Constrained Stochastic Nonconvex Optimization with State-dependent Markov Data

Abhishek Roy* Krishnakumar Balasubramanian† Saeed Ghadimi‡

November 10, 2022

Abstract

We study stochastic optimization algorithms for constrained nonconvex stochastic optimization problems with Markovian data. In particular, we focus on the case when the transition kernel of the Markov chain is state-dependent. Such stochastic optimization problems arise in various machine learning problems including strategic classification and reinforcement learning. For this problem, we study both projection-based and projection-free algorithms. In both cases, we establish that the number of calls to the stochastic first-order oracle to obtain an appropriately defined ϵ-stationary point is of the order $\mathcal{O}(1/\epsilon^{2.5})$. In the projection-free setting we additionally establish that the number of calls to the linear minimization oracle is of order $\mathcal{O}(1/\epsilon^{5.5})$. We also empirically demonstrate the performance of our algorithm on the problem of strategic classification with neural networks.

1 Introduction

We consider the following stochastic optimization problem

$$\operatorname*{argmin}_{\theta\in\Theta} f(\theta) = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\left[F(\theta; x)\right], \tag{1}$$

where (i) the expectation is taken over the stationary distribution, π_θ, of the random vector x, (ii) F (and hence f) is a potentially non-convex function in θ, and (iii) Θ is a compact and convex constraint set. Stochastic approximation algorithms for solving problem (1), given an independent and identically distributed (`iid`) data stream $\{x_k\}_k$ drawn from a distribution π, are well-studied. Such `iid` assumptions are commonly made in various machine learning and statistical problems including empirical risk minimization [SSBD14], sparse recovery [BJMO12] and compressed sensing [FR13, Lan20]. We refer to [MB11, ABRW12, RSS12, GL13, SZ13, LZ16, ACD+19] for a partial list of non-asymptotic upper and lower bounds on the oracle complexity of widely-used stochastic approximation algorithms like the Stochastic Gradient Descent (SGD) and the Stochastic Conditional Gradient Algorithm, in the `iid` setting.

Our focus in this work is on the case when the data sequence $\{x_k\}_k$ is not necessarily `iid`. Such non-`iid` settings arise frequently in several machine learning applications including but not limited to strategic classification [HMPW16, CDP15, MDPZH20, LW22] and reinforcement learning [Bar92, GSK13, ZJM21, KMMW19, QW20]. While sample average approximation (or empirical

*Halıcıoğlu Data Science Institute, University of California, San Diego. abroy@ucdavis.edu.
†Department of Statistics, University of California, Davis. kbala@ucdavis.edu.
‡Department of Management Sciences, University of Waterloo. sghadimi@uwaterloo.ca.

risk minimization) with non-iid data is relatively well-understood (see, for example, [AD12, KM17, DDDJ19, RBE21] and references therein), a deeper understanding of the non-asymptotic oracle complexity of stochastic approximation algorithms for non-iid data is only now starting to emerge.

Towards that, we establish non-asymptotic oracle complexity results for the stochastic conditional gradient algorithm for non-convex constrained stochastic optimization with Markovian data with potentially state-*dependent* transition kernels. To establish our results, from a methodological point-of-view, we leverage the moving-average stochastic gradient estimation technique recently used in [ZSM+20, GRW20, XBG22] in the context of constrained optimization with iid data. This technique avoids having to a use a mini-batch of samples in each iteration, which turns out to be crucial in the non-iid setup we consider. From a theoretical point-of-view, we assume the so-called drift conditions, a classical assumption in Markov Chain literature [AMP05]. This ensures the existence of a solution to the Poisson equation associated with the underlying Markov chain [DMPS18] which enables one to decompose the noise present in the stochastic gradient into three components: a martingale difference sequence, a time-decaying sequence, and a telescopic sum type sequence. The key idea of our paper is to use this decomposition to construct an auxiliary sequence of iterates with a time-decaying noise variance and show that these sequence of iterates are *close* to the iterates of the original sequence produced by our algorithm. This novel technique in then used in combination with a merit-function based analysis to establish the oracle complexity results.

1.1 Motivating Example

Problems of the form in (1) arise in various important applications, e.g., strategic classification, and reinforcement learning as mentioned above. Below we illustrate the motivation of this work through the example of strategic classification with adapted best response [LW22]. In strategic classification, there is a *learner* whose task is to classify a given dataset which is collected from a set of *agents*. Given the knowledge of the classifier, the agents can distort some of their personal features, in order to get classified in a predetermined target class. This scenario arises in various applications, e.g., spam email filtering, and credit score classification. Optimizing the classifier to classify such strategically modified data where the agents modify the data iteratively can be formulated as problem (1).

Formally, let the classifier be $h(u, \theta)$ where $u \in \mathbb{R}^d$ is the feature and θ is the parameter to be optimized. $h(u; \cdot) : \Theta \to \mathbb{R}$ is potentially nonconvex. Let the loss function be logistic loss which for a sample $x := (u, y)$, where $y \in \{-1, 1\}$ denotes the corresponding class, is given by,

$$L(\theta; u, y) = \log\left(1 + \exp\left(-h(u, \theta)\right)\right) + (1 - y)h(u, \theta)/2. \tag{2}$$

We use u_S, and u_{-S} to denote the parts of feature u which are respectively strategically modifiable, and non-modifiable by the agents. Then the modified feature (the best response) u'_S reported by the agent is the solution to the following optimization problem:

$$u'_S = \operatorname*{argmax}_{u_S} \left(h(u; \theta) - c(u_S, u'_S)\right), \tag{3}$$

where $c(u_S, u'_S)$ is the cost of modifying u_S to u'_S. Let the agents iteratively learn u'_S similar to [LW22]. Note that unlike [LW22], where the authors deploy a logistic regression classifier and the closed form solution of the best response is readily known to the agents, it may not be the case in general. In that case the agents have to possibly learn the best response u'_S using some iterative optimization algorithm. For example, if the agents use Gradient Ascent then, at every iteration k, a

set $\mathcal{I}_k$ of $n_1 \leq M$ randomly chosen agents out of M agents modify their features as:

$$u_{S,i}^k = \begin{cases} u_{S,i}^{k-1} + \alpha \left(\nabla h(u_{S,i}^{k-1}; \theta_k) - \nabla c(u_{S,i}^{k-1}, u_{S,i}^0) \right) & i \in \mathcal{I}_k \\ u_{S,i}^{k-1} & i \notin \mathcal{I}_k, \end{cases} \tag{4}$$

where $\alpha > 0$ is the stepsize. With a little abuse of notation, we use $\nabla h(u_{S,i}^{k-1}; \theta)$ in (4) to denote the fact that the gradient is with respect to $u_{S,i}^{k-1}$ while $u_{-S,i}$ remains unchanged. This introduces the state-dependent Markov chain dynamics in the training data. The objective function, analogous to $f(\theta)$ in (1), is

$$\min_{\theta \in \Theta} \mathbb{E}_{\pi_\theta} \left[L(\theta; x) \right],$$

where π_θ is the stationary joint distribution of x, and Θ is a convex and compact set, e.g., sparsity inducing constraint $\|\theta\|_1 \leq R$ from some $R > 0$. The loss evaluated at a single data point x, $L(\theta; x)$, is analogous to $F(\theta; x)$ in (1). [DX20], and [LW22] study this problem theoretically and empirically respectively in an unconstrained strongly convex setting. Our results takes a step forward towards analyzing this problem in constrained nonconvex setting. We empirically show the performance of the stochastic conditional gradient algorithm on a strategic classification problem in Section 3.1.

1.2 Preliminaries and Main Contributions

Before we present our main contributions, we introduce our convergence criterion. In constrained optimization literature, most commonly used convergence criteria are: (i) *Gradient Mapping* (GM), and (ii) *Frank-Wolfe Gap* (FW-gap). The *Gradient Mapping* at a point $\bar{\theta} \in \Theta$ is defined, for some $\beta > 0$, as

$$\mathcal{G}_\Theta(\bar{\theta}, \nabla f(\bar{\theta}), \beta) := \beta \left(\bar{\theta} - \Pi_\Theta \left(\bar{\theta} - \frac{1}{\beta} \nabla f(\bar{\theta}) \right) \right), \tag{5}$$

where $\Pi_\Theta(x)$ denotes the orthogonal projection of the vector x onto the set Θ, i.e.,

$$\Pi_\Theta \left(\bar{\theta} - \frac{1}{\beta} \nabla f(\bar{\theta}) \right) = \operatorname*{argmin}_{y \in \Theta} \left\{ \langle \nabla f(\bar{\theta}), y - \bar{\theta} \rangle + \frac{\beta}{2} \|y - \theta\|_2^2 \right\}.$$

We will use $\Pi_\Theta(\bar{\theta}, \nabla f(\bar{\theta}), \beta)$ to denote $\Pi_\Theta \left(\bar{\theta} - \nabla f(\bar{\theta})/\beta \right)$ when there is no confusion. Note that when $\Theta \equiv \mathbb{R}^d$ we have $\mathcal{G}_\Theta(\bar{\theta}, \nabla f(\bar{\theta}), \beta) = \nabla f(\bar{\theta})$. In other words, for constrained optimization gradient mapping plays an analogous role of the gradient for unconstrained optimization. The gradient mapping is a frequently used measure in the literature as a convergence criterion for nonconvex constrained optimization [Nes18]. We should emphasize here that although the gradient mapping cannot be computed in the stochastic setting, one can still use it as a convergence measure.

[BG22] shows that the above notion of convergence criterion is closely related to the so-called *Frank-Wolfe Gap*. The FW-gap is defined as

$$g_\Theta(\bar{\theta}, \nabla f(\bar{\theta})) := \max_{y \in \Theta} \langle \nabla f(\bar{\theta}), \bar{\theta} - y \rangle. \tag{6}$$

The following proposition from [BG22] establishes the relation between the GM and the FW-gap.

