{\lambda}(k+1)]^{-1})$ follow from the previous result. □

A.13 Proof of Theorem 13

Let $\mathbb{P}$ and $\mathbb{Q}$ denote two probability measures that are absolutely continuous with respect to a third probability measure μ defined on a set Ξ. The square of the Hellinger distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$d_{\text{hel}}(\mathbb{P},\mathbb{Q})^2 = \int_{\Xi}\Big(\sqrt{\frac{p(\xi)}{q(\xi)}}-1\Big)^2 q(\xi)\mathrm{d}\mu(\xi) = \int_{\Xi}(\sqrt{p(\xi)}-\sqrt{q(\xi)})^2\mathrm{d}\mu(\xi),$$

where p and q are the densities of $\mathbb{P}$ and $\mathbb{Q}$, respectively. It is well known [1] that, for any probability distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$d_{\text{hel}}(\mathbb{P},\mathbb{Q})^2 \le 2d_{\text{TV}}(\mathbb{P},\mathbb{Q}) \le 2d_{\text{hel}}(\mathbb{P},\mathbb{Q}). \tag{A.9}$$

We now proceed with a proof of Theorem 13.

Proof: Let p be the density of the distribution $\mathbb{P}$. For simplicity, we suppress the subscript k in λ_k from here on. Since $T_\lambda(x^*;\xi) = x^*$ for any $\xi \in \Xi$, we have that

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\sum_{t=1}^{\top}(x^t - x^*) \mid \mathcal{F}_{t-1}\Big\|\Big] \le & \mathbb{E}\Big[\Big\|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi_t) - T_\lambda(x^*;\xi_t)] \mid \mathcal{F}_{t-1}\Big\|\Big] + \mathbb{E}[\Big\|\sum_{t=1}^{\top}\lambda_{t-1}\epsilon_{t-1}\Big\|\Big] \\
= & \int_\Xi \Big\|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| \mathrm{d}P^t_{t-1}(\xi) + \mathbb{E}\Big[\Big\|\sum_{t=1}^{\top}\lambda_{t-1}\epsilon_{t-1}\Big\|\Big] \\
\le & \int_\Xi \Big\|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| \mathrm{d}(P^t_{t-1} - \mathbb{P})(\xi) \\
& + \int_\Xi \Big\|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| \mathrm{d}\mathbb{P}(\xi) + \mathbb{E}\Big[\Big\|\sum_{t=1}^{\top}\lambda_{t-1}\epsilon_{t-1}\Big\|\Big] \\
\le & \int_\Xi \Big\|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| \mathrm{d}\mathbb{P}(\xi) + R_{T-1}\phi(1) + \mathbb{E}\Big[\Big\|\sum_{t=1}^{\top}\lambda_{t-1}\epsilon_{t-1}\Big\|\Big].
\end{aligned}$$

For any $\tau > 0$,

$$\begin{aligned}
\int_\Xi \|\sum_{t=1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\| \mathrm{d}\mathbb{P}(\xi) \le & \int_\Xi \Big\|\sum_{t=1}^{T-\tau}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\|(p - p^{t+\tau}_{[t-1]})(\xi)\mathrm{d}\xi \\
& + \int_\Xi \Big\|\sum_{t=1}^{T-\tau}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^{t-1+\tau};\xi)]\Big\| p^{t+\tau}_{[t-1]}(\xi)\mathrm{d}\xi \\
& + \int_\Xi \Big\|\sum_{t=1}^{T-\tau}[T_\lambda(x^{t-1+\tau};\xi) - T_\lambda(x^*;\xi)]\Big\| p^{t+\tau}_{[t-1]}(\xi)\mathrm{d}\xi \\
& + \int_\Xi \Big\|\sum_{t=T-\tau+1}^{\top}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| p(\xi)\mathrm{d}\xi \\
= & B_1 + B_2 + B_3 + B_4.
\end{aligned}$$

where the last equality serves to define B_1, B_2, B_3, and B_4. Regarding B_1,

$$\begin{aligned}
B_1 = & \int_\Xi \Big\|\sum_{t=1}^{T-\tau}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\| \big[\sqrt{p(\xi)} + \sqrt{p^{t+\tau}_{[t-1]}(\xi)}\big]\big[\sqrt{p(\xi)} - \sqrt{p^{t+\tau}_{[t-1]}(\xi)}\big]\mathrm{d}\xi \\
\le & \Big\{2\int_\Xi \Big\|\sum_{t=1}^{T-\tau}[T_\lambda(x^{t-1};\xi) - T_\lambda(x^*;\xi)]\Big\|^2 [p(\xi) + p^{t+\tau}_{[t-1]}(\xi)]\mathrm{d}\xi \int [\sqrt{p}(\xi) - \sqrt{p^{t+\tau}_{[t-1]}(\xi)}]^2\mathrm{d}\xi\Big\}^{1/2} \\
\le & \sqrt{2}\sum_{t=1}^{T-\tau}\sqrt{2\mathbb{E}[\|x^{t-1} - x^*\|^2]}\, d_{\mathrm{hel}}(\mathbb{P}, P^{t+\tau}_{[t-1]}) \\
\le & 2\sqrt{2}(T-\tau)r\sqrt{\phi(\tau+1)}.
\end{aligned}$$