Proposition 1.1 ([BG22]). *Let $g_\Theta(\cdot)$ be the Frank-Wolfe gap defined in (6) and $\mathcal{G}_\Theta(\cdot)$ be the gradient mapping defined in (5). Then, we have*

$$\|\mathcal{G}_\Theta(\bar{\theta}, \nabla f(\bar{\theta}), \beta)\|^2 \leq g_\Theta(\bar{\theta}, \nabla f(\bar{\theta})), \qquad \forall \bar{\theta} \in \Theta.$$

Moreover, under standard regularity assumption in smooth optimization (specifically, under Assumptions 2.1 and 2.2), we have

$$g_\Theta(\bar{\theta}, \nabla f(\bar{\theta})) \leq \frac{L}{\beta} \left\| \mathcal{G}_\Theta(\bar{\theta}, \nabla f(\bar{\theta}), \beta) \right\|_2 . \tag{7}$$

In this work we use a suboptimality measure, closely related to both GM and the FW-gap. At point $\bar{\theta} \in \Theta$, we define the suboptimality measure $V(\bar{\theta}, z) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as [GRW20]

$$V(\bar{\theta}, z) := \left\| \Pi_\Theta \left(\bar{\theta} - z/\beta \right) - \bar{\theta} \right\|_2^2 + \left\| z - \nabla f(\bar{\theta}) \right\|_2^2 , \tag{8}$$

where z, formally defined in Algorithm 1, is the moving-average estimate of $\nabla f(\bar{\theta})$. We show the relation among $V(\theta, z)$, and GM $\mathcal{G}_\Theta(\theta, z, \beta)$ in the following proposition.

Proposition 1.2. *Let $\{z_k\}$ be the sequence generated in Algorithm 1. Then, for $k = 1, 2, \cdots, N$,*

$$\|\mathcal{G}_\Theta(\theta_k, z_k, \beta)\|_2^2 \leq \max(2, 2\beta^2) V(\theta_k, z_k).$$

Proof. [Proof of Proposition 1.2] Using properties of projection onto a convex set, we have

$$\begin{aligned} \|\mathcal{G}_\Theta(\theta, z, \beta)\|_2^2 &\leq 2\beta^2 \|\theta - \Pi_\Theta(\theta - z/\beta)\|_2^2 + 2\beta^2 \|\Pi_\Theta(\theta - z/\beta) - \Pi_\Theta(\theta - \nabla f(\theta)/\beta)\|_2^2 \\ &\leq \max(2, 2\beta^2) V(\theta, z). \end{aligned}$$

■

The main objective of this work is to find an ϵ-stationary solution to (1), where an ϵ-stationary solution is defined as follows:

Definition 1. *A point $\bar{\theta}$ is said to be an ϵ-stationary solution to (1), if $\mathbb{E}\left[V(\bar{\theta}, z)\right] \leq \epsilon$, where the expectation is taken over all the randomness involved in the problem.*

For stochastic Frank-Wolfe-type algorithms, the oracle complexity is measured in terms of number of calls to the Stochastic First-order Oracle (SFO) and the Linear Minimization Oracle (LMO) used to the solve the sub-problems of the algorithm which involves minimizing a linear function over the convex constraint set. Formally, we have the following definition.

Definition 2. *For a given point $\theta \in \Theta$, SFO returns the stochastic gradient $\nabla F(\theta, x)$. Given a vector z, LMO returns a vector $v := \operatorname{argmin}_{y \in \Theta} \langle z, y \rangle$.*

Hence, in this work, the oracle complexity is measured in terms of the number of calls to SFO and LMO required by the proposed algorithm to obtain an ϵ-stationary solution as in Definition 1. With the above preliminaries, we now list our **main contributions**:

- In Theorem 2.1, we show that the number of calls to the SFO and LMO required by the stochastic conditional gradient-type method in Algorithm 1, with *state-dependent* Markovian data is of order $\mathcal{O}(\epsilon^{-2.5})$ and $\mathcal{O}(\epsilon^{-5.5})$ respectively in terms of the FW-Gap and the Gradient Mapping criterion. To the best of our knowledge, these are the first oracle complexity results for projection-free one-sample stochastic optimization algorithm for constrained nonconvex optimization in the Markovian setting. Our result also implies an SFO complexity of $\mathcal{O}(\epsilon^{-2.5})$ for projection-based algorithms in terms of the Gradient Mapping criterion.
- In Theorem 2.2, for the sake of completion, we also show that the number of calls to the SFO and LMO required for the case of *state-independent* Markovian data is of the order $\tilde{\mathcal{O}}(\epsilon^{-2})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$ respectively. In particular, this turns out to be of the same order as that of `iid` data ignoring the logarithmic factors.

		iid		non-iid			
				State-independent MC		State-dependent MC	
Algorithm	Criterion	SFO	LMO	SFO	LMO	SFO	LMO
1-SFW [ZSM+20]	FW-gap	$\mathcal{O}\left(\epsilon^{-3}\right)$	$\mathcal{O}\left(\epsilon^{-3}\right)$	✗	✗	✗	✗
(ASA+ICG) [XBG22]	GM	$\mathcal{O}\left(\epsilon^{-2}\right)$	$\mathcal{O}\left(\epsilon^{-3}\right)$	✗	✗	✗	✗
(ASA+ICG) [This paper]	GM	✗	✗	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\epsilon^{-3}\right)$	$\mathcal{O}\left(\epsilon^{-2.5}\right)$	$\mathcal{O}\left(\epsilon^{-5.5}\right)$

Table 1: Oracle complexity of projection-free one-sample stochastic conditional gradient algorithms for constrained non-convex optimization, to find an ϵ-stationary point.

A summary of the our contributions is provided in Table 1. We also empirically evaluate our algorithm on a strategic classification problem with 2-layer neural network classifier and show that the proposed method obtains encouraging results. We provide an experiment on single-index model regression with sparsity-inducing nuclear-norm ball constraint in Section 3.2.

1.3 Related Work

Stochastic Optimization with Dependent Data. Understanding stochastic approximation algorithms like SGD with dependent data in the asymptotic setting has been well-explored in the optimization literature. We refer to [KY03, Bor09, BMP12] for a text-book introduction to such classical results. A few recent results include [AMP05, TD17]. In the unconstrained non-asymptotic setting, [DAJJ12] studies convex optimization with ergodic data sequence. [DL22] uses multi-level gradient estimator and analyzes AdaGrad for nonconvex optimization with Markovian Data. Block coordinate descent with homogeneous Markov chain has been analyzed in [SSXY20] for nonconvex unconstrained optimization. [DX20] studies stochastic optimization with state-dependent Markov data for strongly convex functions in the context of strategic classification.

Sample-average approximation algorithms for constrained convex optimization with ϕ-mixing data was considered in [WPT+21]. [SSY18], and [AL22] analyze projected SGD for constrained nonconvex optimization with time-homogeneous Markov chain. We emphasize that the above works, except for [DX20] do not consider state-dependent Markov data. Furthermore, unlike [DX20], we study constrained nonconvex optimization problems.

There also exists work in the reinforcement learning literature on understanding stochastic optimization with Markovian data; see, for example, [BRS18, SY19, XZL19, KMN+20, DNPR20, XXLZ21, DMN+21]. However, such works are invariably focused on specific objective functions arising in the reinforcement learning setup, while our focus is on obtaining results for a general class of functions.

Conditional Gradient-Type Method. There has been significant recent advancements in understanding conditional gradient algorithm in the machine learning literature; see, for example, [Mig94, Jag13, LJJ15, LJJ15, HJN15, GKS21, BS17], for a non-exhaustive list. [HK12, HL16] provided expected oracle complexity results for stochastic conditional gradient algorithm in the stochastic convex setup. Better rates were provided by a sliding procedure in [LZ16]. In the non-convex setting, [RSPS16, YSC19, HL16] considered variance reduced stochastic conditional gradient algorithms, and provided expected oracle complexities. [QLX18] analyzed the sliding algorithm in the non-convex setting and provided results for the gradient mapping criterion. All of the above works use increasing orders of mini-batch based gradient-estimate.

To avoid mini-batches, a moving-average gradient estimator based on only one-sample in each iteration for a stochastic conditional gradient-type algorithm was proposed in [MHK20] and [ZSM+20]

for the convex and non-convex setting. However, several restrictive assumptions have been made in [MHK20] and [ZSM+20]. Specifically, [ZSM+20] requires that the stochastic gradient $G_1(x, \xi_1)$ has uniformly bounded function value, gradient-norm, and Hessian spectral-norm, and the distribution of the random vector ξ_1 has an absolutely continuous density p such that the norm of the gradient of $\log p$ and spectral norm of the Hessian of $\log p$ has finite fourth and second-moments respectively. A recent work [XBG22] provided similar convergence results under significantly weaker conditions.

2 Main Results

2.1 Methodology

We use the moving-average based single-time scale algorithm as proposed by [GRW20] for constrained optimization. The overall procedure is provided in Algorithm 1, and 2. It is worth emphasizing

Algorithm 1 Inexact Averaged Stochastic Approximation (I-ASA)

Input: $z_0, \theta_0 \in \mathbb{R}^d$, $\eta_k = (N+k)^{-a}$, $1/2 < a < 1$, β.

for $k = 1, 2, \cdots, N$ **do**

$$y_k = \begin{cases} \min\limits_{y \in \Theta} \left\{ \langle z_k, y - \theta_k \rangle + \frac{\beta}{2} \|y - \theta_k\|_2^2 \right\} & \text{(Projection)} \\ \text{ICG}(z_k, \theta_k, \beta, t_k, \omega) & \text{(No Projection)} \end{cases}$$

$\theta_{k+1} = \theta_k + \eta_{k+1}(y_k - \theta_k)$

$z_{k+1} = (1 - \eta_{k+1}) z_k + \eta_{k+1} \nabla F(\theta_k, x_{k+1})$

end for

Output: θ_R where $P(R = i) = \frac{\eta_i}{\sum_{j=1}^N \eta_j}$ for $i = 1, 2, \cdots, N$.

Algorithm 2 Inexact Conditional Gradient (ICG)

Input: $z, \theta, \beta, t, \omega$.

Set $w_0 = \theta$

for $i = 1, 2, \cdots, t-1$ **do**

Find v_i such that

$$\langle v_i, z + \beta(w_i - \theta) \rangle \leq \operatorname*{argmin}_{v \in \Theta} \langle v, z + \beta(w_i - \theta) \rangle + \beta \omega \mathcal{D}_\Theta^2 / (i+2)$$

$w_{i+1} = (1 - \mu_i) w_i + \mu_i v_i$ where $\mu_i = \frac{2}{i+2}$

end for

Output: w_t

that the above approach is similar to `ASA` algorithm introduced in [GRW20] except that we do not assume the knowledge of the exact minimizer, which is the projection of $\theta_k - z_k/\beta$ on to Θ, of the following subproblem:

$$\min_{y \in \Theta} \left\{ \langle z_k, y - \theta_k \rangle + \frac{\beta}{2} \|y - \theta_k\|_2^2 \right\}. \tag{9}$$

Instead, at iteration k, Algorithm 2 finds an approximate solution to (9) based on the conditional gradient algorithm. The main idea is to replace costly projection operator by the Inexact Conditional

Gradient (ICG) method which solves (9) approximately with access to LMO which is often much cheaper and simpler to compute. Such an approach was also used recently in [XBG22] in the context of `iid` data.

2.2 Assumptions

We now introduce the assumption that we make on the optimization problem (1). We refer to [DMPS18] for a textbook introduction to additional details regarding several assumptions below. Let $\mathcal{F}_k$ be the filtration generated by $\{\theta_0, \cdots, \theta_k, z_0, \cdots, z_k, x_1, \cdots, x_k\}$. For any mapping $g : \mathbb{R}^d \to \mathbb{R}^d$ define the norm with respect to a function $\mathcal{V} : \mathbb{R}^d \to [1, \infty)$ as

$$\|g\|_{\mathcal{V}} = \sup_{x \in \mathbb{R}^d} \frac{\|g(x)\|_2}{\mathcal{V}(x)},$$

and let $L_{\mathcal{V}} = \{g : \mathbb{R}^d \to \mathbb{R}^d, \sup_{x \in \mathbb{R}^d} \|g\|_{\mathcal{V}} < \infty\}$.

Assumption 2.1 (Constraint set)**.** *The set $\Theta \subset \mathbb{R}^d$ is convex and closed with $\max_{x,y \in \Theta} \|x - y\|_2 \leq D_\Theta$, for some $D_\Theta > 0$.*

Assumption 2.2. *Let f be a continuously differentiable function.*

Assumption 2.3. *Let $\xi_{k+1}(\theta_k, x_{k+1}) \coloneqq \nabla F(\theta_k, x_{k+1}) - \nabla f(\theta_k)$. Then,*

$$\mathbb{E}\left[\|\xi_{k+1}(\theta_k, x_{k+1})\|_2^2 \,|\mathcal{F}_k\right] \leq \sigma_1^2, \qquad \mathbb{E}\left[\|\nabla F(\theta_k, x_{k+1})\|_2^2 \,|\mathcal{F}_k\right] \leq \sigma_2^2.$$

Let $\sigma^2 \coloneqq \max(\sigma_1^2, \sigma_2^2)$.

Assumption 2.4. *Let $\{x_k\}_k$ be a Markov chain controlled by θ, i.e., there exists a transition probability kernel $P_\theta(\cdot, \cdot)$ such that*

$$\mathbb{P}(x_{k+1} \in B | \theta_0, x_0, \cdots, \theta_k, x_k) = P_{\theta_k}(x_k, B),$$

almost surely for any Borel-measurable set $B \subseteq \mathbb{R}^d$ for $k \geq 0$. For any $\theta \in \Theta$, P_θ is irreducible and aperiodic. Additionally, there exists a function $\mathcal{V} : \mathbb{R}^d \to [1, \infty)$ and a constant $\alpha_0 \geq 2$ such that for any compact set $\Theta' \subset \Theta$, we have the following.

(a) There exist a set $C \subset \mathbb{R}^d$, an integer l, constants $0 < \lambda < 1$, b, κ, $\delta > 0$, and a probability measure ν such that,

$$\sup_{\theta \in \Theta'} P_\theta^l \mathcal{V}^{\alpha_0}(x) \leq \lambda \mathcal{V}^{\alpha_0}(x) + bI(x \in C) \quad \forall x \in \mathbb{R}^d, \tag{10}$$

$$\sup_{\theta \in \Theta'} P_\theta \mathcal{V}^{\alpha_0}(x) \leq \kappa \mathcal{V}^{\alpha_0}(x) \quad \forall x \in \mathbb{R}^d, \tag{11}$$

$$\inf_{\theta \in \Theta'} P_\theta^l(x, A) \geq \delta \nu(A) \forall x \in C, \forall A \in \mathcal{B}_{\mathbb{R}^d}. \tag{12}$$

where $\mathcal{B}_{\mathbb{R}^d}$ is the Borel σ-algebra over $\mathbb{R}^d$, and for a function m, $P_\theta m(x) \coloneqq \int P_\theta(x, y) m(y)\, dy$.

(b) There exists a constant $c > 0$, such that, for all $x \in \mathbb{R}^d$,

$$\sup_{\theta \in \Theta'} \|\nabla F(\theta, x)\|_{\mathcal{V}} \leq c, \tag{13}$$

$$\left\|\nabla F(\theta, x) - \nabla F(\theta', x)\right\|_{\mathcal{V}} \leq c \left\|\theta - \theta'\right\|_2 \qquad \forall (\theta, \theta') \in \Theta'. \tag{14}$$

(c) There exists a constant $c > 0$, such that, for all $(\theta, \theta') \in \Theta' \times \Theta'$,

$$\|P_\theta g - P_{\theta'} g\|_{\mathcal{V}} \leq c \|g\|_{\mathcal{V}} \left\|\theta - \theta'\right\|_2 \quad \forall g \in L_{\mathcal{V}} \tag{15}$$

$$\|P_\theta g - P_{\theta'} g\|_{\mathcal{V}^{\alpha_0}} \leq c \|g\|_{\mathcal{V}^{\alpha_0}} \left\|\theta - \theta'\right\|_2 \quad \forall g \in L_{\mathcal{V}^{\alpha_0}}. \tag{16}$$

Some comments regarding the assumptions are in order. Assumption 2.1, and Assumption 2.2 are common in the literature on smooth constrained optimization [GRW20, XBG22, AL22, ZSM+20]. Assumption 2.1, and Assumption 2.2 together imply the Lipschitz continuity of $f(\cdot)$, i.e., there is a constant $L > 0$ such that for any $\theta_1, \theta_2 \in \Theta$, we have $|f(\theta_1) - f(\theta_2)| \leq L \|\theta_1 - \theta_2\|_2$. Assumption 2.3 is common in stochastic optimization literature. Assumption 2.4(a) is a frequently used assumption in Markov chain literature. It implies that for every $\theta \in \Theta$, there exists a stationary distribution $\pi_\theta(x)$, and the chain is $\mathcal{V}^{\alpha_0}$-uniformly ergodic [AMP05]. Assumption 2.4(c) provides smoothness guarantee on the function $f(\cdot)$. More formally, we have the following proposition.

Proposition 2.1 (Lipschitz continuous gradient [AMP05])**.** *Let Assumption 2.4 be true. Then $f(\cdot)$ has Lipschitz continuous gradient, i.e., there is a constant $L_G > 0$ such that for any $\theta_1, \theta_2 \in \Theta$:*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\|_2 \leq L_G \|\theta_1 - \theta_2\|_2 . \tag{17}$$

Finally, the most important implication of Assumption 2.4 is that it ensures the existence and regularity of a solution $u(\theta, x)$ to the Poisson equation,

$$u(\theta, x) - P_\theta u(\theta, x) = \nabla F(\theta, x) - \nabla f(\theta), \tag{18}$$

associated to the transition kernel P_θ, where $P_\theta u(\theta, x) := \int_{\mathbb{R}^d} u(\theta, x') P_\theta(x, x') dx'$. Solutions of Poisson equation have been crucial in analyzing additive functionals of Markov chain (see [AMP05] for details). In this work, the Poisson equation solution facilitates a decomposition of the noise as presented in Lemma 2.1 which is a key component of our analysis.

2.3 State-dependent Markov Chain

We now present our main result on the oracle complexity to establish a bound on $\mathbb{E}[V(\theta_k, z_k)]$. We emphasize that our results are not limited to `ICG` method but are valid for any method which can solve (9) within an error of the order of $\{\eta_k\}$.

Theorem 2.1. *Let Assumption 2.1-2.4 be true. Then, for Algorithm 1,*

(a) when a projection operator is available, choosing

$$\eta_k = (N + k)^{-3/5}, \quad \beta = 1 \tag{19}$$

for $k = 1, 2, \cdots, N$ we have

$$\mathbb{E}[V(\theta_R, z_R)] = \mathcal{O}\left(N^{-\frac{2}{5}}\right),$$

(b) when Algorithm 2 is used to solve (9),choosing

$$\eta_k = (N + k)^{-3/5}, \quad t_k = \eta_k^{-2}, \quad \beta = 1, \quad \omega = 1, \quad \mu_i = 2/(i + 2) \tag{20}$$

for $k = 1, 2, \cdots, N$ we have

$$\mathbb{E}[V(\theta_R, z_R)] = \mathcal{O}\left(N^{-\frac{2}{5}}\right),$$

where the expectations are taken with respect to all the randomness of the algorithm, and an independent integer random variable $R \in \{1, 2, \cdots, N\}$ with probability mass function,

$$P(R = k) = \eta_k / \sum_{k=1}^{N} \eta_k \quad k \in \{1, 2, \cdots, N\}.$$

Remark 1. *Note that total number of LMO calls are $\sum_{k=1}^{N} t_k = \sum_{k=1}^{N} t_k = \sum_{k=1}^{N} (N+k)^{2a} = \mathcal{O}(N^{11/5})$. In other words, to achieve $\|\mathcal{G}_\Theta(\theta_R, \nabla f(\theta_R), \beta)\|_2^2 \leq \mathbb{E}[V(\theta_R, z_R)] \leq \epsilon$, SFO and LMO complexities are respectively $\epsilon^{-2.5}$, and $\epsilon^{-5.5}$. Note that the SFO complexity will be $\epsilon^{-2.5}$ as long as one has an approximation of the projection operator with approximation error $\mathcal{O}(\eta_k)$.*

Remark 2. *In Theorem 2.1, one obtains sublinear rate $\max(N^{a-1}, N^{2-4a})$ with $\eta_k = (N+k)^{-a}$ for $1/2 < a < 1$. Choosing $a = 3/5$ provides the fastest rate of convergence.*

Before sketching the outline of the proof, we present the following lemma which provides a decomposition of the noise $\xi_k(\theta_{k-1}, x_k)$ – one of the key results used in the proof of the main theorem. The lemma and its proof are almost same as Lemma A.5 in [Lia10] with the only difference that unlike [Lia10], where the iterates are of SGD, we need to prove it for the iterates of Algorithm 1. We provide the proof in Appendix A.