The second inequality above holds by the contraction property of the averaged operator T_λ, while the third follows from (A.9). Now regarding B_2,

$$
\begin{aligned}
B_2 &\le \sum_{t=1}^{T-\tau} \mathbb{E}[\|T_\lambda(x^{t-1};\xi_{t+\tau}) - T_\lambda(x^{t-1+\tau};\xi_{t+\tau})\| \mid \mathcal{F}_{t-1}] \\
&\le \sum_{t=1}^{T-\tau}\sum_{l=0}^{\tau-1} \mathbb{E}[\|T_\lambda(x^{t+l-1};\xi_{t+\tau}) - T_\lambda(x^{t+l};\xi_{t+\tau})\| \mid \mathcal{F}_{t-1}] \\
&= \sum_{t=1}^{T-\tau}\sum_{l=t-1}^{t+\tau-2} \mathbb{E}[\|T_\lambda(x^{l};\xi_{t+\tau}) - T_\lambda(x^{l+1};\xi_{t+\tau})\| \mid \mathcal{F}_{t-1}] \\
&\le \sum_{t=1}^{T-\tau}\sum_{l=t-1}^{t+\tau-2} \mathbb{E}[\|x^{l} - x^{l+1}\| \mid \mathcal{F}_{t-1}] \\
&\le \tau \sum_{t=1}^{T-\tau} \kappa(t-1).
\end{aligned}
$$

Regarding B_3,

$$
\begin{aligned}
B_3 &= \mathbb{E}\Big[\Big\| \sum_{t=\tau}^{T-1} [T_\lambda(x^t;\xi_{t+1}) - T_\lambda(x^*;\xi_{t+1})]\Big\| \mid \mathcal{F}_{t-\tau}\Big] \\
&= \mathbb{E}\Big\{\mathbb{E}\Big[\Big\| \sum_{t=\tau}^{T-1} [T_\lambda(x^t;\xi_{t+1}) - T_\lambda(x^*;\xi_{t+1})]\Big\| \mid \mathcal{F}_{t}\Big] \mid \mathcal{F}_{t-\tau}\Big\} \\
&\le \sum_{t=\tau}^{T-1} \mathbb{E}[\|x^t - x^*\| \mid \mathcal{F}_{t-\tau}] \\
&\le \mathbb{E}[R_{T-1} - R_{\tau-1}].
\end{aligned}
$$

Lastly, regarding B_4,

$$
B_4 \le \sum_{t=T-\tau+1}^{\top} \mathbb{E}[\|x^{t-1} - x^*\|] \le \tau r.
$$

The result holds by the above bounds on B_1, B_2, B_3, and B_4. □

A.14 Proof of Theorem 14

Proof: Recall that $x^k - x^{k+1} = \gamma_k \widetilde{\nabla} g(x^{k+1}) + \gamma_k \nabla f(x^k) \in \gamma_k \partial g(x^{k+1}) + \gamma_k \nabla f(x^k)$ for all $k \ge 0$. For simplicity, we assume a constant step size $\gamma = \gamma_k$ for all k. By the joint descent theorem, for all $x \in \mathrm{dom}(f)$,

$$
\begin{aligned}
&\mathbb{E}^{\mathbb{P}}[f(x^k;\xi) + g(x^k;\xi)] - [f(x^\star;\xi) + g(x^\star;\xi)] \\
&\qquad \le \mathbb{E}[\langle x^k - x^\star, \nabla f(x^{k-1};\xi) + \partial g(x^k;\xi)\rangle] + \frac{1}{2\beta}\mathbb{E}[\|x^k - x^{k-1}\|^2] \\
&\qquad \le \frac{1}{\gamma}\mathbb{E}[\langle x^k - x^\star, (x^{k-1} - x^k)\rangle] + \frac{1}{2\beta}\mathbb{E}[\|x^k - x^{k-1}\|^2] \\
&\qquad = \frac{1}{2\gamma}\mathbb{E}[\|x^k - x^\star\|^2 - \|x^{k-1} - x^\star\|^2] + \Big(\frac{1}{2\beta} - \frac{1}{2\gamma}\Big)\mathbb{E}[\|x^{k-1} - x^k\|^2].
\end{aligned}
$$