Lemma 2.1. *Let Assumption 2.1-2.4 be true. Then the following decomposition takes place:*

$$\xi_k(\theta_{k-1}, x_k) = e_k + \nu_k + \zeta_k,$$

where, $\{e_k\}_k$ is martingale difference sequence, $\mathbb{E}[\|\nu_k\|_2] \leq \eta_k$, and $\zeta_k = (\tilde{\zeta}_k - \tilde{\zeta}_{k+1})/\eta_k$, where $\{\tilde{\zeta}_k\}_k$ is a sequence such that $\mathbb{E}\left[\|\tilde{\zeta}_k\|_2\right] \leq \eta_k$. The exact expressions of $\{e_k\}_k$, $\{\nu_k\}_k$, and $\{\tilde{\zeta}_k\}_k$ are provided in (30).

Outline of the proof of Theorem 2.1: The key step in the analysis of Algorithm 1 involves controlling the expectation of interaction with noise of the form $\langle \nabla f(\theta_k) - \nabla f(\theta_{k-1}), \xi_{k+1}(\theta_k, x_{k+1}) \rangle$. For `iid` or martingale difference data it is easy to control because

$$\mathbb{E}[\langle \nabla f(\theta_k) - \nabla f(\theta_{k-1}), \xi_{k+1}(\theta_k, x_{k+1}) \rangle | \mathcal{F}_k] = 0.$$

However, this is no longer true for Markov chain data. To resolve the issue, first notice that under our assumptions, the noise sequence ξ_k can be decomposed into the sum of a martingale difference sequence $\{e_k\}$ and some residual terms $\{\nu_k\}$, and $\{\zeta_k\}$ as shown in Lemma 2.1. Then the key step is to introduce a different sequence of hypothetical iterates $(\tilde{\theta}_k, \tilde{y}_k, \tilde{z}_k)$ for which the noise is small enough so that we can bound $\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right]$, and then show that these hypothetical iterates and the original sequence generated by Algorithm 1 are close enough so that $\mathbb{E}[V(\theta_k, z_k)]$ is of the same order as $\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right]$. This forms the main novelty in our analysis.

Specifically, the hypothetical sequence that we consider is given by:

$$\tilde{\theta}_0 = \theta_0 \quad \tilde{z}_0 = z_0 \tag{21}$$

$$\tilde{y}_k = \underset{y \in \Theta}{\operatorname{argmin}} \left\{ \left\langle \tilde{z}_k, y - \tilde{\theta}_k \right\rangle + \frac{\beta}{2} \left\| y - \tilde{\theta}_k \right\|_2^2 \right\} \tag{22}$$

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k + \eta_{k+1}(\tilde{y}_k - \tilde{\theta}_k) \tag{23}$$

$$\tilde{z}_{k+1} = z_{k+1} + \tilde{\zeta}_{k+2}. \tag{24}$$

This also means that we have

$$\tilde{z}_{k+1} = (1 - a\eta_{k+1})\tilde{z}_k + a\eta_{k+1}\left(\nabla f(\theta_k) + \tilde{\epsilon}_{k+1}\right), \tag{25}$$

where, $\tilde{\epsilon}_k = e_k + \nu_k + \tilde{\zeta}_k$. Note that by Lemma 2.1, $\mathbb{E}\left[e_k\right] = 0$, and $\mathbb{E}\left[\left\|\nu_k + \tilde{\zeta}_k\right\|_2\right] \leq \eta_k$. First we show that by choosing

$$\eta_k = (N+k)^{-a}, \quad 1/2 < a < 1 \quad \text{and} \quad t_k = 1/\eta_k^2,$$

one has that

$$\mathbb{E}\left[\|\tilde{\theta}_k - \theta_k\|_2^2\right] = \mathcal{O}\left(N^{2-4a}\right), \quad \text{and} \quad \mathbb{E}\left[V(\theta_k, z_k)\right] \leq 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + \mathcal{O}\left(N^{2-4a}\right).$$

Then we establish a similar bound on $V(\tilde{\theta}_k, \tilde{z}_k)$. Combining the above two facts proves Theorem 2.1. The full proof is deferred to Appendix A.1.

2.4 State-independent Markov Chain

While our main goal in this work is to analyze Algorithm 1 for constrained nonconvex optimization with state-dependent Markov chain data, we provide the following result on the complexity of Algorithm 1 for Markov chain data with state-independent transition kernel for the sake of completion. Here we use P to denote the transition kernel (as opposed to P_θ for state-dependent kernel). Note that under Assumption 2.4(a), for each θ, the chain is $\mathcal{V}$-uniformly ergodic, and hence, exponentially mixing [MT12] in the following sense.

Definition 3. *A Markov chain is said to be exponentially mixing, if there exists $C, r > 0$ such that, for any initial state x,*

$$\|P^n(x, \cdot) - \pi\|_{\mathcal{V}} \leq C\exp(-rn), \tag{26}$$

where $P^n(x, \cdot)$ is the distribution of X_n with initial state $X_0 = x$.

Now we present our result on the complexity of Algorithm 1 to find an ϵ-stationary solution to (1) for exponentially-mixing Markov chain data with state-independent transition kernel.

Theorem 2.2. *Let Assumption 2.1-2.3 be true. Let Assumption 2.4(a)-(b) be true with P_θ replaced by P. Then, for Algorithm 1,*

(a) when the projection operator is available, choosing

$$\eta_k = 1/\sqrt{N}, \quad \beta = 1 \tag{27}$$

for $k = 1, 2, \cdots, N$ we have

$$\mathbb{E}\left[V(\theta_R, z_R)\right] = \mathcal{O}\left(\log N/\sqrt{N}\right),$$

(b) when Algorithm 2 is used, choosing

$$\eta_k = 1/\sqrt{N}, \quad t_k = \lceil\sqrt{k}\rceil, \quad \beta = 1, \quad \omega = 1, \quad \mu_i = 2/(i+2) \tag{28}$$

for $k = 1, 2, \cdots, N$ we have

$$\mathbb{E}\left[V(\theta_R, z_R)\right] = \mathcal{O}\left(\log N/\sqrt{N}\right),$$

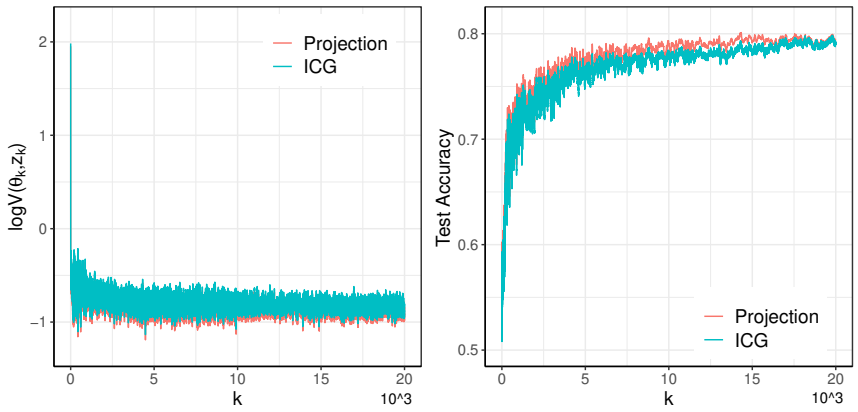

Figure 1: Strategic Classification: (*Left*): Performance of Algorithm 1 with and without the projection operator. (*Right*): Test Accuracy with Algorithm 1 with and without the projection operator.

where the expectation is taken with respect to all the randomness of the algorithm, and an independent integer random variable $R \in \{1, 2, \cdots, N\}$ *whose probability mass function is given by,*

$$P(R = k) = \eta_k / \sum_{k=1}^{N} \eta_k \quad k \in \{1, 2, \cdots, N\}.$$

We defer the proof to the Appendix.

Remark 3. *To find an* ϵ*-stationary point, the total number of calls to SFO and LMO are* $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$*, and* $\tilde{\mathcal{O}}\left(\epsilon^{-3}\right)$*, where* $\tilde{\mathcal{O}}(\cdot)$ *denotes the order ignoring logarithmic factors.*

Remark 4. *The authors of [AL22] obtain the same rate as in Theorem 2.2 for constrained (but projection-based) nonconvex optimization with state-independent exponentially mixing data. In the state-dependent case, since the transition kernel of the Markov chain is controlled by* θ_k*, and the transition kernel is assumed to be only Lipschitz smooth in* θ *(16), the chain does not necessarily exponentially mix. In the state-independent case, since the chain mixes exponentially we obtain the same rate as well. While their results are for projection-based algorithms, we analyze a projection-free LMO-based algorithm since LMO is often computationally cheaper than projection.*

3 Experimental Evaluation

3.1 Strategic Classification

In this section we illustrate our algorithm on the strategic classification problem as described in Section 1.1 with the `GiveMeSomeCredit`[1] dataset. The main task is a credit score classification problem where the bank (learner) has to decide whether a loan should be granted to a client. Given the knowledge of the classifier the clients (agents) can distort some of their personal traits in order to get approved for a loan. Here we use a 2-layer neural network with width m as the classifier, given by

$$h(u; \mathcal{W}, \mathcal{A}, \mathcal{B}) = \sum_{i=1}^{m} \mathcal{A}_i \upsilon(\mathcal{W}_i^\top u + \mathcal{B}_i),$$

[1] Available at https://www.kaggle.com/c/GiveMeSomeCredit/data

where $\upsilon(\cdot)$ is the activation function, $\mathcal{W}_i \in \mathbb{R}^d$ and

$$\mathcal{W} = [\mathcal{W}_1, \mathcal{W}_2, \cdots, \mathcal{W}_m]^\top \in \mathbb{R}^{m\times d}, \mathcal{A} = \left(\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_m\right) \in \mathbb{R}^m, \mathcal{B} = \left(\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_m\right) \in \mathbb{R}^m.$$

We will use θ to collectively denote $(\mathcal{W}, \mathcal{A}, \mathcal{B})$. We impose the constraint of sparsity on the classifier given by $\|\theta\|_1 \leq R$ for some $R > 0$. As loss function we consider logistic loss as shown in (2). We consider a quadratic cost given by $c(u, u') = \|u_S - u'_S\|_2^2/(2\lambda)$ where λ is the sensitivity of the underlying distribution on θ. We assume that the agents iteratively learn u'_S similar to [LW22]. Note that unlike [LW22], the closed form of best response is not known here. So we assume that the agents use Gradient Ascent (GA) to learn the best response. For $\|\theta\|_1 \leq R$ constraint, the LMO in Algorithm 2 at iteration k is given by,

$$i = \underset{j=1,\cdots,d}{\operatorname{argmax}} |q_j| \quad \text{LMO} = -R\operatorname{sign}(q_i),$$

where $q = z + \beta(w_k - \theta)$, and q_j is the j-th coordinate of q. We select a subset of randomly chosen $M = 2000$ samples (agents) such that the dataset is balanced. Each agent has 10 features. Note that since Algorithm 1 computes the gradient on one sample at every iterate, the computation time is independent of the total number of agents. We assume that the agents can modify Revolving Utilization, Number of Open Credit Lines, and Number of Real Estate Loans or Lines. In this experiment we set $n_1 = 200$. Similar to [LW22], we set $\alpha = 0.5\lambda$, and $\lambda = 0.01$. For the classifier, the activation function is chosen as *sigmoidal*, and $m = 400$. We set $N = 20000$, and $R = 4000$. All the parameters of Algorithm 1 are chosen as described in (20). Figure 3.1 shows that Algorithm 1 finds an ϵ-stationary point of the strategic classification problem. We show that Algorithm 1 performs comparably with Averaged Stochastic Approximation with the projection operator. Each curve in Figure 3.1 is an average of 50 repetitions.

3.2 Single Index Model with Trace-norm Ball Constraint

In this section we illustrate our algorithm on a synthetic example on single-index model regression with a nuclear-norm constraint on the model parameter. Let $\|\cdot\|_*$ denote the nuclear norm. The features $\{x_k\}_k \in \mathbb{R}^{d_1\times d_2}$ are a matrix-valued time-series given by,

$$x_k = Ax_{k-1} + E_k + W_k\upsilon\theta_k,$$

where $A \in \mathbb{R}^{d_1\times d_1}$ matrix with spectral radius less than 1, $E_k \in \mathbb{R}^{d_1\times d_2}$ is the noise matrix with each entry of E_k is `iid` $N(0,1)$ random variable, W_k is a $Bernoulli(0.5)$ random variable, and $\upsilon \in \mathbb{R}$. For a fixed $\theta_k = \theta$, $\{x_k\}_k$ has a stationary distribution as shown in Proposition 1 of [**?**]. $\{E_k\}_k$, and $\{W_k\}_k$ are `iid` sequence. This Markov chain follows conditions (b) and (c) of Assumption 2.4 since the evolution of x_k only involves linear terms in θ_k. The responses $\{y_k\}_k$ are generated according to the following single index model,

$$y_k = g(x_k^\top\theta^*) + \tilde{E}_k,$$

where $\{\tilde{E}_k\}_k$ is an `iid` sequence of standard normal random variables, $\theta^* \in \mathbb{R}^{d_1\times d_2}$ is a matrix with $\|\theta^*\|_* \leq 1$, and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is the link function. For this experiment we choose $g(x) = 3x + 5\sin(x)$. Since y_k only depends on x_k, and g is a Lipschitz continuous function of θ, Assumption 2.4 holds for (x_k, y_k). It is easy to see that Assumptions 2.1 - 2.3 holds for this example. The constraint set is given by $\|\theta\|_* \leq 1$, i.e., we assume that θ^* has a low-rank structure. The goal is to minimize the expected squared loss with the constraint $\|\theta\|_* \leq 1$, i.e.,

$$\min_{\|\theta\|_*\leq 1} \mathbb{E}\left[(y - g(x^\top\theta))^2\right]. \tag{29}$$

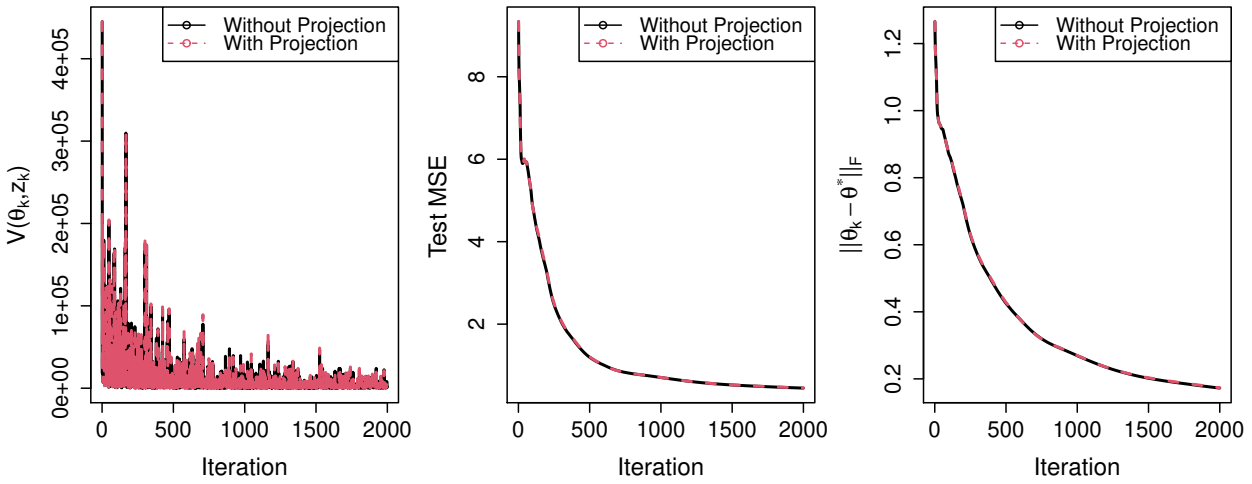

Figure 2: Single-index model with nuclear-norm constraint: (*Left*): Performance of Algorithm 1 with and without the projection operator. (*Middle*): Test Mean Squared Error (MSE) with Algorithm 1 with and without the projection operator. (*Right*): $\|\theta - \theta^*\|_F$ with Algorithm 1 with and without the projection operator.

The advantages of conditional-gradient based method for nuclear-norm ball constrained problems have been studied extensively [**?**, Jag13, HJN15]. The main advantage of ICG-based method is that calculating the LMO in this case requires computation of the leading singular vector of gradient matrix whereas to calculate the projection on the trace-norm ball one needs to compute the complete singular value decomposition. Let u_1, v_1 are the leading left and right singular vectors of the noisy gradient matrix evaluated at $(\theta; x, y)$, $-2(y - g(x^\top\theta))g'(x^\top\theta)x$. Then the LMO is given by,

$$\text{LMO} = -u_1 v_1^\top.$$

For this experiment we choose $d_1 = 10$, $d_2 = 20$, $\upsilon = 0.1$, and $N = 2000$. Rest of the parameters of Algorithm 1 are chosen according to Theorem 2.1. In Figure 2, we compare the projection-based and ICG based version of Algorithm 1 with respect to $V(\theta_k, z_k)$, test Mean Squared Error (MSE), and $\|\theta_k - \theta^*\|_F$ where $\|\cdot\|_F$ is the Fröbenius norm. Figure 2 shows that the performance of projection-based and the ICG-based versions of Algorithm 1 are almost same. Each plot in Figure 2 is the average of 50 repetitions.

4 Discussion

In this work we provide oracle complexity results for the stochastic conditional gradient algorithm to find an ϵ-stationary point of a constrained nonconvex optimization problem with state-dependent Markovian data. In Theorem 2.1, we show that the number of calls to the SFO and LMO required by the stochastic conditional gradient-type method in Algorithm 1, with *state-dependent* Markovian data, is $\mathcal{O}(\epsilon^{-2.5})$ and $\mathcal{O}(\epsilon^{-5.5})$ respectively. To the best of our knowledge, these are the first oracle complexity results in this setting. In Theorem 2.2, we show also that the SFO and LMO complexity in the case of state-independent Markovian data is $\tilde{\mathcal{O}}(\epsilon^{-2})$ and $\tilde{\mathcal{O}}(\epsilon^{-3})$ respectively, which matches the corresponding results in the `iid` setting.

There are various avenues for further extensions. Improving the established complexities and/or proving matching lower bounds on the oracle complexity of projection-free algorithms in the

Markovian setting is extremely interesting. It is also intriguing to establish upper and lower bounds on the oracle complexity for more general types of dependent data sequences arising in applications, including ϕ and α-mixing sequences. Yet another exciting direction is that of designing algorithms adaptive to the dependency in the data that achieve potentially better oracle complexity bounds.

A Proofs for the State-dependent Case

Before proving Lemma 2.1, we present the following result on Poisson equation solution from [Lia10], and [AMP05] which is crucial to the proof of Lemma 2.1.

Lemma A.1 ([Lia10]). *Let Assumption 2.4 be true. Then we have the following:*

(a) For any $\theta \in \Theta$, the Markov kernel P_θ has a unique stationary distribution π_θ. Moreover, $\nabla F(\theta, x) : \Theta \times \mathbb{R}^d \to \Theta$ is measurable for all $\theta \in \Theta$, and $\mathbb{E}_{x\sim\pi_\theta}[\nabla F(\theta, x)] < \infty$.

(b) For any $\theta \in \Theta$, the solution to the Poisson equation (18) exists. Furthermore, there exist a function $\mathcal{V} : \mathbb{R}^d \to [1, \infty)$ such that for all $\theta \in \Theta$, the following holds:

(i) $\sup_{\theta\in\Theta} \|\nabla F(\theta, x)\|_{\mathcal{V}} < \infty$,
(ii) $\sup_{\theta\in\Theta} (\|u(\theta, x)\|_{\mathcal{V}} + \|P_\theta u(\theta, x)\|_{\mathcal{V}}) < \infty$,
(iii) $\sup_{\theta\in\Theta} (\|u(\theta, x) - u(\theta', x)\|_{\mathcal{V}} + \|P_\theta u(\theta, x) - P_{\theta'} u(\theta', x)\|_{\mathcal{V}}) < \|\theta - \theta'\|_2$.

We now prove Lemma 2.1.