Additionally,

$$\begin{aligned}
\|T(x^{k-1};\xi_k) - x^{k-1}\|^2 =& \|T(x^{k-2};\xi_{k-1}) - x^{k-2}\|^2 + \|T(x^{k-1};\xi_k) - x^{k-1} - T(x^{k-2};\xi_{k-1}) + x^{k-2}\|^2 \\
&+ 2\langle T(x^{k-1};\xi_k) - x^{k-1} - T(x^{k-2},\xi_{k-1}) + x^{k-2}, T(x^{k-2};\xi_{k-1}) - x^{k-2}\rangle \\
=& \|T(x^{k-2};\xi_{k-1}) - x^{k-2}\|^2 + \|(T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2})\|^2 \\
&+ \frac{2}{\lambda_{k-2}}\langle (T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2}), x^{k-1} - x^{k-2} - \lambda_{k-2}\epsilon_{k-2}\rangle.
\end{aligned}$$

and

$$\begin{aligned}
2\langle (T(x^{k-1};\xi_k)-&T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2}), x^{k-1} - x^{k-2}\rangle \\
&= \|T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})\|^2 - \|x^{k-1} - x^{k-2}\|^2 - 2\|x^{k-1} - x^{k-2}\|^2 \\
&\quad - \|(T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2})\|^2 \\
&\leq - \|(T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2})\|^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\|T(x^{k-1};\xi_k)-&x^{k-1}\|^2 \\
\leq& \|T(x^{k-2};\xi_{k-1}) - x^{k-2}\|^2 - \frac{1-\lambda_{k-2}}{\lambda_{k-2}}\|(T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2})\|^2 \\
&- 2\langle (T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2}), \epsilon_{k-2}\rangle \\
=& \|T(x^{k-2};\xi_{k-1}) - x^{k-2}\|^2 - \frac{1-\lambda_{k-2}}{\lambda_{k-2}}\|(T(x^{k-1};\xi_k) - T(x^{k-2};\xi_{k-1})) - (x^{k-1} - x^{k-2}) \\
&- \frac{\lambda_{k-2}}{1-\lambda_{k-2}}\epsilon_{k-2}\|^2 + \frac{\lambda_{k-2}}{1-\lambda_{k-2}}\|\epsilon_{k-2}\|^2 \\
\leq& \|T(x^{k-2};\xi_{k-1}) - x^{k-2}\|^2 + \frac{\lambda_{k-2}^2}{\tau_{k-2}}\|\epsilon_{k-2}\|^2 \\
\leq& 4\|x^0 - x^\star\|^2 + \sum_{l=1}^{k-2}\lambda_l^2\|\epsilon_l\|^2/\underline{\tau}.
\end{aligned} \tag{A.10}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[(f(x^k;\xi)+&g(x^k;\xi)) - (f(x^\star;\xi) + g(x^\star;\xi))] \\
\leq& \frac{1}{2\gamma}(\|x^k - x^\star\|^2 - \|x^{k-1} - x^\star\|^2) + \Big(\frac{1}{2\beta} - \frac{1}{2\gamma}\Big)\|x^{k-1} - x^k\|^2 \\
\leq& \frac{1}{2\gamma}(\|x^k - x^\star\|^2 - \|x^{k-1} - x^\star\|^2) + \Big(\frac{1}{\beta} - \frac{1}{\gamma}\Big)\lambda_{k-1}^2\|T(x^{k-1};\xi_k) - x^{k-1}\|^2 \\
&+ \Big(\frac{1}{\beta} - \frac{1}{\gamma}\Big)\|\lambda_{k-2}\epsilon_{k-2}\|^2 \\
\leq& \frac{1}{2\gamma}(\|x^k - x^\star\|^2 - \|x^{k-1} - x^\star\|^2) + \Big(\frac{4}{\beta} - \frac{4}{\gamma}\Big)\|x^0 - x^\star\|^2 + \Big(\frac{1}{\beta} - \frac{1}{\gamma}\Big)\sum_{l=1}^{k-1}\lambda_l^2\|\epsilon_l\|^2/\underline{\tau}.
\end{aligned}$$

and so

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K}[(f(x^k;\xi)+g(x^k;\xi))-(f(x^\star;\xi)+g(x^\star;\xi))]\Big]\\
&\qquad\le \frac{1}{2\gamma}\mathbb{E}[\|x^K-x^\star\|^2-\|x^0-x^\star\|^2]+\Big(\frac{4}{\beta}-\frac{4}{\gamma}\Big)Kr^2+\Big(\frac{1}{\beta}-\frac{1}{\gamma}\Big)\sum_{k=1}^{K}\sum_{l=1}^{k-1}\mathbb{E}[\|\lambda_l\epsilon_l\|^2]/\underline{\tau}\\
&\qquad\le \frac{r^2}{2\gamma}+\Big(\frac{1}{\beta}-\frac{1}{\gamma}\Big)4Kr^2+\Big(\frac{1}{\beta}-\frac{1}{\gamma}\Big)K^2\Delta^2/\underline{\tau}.
\end{aligned}$$