Proof. [Proof of Lemma 2.1] Let,

$$\begin{aligned}
e_{k+1} &= u(\theta_k, x_{k+1}) - P_{\theta_k} u(\theta_k, x_k) \\
\nu_{k+1} &= P_{\theta_{k+1}} u(\theta_{k+1}, x_{k+1}) - P_{\theta_k} u(\theta_k, x_{k+1}) + \frac{\eta_{k+2} - \eta_{k+1}}{\eta_{k+1}} P_{\theta_{k+1}} u(\theta_{k+1}, x_{k+1}) \\
\tilde{\zeta}_{k+1} &= \eta_{k+1} P_{\theta_k} u(\theta_k, x_k) \\
\zeta_{k+1} &= \frac{\tilde{\zeta}_{k+1} - \tilde{\zeta}_{k+2}}{\eta_{k+1}}.
\end{aligned} \tag{30}$$

Now, one has,

$$\mathbb{E}[e_{k+1}|\mathcal{F}_k] = \mathbb{E}[u(\theta_k, x_{k+1})|\mathcal{F}_k] - P_{\theta_k} u(\theta_k, x_k) = 0$$

We also have $\mathbb{E}[|e_{k+1}|] < \infty$. So e_{k+1} is a martingale difference sequence. We also have, using Lemma A.1, and the fact that Θ is compact,

$$\mathbb{E}[\|\nu_{k+1}\|_2] \le c_1 \|\theta_k - \theta_{k+1}\|_2 + c_2\eta_{k+2} \le c_1\eta_{k+1} \|y_k - \theta_k\|_2 + c_2\eta_{k+2} \le c_3\eta_{k+1}.$$

Again, using Lemma A.1, we have

$$\mathbb{E}\left[\left\|\tilde{\zeta}_{k+1}\right\|_2\right] \le \eta_{k+1}\mathbb{E}\left[\|P_{\theta_k} u(\theta_k, x_k)\|_2\right] \le c_4\eta_{k+1},$$

where $c_i,\ i = 1, 2, 3, 4$ are constants. ∎

A.1 Proof of Theorem 2.1

Before proving Theorem 2.1, we start with a few preliminaries. Let

$$y_k' = \operatorname*{argmin}_{y\in\Theta} \left\{ \langle z_k, y-\theta_k\rangle + \frac{\beta}{2}\|y-\theta_k\|_2^2 \right\},$$

and let $\|y_k - y_k'\|_2 \le \delta_k$. We need the following result from [Jag13] which bounds the distance between iterates generated by Algorithm 2, y_k, and y_k':

Lemma A.2 ([Jag13]). *Under Assumption 2.1,*

$$\left\|y_k - y_k'\right\|_2^2 \le \frac{4\mathcal{D}_\Theta^2(1+\omega)}{t_k+2},$$

where ω is the accuracy of the LMO.

Consider the following system:

$$\tilde\theta_0 = \theta_0 \quad \tilde z_0 = z_0 \tag{31}$$

$$\tilde y_k = \operatorname*{argmin}_{y\in\Theta} \left\{ \left\langle \tilde z_k, y-\tilde\theta_k\right\rangle + \frac{\beta}{2}\left\|y-\tilde\theta_k\right\|_2^2 \right\} \tag{32}$$

$$\tilde\theta_{k+1} = \tilde\theta_k + \eta_{k+1}(\tilde y_k - \tilde\theta_k) \tag{33}$$

$$\tilde z_{k+1} = z_{k+1} + \tilde\zeta_{k+1}. \tag{34}$$

Equivalently one can also write:

$$\tilde y_k = \Pi_\Theta\left(\tilde\theta_k - \frac{1}{\beta}\tilde z_k\right), \tag{35}$$

where Π_Θ is the orthogonal projection on the set Θ. Let $\phi(\theta, z)$ be the following function:

$$\phi(\theta, z) = \min_{y\in\Theta}\left(\langle z, y-\theta\rangle + \frac{\beta}{2}\|y-\theta\|_2^2\right). \tag{36}$$

By [GRW20, Lemma 4], the function $\phi(\theta, z)$ has Lipschitz continuous gradient with Lipschitz constant L_ϕ for some $L_\phi > 0$. Define the merit function

$$W(\theta, z) \coloneqq (f(\theta) - f^*) - \phi(\theta, z). \tag{37}$$

Recall that as optimality measure we use the following:

$$V(\theta_k, z_k) = \left\|\Pi_\Theta\left(\theta_k - \frac{z_k}{\beta}\right) - \theta_k\right\|_2^2 + \|z_k - \nabla f(\theta_k)\|_2^2. \tag{38}$$

We now introduce a few intermediate results that are crucial in proving Theorem 2.1. Specifically, in Lemma A.3, we show that the iterates generated by the auxiliary updates $\tilde\theta_k$ are close to the original updates θ_k of Algorithm 1. Next, in Lemma A.4 we show that $V(\theta_k, z_k)$ is close to $V(\tilde\theta_k, \tilde z_k)$. Next, note that from (38), we have

$$\sum_{k=1}^N \eta_k V(\tilde\theta_k, \tilde z_k) = \sum_{k=1}^N \eta_k \left\|\Pi_\Theta\left(\tilde\theta_k - \frac{\tilde z_k}{\beta}\right) - \tilde\theta_k\right\|_2^2 + \sum_{k=1}^N \eta_k \left\|\tilde z_k - \nabla f(\tilde\theta_k)\right\|_2^2$$

We bound the first and the second term in the right hand side of the above equation in Lemma A.5, Lemma A.6 and (47) respectively.

Lemma A.3. *Let the conditions of Lemma 2.1 hold. Then, for $k \geq 1$, and for any $\gamma \in \mathbb{R}$, we have*

$$\mathbb{E}\left[\left\|\tilde{\theta}_k - \theta_k\right\|_2^2\right] \leq \eta_k^3(1+\eta_k^{-\gamma}) + 2\sum_{i=1}^{k-1}\eta_i^3\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k}\left(1+\eta_j^{1+\gamma}\right) + 2\sum_{i=1}^{k-1}\eta_i\mathbb{E}\left[\delta_{i-1}^2\right]\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k}\left(1+\eta_j^{1+\gamma}\right). \tag{39}$$

Lemma A.4. *Let the conditions of Lemma 2.1 be true. Then, choosing $\eta_k = (N+k)^{-a}$, $a > 1/2$, and setting γ of Lemma A.3 to $\gamma = 1/a - 1$, for $\delta_k \leq \eta_k$ we get*

$$\mathbb{E}\left[V(\theta_k, z_k)\right] \leq 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9+4L_G)\left(N^{1-4a} + 8N^{2-4a}\right) + 12N^{-2a}.$$

Lemma A.5. *Let Assumption 2.1, Assumption 2.2, and Assumption 2.4 be true. Let $\{\tilde{\theta}_k, \tilde{z}_k, \tilde{y}_k, \}_{k\geq 0}$ be the sequence generated by (31)-(34). Then $\forall k \geq 0$,*

$$\frac{\beta}{2}\sum_{k=0}^{N-1}\eta_{k+1}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 \leq W(x_0, z_0) + \sum_{k=0}^{N-1} r_{k+1} \quad \forall N \geq 1, \tag{40}$$

where for $k \geq 0$, and

$$r_{k+1} = \frac{(L_G + L_\phi)\eta_{k+1}^2}{2}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 + \frac{L_\phi}{2}\left\|\tilde{z}_{k+1} - \tilde{z}_k\right\|_2^2 + \eta_{k+1}\left\langle\tilde{\theta}_k - \tilde{y}_k, \tilde{\epsilon}_{k+1}\right\rangle + \frac{\eta_{k+1}L_G^2}{\beta}\left\|\theta_k - \tilde{\theta}_k\right\|_2^2.$$

Lemma A.6. *Let $\{\tilde{\theta}_k, \tilde{z}_k, \tilde{y}_k, \}_{k\geq 0}$ be the sequence generated by (31)-(34), and Assumption 2.1-2.4 hold. Then,*

1. *If $\eta_0 = 1$, we have,*

$$\beta^2\mathbb{E}\left[\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 |\mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[\left\|\tilde{z}_k\right\|_2^2 |\mathcal{F}_{k-1}\right] \leq \sigma^2 \quad \forall k \geq 1; \tag{41}$$

2. *If $\eta_k \leq 1/\sqrt{2}$ for all $k \geq 1$, then,*

$$\sum_{k=0}^{\infty}\mathbb{E}\left[\left\|\tilde{z}_{k+1} - \tilde{z}_k\right\|_2^2 |\mathcal{F}_k\right] \leq 2\left(\left\|\tilde{z}_0\right\|_2^2 + 24\sigma^2\sum_{k=0}^{\infty}\eta_k^2\right), \tag{42}$$

$$\sum_{k=0}^{\infty}\mathbb{E}\left[r_{k+1}|\mathcal{F}_k\right] \leq \sigma_3^2\sum_{k=0}^{\infty}\eta_k^2 + \frac{L_G^2}{\beta}\sum_{k=0}^{\infty}\eta_{k+1}\mathbb{E}\left[\left\|\theta_k - \tilde{\theta}_k\right\|_2^2\right], \tag{43}$$

where

$$\sigma_3^2 = \frac{1}{2}\left((3L_G + L_\phi)\frac{\sigma^2}{\beta^2} + 4L_\phi(\|z_0\|_2^2 + 24\sigma^2) + 2\right).$$

The proofs of Lemma 2.1, Lemma A.3, Lemma A.5, and Lemma A.6 are provided in Section A.2. We now prove Theorem 2.1.

Proof. [Proof of Theorem 2.1] Define,

$$\Gamma_1 \coloneqq 1 \qquad \Gamma_k \coloneqq \prod_{i=0}^{k-1}(1-\eta_{i+1}) \quad \forall k \geq 2. \tag{44}$$

Using the update of Algorithm 1 we have

$$\begin{aligned}\nabla f(\tilde{\theta}_{k+1}) - \tilde{z}_{k+1} = (1-\eta_{k+1})\left(\nabla f(\tilde{\theta}_k) - \tilde{z}_k + \nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k)\right) \\ + \eta_{k+1}\left(\nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k) - \tilde{\epsilon}_{k+1}\right) + \eta_{k+1}\left(\nabla f(\tilde{\theta}_k) - \nabla f(\theta_k)\right).\end{aligned}$$

Dividing both sides of the above equation by Γ_{k+1} we obtain

$$\begin{aligned}\frac{\nabla f(\tilde{\theta}_{k+1}) - \tilde{z}_{k+1}}{\Gamma_{k+1}} =& \frac{1}{\Gamma_k}\left(\nabla f(\tilde{\theta}_k) - \tilde{z}_k + \nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k)\right) + \frac{\eta_{k+1}}{\Gamma_{k+1}}\left(\nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k) - \tilde{\epsilon}_{k+1}\right) \\ &+ \frac{\eta_{k+1}}{\Gamma_{k+1}}\left(\nabla f(\tilde{\theta}_k) - \nabla f(\theta_k)\right) \\ =& \frac{1}{\Gamma_k}\left(\nabla f(\tilde{\theta}_k) - \tilde{z}_k\right) + \frac{1}{\Gamma_{k+1}}\left(\nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k)\right) - \frac{\eta_{k+1}}{\Gamma_{k+1}}\left(\tilde{\epsilon}_{k+1} + \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k)\right)\end{aligned}$$

Summing both sides from $k=1$ to $k=i-1$ we obtain

$$\nabla f(\tilde{\theta}_i) - \tilde{z}_i = \sum_{k=0}^{i-1}\frac{\Gamma_i}{\Gamma_{k+1}}\left(\nabla f(\tilde{\theta}_{k+1}) - \nabla f(\tilde{\theta}_k)\right) - \sum_{k=0}^{i-1}\frac{\eta_{k+1}\Gamma_i}{\Gamma_{k+1}}\left(\tilde{\epsilon}_{k+1} + \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k)\right).$$

Hence,

$$\begin{aligned}\nabla f(\tilde{\theta}_i) - \tilde{z}_i =& \frac{\Gamma_i}{\Gamma_{i-1}}\left(\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right) + \left(\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right) - \eta_i\left(\tilde{\epsilon}_i + \nabla f(\theta_{i-1}) - \nabla f(\tilde{\theta}_{i-1})\right) \\ =& (1-\eta_i)\left(\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right) + \frac{\eta_i}{\eta_i}\left(\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right) - \eta_i\left(\tilde{\epsilon}_i + \nabla f(\theta_{i-1}) - \nabla f(\tilde{\theta}_{i-1})\right).\end{aligned}$$

Using Young's inequality and Jensen's inequality,

$$\begin{aligned}\left\|\nabla f(\tilde{\theta}_i) - \tilde{z}_i\right\|_2^2 \leq& \frac{1-\eta_i/4}{1-\eta_i/2}\left\|(1-\eta_i)\left(\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right) + \frac{\eta_i}{\eta_i}\left(\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right) - \eta_i\tilde{\epsilon}_i\right\|_2^2 \\ &+ \frac{4-\eta_i}{\eta_i}\eta_i^2\left\|\nabla f(\theta_{i-1}) - \nabla f(\tilde{\theta}_{i-1})\right\|_2^2 \\ \leq& \frac{1-\eta_i/4}{1-\eta_i/2}\cdot I_1 + 4L_G^2\eta_i\left\|\theta_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2, \end{aligned} \tag{45}$$

where

$$\begin{aligned}I_1 \coloneqq& (1-\eta_i)\left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2 + \frac{\left\|\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right\|_2^2}{\eta_i} + \eta_i^2\left\|\tilde{\epsilon}_i\right\|_2^2 \\ &- 2\eta_i\left\langle(1-\eta_i)\left(\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right) + \left(\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right), \tilde{\epsilon}_i\right\rangle.\end{aligned}$$

Taking conditional expectation of I_1 with respect to $\mathcal{F}_{i-1}$, using (17), Assumption 2.4, and (33), we then obtain

$$\begin{aligned}\mathbb{E}\left[I_1 \mid \mathcal{F}_{i-1}\right] \leq& (1-\eta_i)\left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2 + \eta_i L_G^2\left\|\tilde{y}_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2 + \eta_i^2\sigma^2 \\ &- 2\eta_i\mathbb{E}\left[\left\langle(1-\eta_i)\left(\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right) + \left(\nabla f(\tilde{\theta}_i) - \nabla f(\tilde{\theta}_{i-1})\right), \tilde{\zeta}_i\right\rangle|\mathcal{F}_{i-1}\right] \\ \leq& (1-\eta_i)\left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2 + \eta_i L_G^2\mathbb{E}\left[\left\|\tilde{y}_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2|\mathcal{F}_{i-1}\right] + \eta_i^2\sigma^2\end{aligned}$$

$$+ 2\eta_i^2(1-\eta_i)^2 \left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2 + 2\eta_i^4 L_G^2 \left\|\tilde{y}_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2 + \eta_i^2$$
$$\leq \left(1 - \frac{\eta_i}{2}\right) \left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2 + 2\eta_i L_G^2 \mathbb{E}\left[\left\|\tilde{y}_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2 |\mathcal{F}_{i-1}\right] + \eta_i^2(1+\sigma^2).$$

Now, taking expectation on both sides of (45),

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_i) - \tilde{z}_i\right\|_2^2\right] \leq& \left(1 - \frac{\eta_i}{4}\right) \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_{i-1}) - \tilde{z}_{i-1}\right\|_2^2\right] + 4\eta_i L_G^2 \mathbb{E}\left[\left\|\tilde{y}_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2\right] \\
&+ 2\eta_i^2(1+\sigma^2) + 4L_G^2\eta_i \mathbb{E}\left[\left\|\theta_{i-1} - \tilde{\theta}_{i-1}\right\|_2^2\right] \\
\leq& Y_0^i \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0) - \tilde{z}_0\right\|_2^2\right] + 4L_G^2 \sum_{k=1}^{i} Y_k^i \eta_k \mathbb{E}\left[\left\|\tilde{y}_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right] + 2\sum_{k=1}^{i} Y_{k-1}^i \eta_k^2(1+\sigma^2) \\
&+ 4L_G^2 \sum_{k=1}^{i} Y_{k-1}^i \eta_k \mathbb{E}\left[\left\|\theta_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right],
\end{aligned}$$

where

$$Y_i^i = 1 \quad \text{for all } i, \qquad \text{and} \qquad Y_k^i = \prod_{j=k+1}^{i} \left(1 - \frac{\eta_j}{4}\right) \quad \text{for } i > k. \tag{46}$$

Then,

$$\begin{aligned}
\sum_{i=1}^{N} \eta_i \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_i) - \tilde{z}_i\right\|_2^2\right] \leq& \sum_{i=1}^{N} \eta_i Y_0^i \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0) - \tilde{z}_0\right\|_2^2\right] + 4L_G^2 \sum_{i=1}^{N}\sum_{k=1}^{i} Y_k^i \eta_i \eta_k \mathbb{E}\left[\left\|\tilde{y}_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right] \\
&+ 2\sum_{i=1}^{N}\sum_{k=1}^{i} Y_{k-1}^i \eta_i \eta_k^2(1+\sigma^2) + 4L_G^2 \sum_{i=1}^{N}\sum_{k=1}^{i} Y_{k-1}^i \eta_i \eta_k \mathbb{E}\left[\left\|\theta_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right] \\
=& \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0) - \tilde{z}_0\right\|_2^2\right] + 4L_G^2 \sum_{k=1}^{N}\sum_{i=k}^{N} Y_k^i \eta_i \eta_k \mathbb{E}\left[\left\|\tilde{y}_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right] \\
&+ 2\sum_{k=1}^{N}\sum_{i=k}^{N} Y_{k-1}^i \eta_i \eta_k^2(1+\sigma^2) + 4L_G^2 \sum_{k=1}^{N}\sum_{i=k}^{N} Y_{k-1}^i \eta_i \eta_k \mathbb{E}\left[\left\|\theta_{k-1} - \tilde{\theta}_{k-1}\right\|_2^2\right] \\
\leq& \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0) - \tilde{z}_0\right\|_2^2\right] + 4L_G^2 \sum_{k=0}^{N-1} \eta_k \mathbb{E}\left[\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2\right] \\
&+ 2\sum_{k=1}^{N} \eta_k^2(1+\sigma^2) + 4L_G^2 \sum_{k=0}^{N-1} \eta_{k+1} \mathbb{E}\left[\left\|\theta_k - \tilde{\theta}_k\right\|_2^2\right].
\end{aligned}$$

The last inequality follows by Lemma A.7. Combining (40), and (43), we get,

$$\begin{aligned}
\sum_{i=1}^{N} \eta_i \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_i) - \tilde{z}_i\right\|_2^2\right] \leq& \mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0) - \tilde{z}_0\right\|_2^2\right] + \sum_{k=1}^{N} \eta_k^2(1+\sigma^2) + 4L_G^2 \sum_{k=0}^{N-1} \eta_{k+1} \mathbb{E}\left[\left\|\theta_k - \tilde{\theta}_k\right\|_2^2\right] \\
&+ 4L_G^2 \left(W(x_0, z_0) + \sigma^2 \sum_{k=0}^{N} \eta_k^2 + \frac{L_G^2}{\beta} \sum_{k=0}^{\infty} \eta_{k+1} \mathbb{E}\left[\left\|\theta_k - \tilde{\theta}_k\right\|_2^2\right]\right).
\end{aligned} \tag{47}$$

Combining (47), and (40), we get,

$$\left(\sum_{k=1}^{N}\eta_k\right)\mathbb{E}\left[V(\tilde{\theta}_k,\tilde{z}_k)\right]\leq\mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0)-\tilde{z}_0\right\|_2^2\right]+4L_G^2W(x_0,z_0)+2\sum_{k=1}^{N}\eta_k^2(1+\sigma^2)$$
$$+(4L_G^2+4L_G^4/\beta)\sum_{k=0}^{N-1}\eta_{k+1}\mathbb{E}\left[\left\|\theta_k-\tilde{\theta}_k\right\|_2^2\right]. \tag{48}$$

Choosing $\eta_k=(N+k)^{-a}$, using Lemma A.3, for $\gamma=1/a-1$, we get,

$$\begin{aligned}
\sum_{k=0}^{N-1}\eta_{k+1}\mathbb{E}\left[\left\|\theta_k-\tilde{\theta}_k\right\|_2^2\right]&\leq\sum_{k=0}^{N-1}\eta_{k+1}\eta_k^3(1+\eta_k^{-\gamma})+\sum_{k=0}^{N-1}\eta_{k+1}\sum_{i=1}^{k-1}\eta_i^3\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k}\left(1+\eta_j^{1+\gamma}\right)\\
&\leq 2\sum_{k=0}^{N-1}(N+k+1)^{-a}\sum_{i=1}^{k-1}(N+i)^{-3a}\left(1+(N+i)^{a\gamma}\right)\prod_{j=i+1}^{k}\left(1+(N+j)^{-a(1+\gamma)}\right)\\
&\leq 2\sum_{k=0}^{N-1}N^{-4a}\sum_{i=1}^{k-1}\left(1+(2N)^{1-a}\right)\left(1+N^{-1}\right)^{k-i}\\
&\leq 2\sum_{k=0}^{N-1}N^{1-4a}\left(1+(2N)^{1-a}\right)\left((1+N^{-1})^k-1\right)\\
&\leq 8N^{2-5a}\left(N(1+N^{-1})^N-2N\right)\\
&\leq 8N^{3-5a}.
\end{aligned}$$