The result of theorem follows since $f(K^{-1}\sum_{k=1}^{K}x^k)\le K^{-1}\sum_{k=1}^{K}f(x^k)$ by Jensen's inequality. □

A.15 Proof of Theorem 15

Proof: We first have that

$$\begin{aligned}
&\mathbb{E}^{\mathbb{P}}[(f(x_f^k;\xi)+g(x_g^k;\xi))-(f(x^\star;\xi)+g(x^\star;\xi))]\\
&\qquad\le \mathbb{E}[\langle x_f^k-x^\star,\nabla f(x_f^k)+\epsilon_f\rangle+\langle x_g^k-x^\star,\partial g(x_g^k)+\epsilon_g\rangle]\\
&\qquad= \mathbb{E}[\langle x_f^k-x_g,\nabla f(x_f^k)+\epsilon_f\rangle+\langle x_g^k-x^\star,\nabla f(x_f^k)+\partial g(x_g^k)+\epsilon_g\rangle]\\
&\qquad= \frac{1}{2\gamma}\mathbb{E}[\langle x^{k+1}-x^k,\gamma(\nabla f(x_f^k)+\epsilon_f)\rangle+\langle x^\star-x_g^k,x^{k+1}-x^k\rangle]\\
&\qquad= \frac{1}{2\gamma}\mathbb{E}[\langle x^{k+1}-x^k,x^\star+(x^k-x_g^k+\gamma(\nabla f(x_f^k)+\epsilon_f))-x^k\rangle]\\
&\qquad= \frac{1}{2\gamma}\mathbb{E}[\langle x^{k+1}-x^k,x^\star+\gamma(\partial g(x_g^k)+\epsilon_g+\nabla f(x_f^k)+\epsilon_f)-x^k\rangle]\\
&\qquad= \frac{1}{2\gamma}\mathbb{E}\big[\langle x^{k+1}-x^k,x^\star-\frac{1}{2\lambda}(x^{k+1}-x^k)-x^k\rangle\big]\\
&\qquad= \frac{1}{4\gamma\lambda}\mathbb{E}\big[\|x^k-x^\star\|^2-\|x^{k+1}-x^\star\|^2+\big(1-\frac{1}{\lambda}\big)\|x^{k+1}-x^k\|^2\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}^{\mathbb{P}}\Big[\sum_{k=1}^{K}[(f(x_f^k;\xi)+g(x_g^k;\xi))-(f(x^\star;\xi)+g(x^\star;\xi))]\Big]\\
&\qquad\le\frac{1}{4\gamma\lambda}\mathbb{E}[\|x^0-x^\star\|^2-\|x^K-x^\star\|^2]+\frac{1}{2\gamma\lambda}\big(1-\frac{1}{\lambda}\big)\sum_{k=1}^{K}\|T(x^k;\xi_{k+1})-x^k\|^2\\
&\qquad\quad+\frac{1}{2\gamma\lambda}\big(1-\frac{1}{\lambda}\big)\sum_{k=1}^{K}\|\lambda\epsilon_k\|^2.
\end{aligned}$$

Replacing the proximal gradient operator in (A.10) with the Peaceman–Rachford operator yields

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K}\{[f(x_f^k;\xi)+g(x_g^k;\xi)]-[f(x^\star;\xi)+g(x^\star;\xi)]\}\Big]\\
&\qquad\le\frac{1}{4\gamma\lambda}\mathbb{E}[\|x^0-x^\star\|^2-\|x^K-x^\star\|^2]+\frac{1}{\gamma\lambda}\big(2-\frac{2}{\lambda}\big)Kr^2\\
&\qquad\quad+\frac{1}{2\gamma\lambda}\big(1-\frac{1}{\lambda}\big)\sum_{k=1}^{K}\sum_{l=1}^{k-1}\mathbb{E}[\lambda_l^2\|\epsilon_l\|^2]/\underline{\tau}+\frac{1}{2\gamma\lambda}\big(1-\frac{1}{\lambda}\big)\sum_{k=1}^{K}\mathbb{E}[\|\lambda\epsilon_k\|^2]\\
&\qquad\le\frac{r^2}{4\gamma\lambda}+\frac{1}{\gamma\lambda}\big(2-\frac{2}{\lambda}\big)Kr^2+\frac{1}{2\gamma\lambda}\big(1-\frac{1}{\lambda}\big)K^2\Delta^2/\underline{\tau}.
\end{aligned}$$

The result of theorem follows since $f(K^{-1}\sum_{k=1}^{K}x^k)\le K^{-1}\sum_{k=1}^{K}f(x^k)$ by Jensen's inequality. □