Hence, we have

$$\frac{\sum_{k=0}^{N-1}\eta_{k+1}\mathbb{E}\left[\left\|\theta_k-\tilde{\theta}_k\right\|_2^2\right]}{\sum_{k=1}^{N}\eta_k}\leq\frac{8N^{3-5a}}{\sum_{k=1}^{N}(2N)^{-a}}\leq 16N^{2-4a}. \tag{49}$$

Using (49), and (48), we get

$$\mathbb{E}\left[V(\tilde{\theta}_k,\tilde{z}_k)\right]\leq\left(\mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0)-\tilde{z}_0\right\|_2^2\right]+4L_G^2W(x_0,z_0)\right)N^{a-1}+2(1+\sigma^2)N^{-a}+16N^{2-4a}.$$

Then using Lemma A.4, we get,

$$\begin{aligned}
\mathbb{E}\left[V(\theta_k,z_k)\right]\leq&\left(\mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0)-\tilde{z}_0\right\|_2^2\right]+4L_G^2W(x_0,z_0)\right)N^{a-1}\\
&+2(1+\sigma^2)N^{-a}+(9+4L_G)\left(N^{1-4a}+8N^{2-4a}\right)+12N^{-2a}+16N^{2-4a}.
\end{aligned} \tag{50}$$

Now choosing, $a=3/5$, we get,

$$\begin{aligned}
\mathbb{E}\left[V(\theta_k,z_k)\right]\leq&\left(\mathbb{E}\left[\left\|\nabla f(\tilde{\theta}_0)-\tilde{z}_0\right\|_2^2\right]+4L_G^2W(x_0,z_0)\right)N^{-2/5}\\
&+2(1+\sigma^2)N^{-3/5}+(9+4L_G)\left(N^{-7/5}+8N^{-2/5}\right)+12N^{-6/5}+16N^{-2/5}\\
=&\mathcal{O}\left(N^{-\frac{2}{5}}\right).
\end{aligned}$$

■

Now we the provide the proofs of the Lemmas required to prove Theorem 2.1.

A.2 Proof of Lemmas related to Theorem 2.1

Proof. [Proof of Lemma A.3] By Jensen's inequality, contraction property of the projection operator, and Young's inequality, we get

$$\begin{aligned}\left\|\tilde{\theta}_{k+1}-\theta_{k+1}\right\|_2^2 \leq & (1-\eta_{k+1})\left\|\tilde{\theta}_k-\theta_k\right\|_2^2+\eta_{k+1}\left\|\tilde{y}_k-y_k\right\|_2^2 \\ \leq & (1-\eta_{k+1})\left\|\tilde{\theta}_k-\theta_k\right\|_2^2+\eta_{k+1}\left(\left\|\tilde{\theta}_k-\theta_k\right\|_2+\left\|\tilde{z}_k/\beta-z_k/\beta\right\|_2+\delta_k\right)^2 \\ \leq & (1-\eta_{k+1})\left\|\tilde{\theta}_k-\theta_k\right\|_2^2+\eta_{k+1}(1+\eta_{k+1}^{\gamma})\left\|\tilde{\theta}_k-\theta_k\right\|_2^2 \\ & +2\eta_{k+1}(1+\eta_{k+1}^{-\gamma})\left\|\tilde{z}_k/\beta-z_k/\beta\right\|_2^2+2\eta_{k+1}(1+\eta_{k+1}^{-\gamma})\delta_k^2.\end{aligned}$$

Now taking expectation on both sides, and using Lemma 2.1, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\theta}_{k+1}-\theta_{k+1}\right\|_2^2\right] \leq & \left(1+\eta_{k+1}^{1+\gamma}\right)\mathbb{E}\left[\left\|\tilde{\theta}_k-\theta_k\right\|_2^2\right]+2\eta_{k+1}^3(1+\eta_{k+1}^{-\gamma})+2\eta_{k+1}(1+\eta_{k+1}^{-\gamma})\mathbb{E}\left[\delta_k^2\right] \\ \leq & \eta_{k+1}^3(1+\eta_{k+1}^{-\gamma})+2\sum_{i=1}^{k}\eta_i^3\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k+1}\left(1+\eta_j^{1+\gamma}\right) \\ & +2\sum_{i=1}^{k}\eta_i\mathbb{E}\left[\delta_{i-1}^2\right]\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k+1}\left(1+\eta_j^{1+\gamma}\right).\end{aligned}$$

■

Proof. [Proof of Lemma A.4] Using (17), and contraction property of the projection operator,

$$\begin{aligned}V(\theta_k,z_k) = & \left\|\Pi_\Theta\left(\theta_k-z_k\right)-\theta_k\right\|_2^2+\left\|z_k-\nabla f(\theta_k)\right\|_2^2 \\ \leq & 2\left\|\Pi_\Theta\left(\theta_k-z_k\right)-\theta_k-\Pi_\Theta\left(\tilde{\theta}_k-\tilde{z}_k\right)+\tilde{\theta}_k\right\|_2^2+2\left\|\tilde{z}_k-\nabla f(\tilde{\theta}_k)-z_k+\nabla f(\theta_k)\right\|_2^2 \\ & +2\left\|\Pi_\Theta\left(\tilde{\theta}_k-\tilde{z}_k\right)-\tilde{\theta}_k\right\|_2^2+2\left\|\tilde{z}_k-\nabla f(\tilde{\theta}_k)\right\|_2^2 \\ \leq & 2V(\tilde{\theta}_k,\tilde{z}_k)+(8+4L_G)\left\|\theta_k-\tilde{\theta}_k\right\|_2^2+12\left\|z_k-\tilde{z}_k\right\|_2^2.\end{aligned}$$

Using Lemma A.3, and Lemma 2.1, we get,

$$\begin{aligned}\mathbb{E}\left[V(\theta_k,z_k)\right] \leq & 2\mathbb{E}\left[V(\tilde{\theta}_k,\tilde{z}_k)\right]+(8+4L_G)\mathbb{E}\left[\left\|\theta_k-\tilde{\theta}_k\right\|_2^2\right]+12\mathbb{E}\left[\left\|z_k-\tilde{z}_k\right\|_2^2\right] \\ \leq & 2\mathbb{E}\left[V(\tilde{\theta}_k,\tilde{z}_k)\right]+(8+4L_G)\left(\eta_k^3(1+\eta_k^{-\gamma})+2\sum_{i=1}^{k-1}\eta_i^3\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k}\left(1+\eta_j^{1+\gamma}\right)\right. \\ & \left.+2\sum_{i=1}^{k-1}\eta_i\mathbb{E}\left[\delta_{i-1}^2\right]\left(1+\eta_i^{-\gamma}\right)\prod_{j=i+1}^{k}\left(1+\eta_j^{1+\gamma}\right)\right)+12\eta_{k+1}^2.\end{aligned}$$

For $\delta_{k-1}\leq\eta_k$, choosing $\eta_k=\frac{1}{(N+k)^a}$ with $a>1/2$, we get,

$$\mathbb{E}\left[V(\theta_k,z_k)\right] \leq 2\mathbb{E}\left[V(\tilde{\theta}_k,\tilde{z}_k)\right]+\frac{12}{(N+k+1)^{2a}}$$

$$
\begin{aligned}
&+ (8 + 4L_G) \left(\frac{1 + (N+k)^{a\gamma}}{(N+k)^{3a}} + 4 \sum_{i=1}^{k-1} \frac{1 + (N+i)^{a\gamma}}{(N+i)^{3a}} \prod_{j=i+1}^{k} \left(1 + (N+j)^{-a(1+\gamma)}\right) \right) \\
\leq & 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9 + 4L_G) \left(\frac{1}{N^{3a-a\gamma}} + \sum_{i=1}^{k-1} \frac{4}{N^{3a-a\gamma}} \left(1 + \frac{1}{N^{a(1+\gamma)}}\right)^i \right) + \frac{12}{N^{2a}} \\
\leq & 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9 + 4L_G) \left(\frac{1}{N^{3a-a\gamma}} + \frac{4}{N^{2a-2a\gamma}} \left[\left(1 + \frac{1}{N^{a(1+\gamma)}}\right)^N - 1 \right] \right) + \frac{12}{N^{2a}} \\
\leq & 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9 + 4L_G) \left(\frac{1}{N^{3a-a\gamma}} + \frac{4}{N^{2a-2a\gamma}} \left[\exp\left(N^{1-a(1+\gamma)}\right) - 1 \right] \right) + \frac{12}{N^{2a}} \\
\leq & 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9 + 4L_G) \left(\frac{1}{N^{3a-a\gamma}} + 8N^{1-3a+a\gamma} \right) + \frac{12}{N^{2a}}. \qquad (51)
\end{aligned}
$$

Setting $\gamma = 1/a - 1$, we get,

$$
\mathbb{E}\left[V(\theta_k, z_k)\right] \leq 2\mathbb{E}\left[V(\tilde{\theta}_k, \tilde{z}_k)\right] + (9 + 4L_G)\left(N^{1-4a} + 8N^{2-4a}\right) + 12N^{-2a}.
$$

■

Proof. [Proof of Lemma A.5]
Recall that,

$$
\phi(\theta, z) = \min_{y \in \Theta} \left(\langle z, y - \theta \rangle + \frac{\beta}{2} \|y - \theta\|_2^2 \right). \qquad (52)
$$

It is easy to verify that $\phi(\theta, z)$ has a L_ϕ-Lipschitz continuous gradient [GRW20, Lemma 3] where

$$
L_\phi = 2\sqrt{(1+\beta)^2 + (1 + 1/(2\beta))^2}.
$$

Using the definition of $\phi(\theta, z)$ in (52), and Lipschitz continuity of its gradient, we have

$$
\begin{aligned}
\phi(\tilde{\theta}_k, \tilde{z}_k) - \phi(\tilde{\theta}_{k+1}, \tilde{z}_{k+1}) \leq & \left\langle \tilde{z}_k + \beta(\tilde{y}_k - \tilde{\theta}_k), \tilde{\theta}_{k+1} - \tilde{\theta}_k \right\rangle - \left\langle \tilde{y}_k - \tilde{\theta}_k, \tilde{z}_{k+1} - \tilde{z}_k \right\rangle \\
& + \frac{L_\phi}{2} \left[\left\| \tilde{\theta}_{k+1} - \tilde{\theta}_k \right\|_2^2 + \|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2 \right]. \qquad (53)
\end{aligned}
$$

By the optimality condition of the subproblem (32) we have,

$$
\left\langle \tilde{z}_k + \beta(\tilde{y}_k - \tilde{\theta}_k), y - \tilde{y}_k \right\rangle \geq 0 \quad \forall y \in \Theta. \qquad (54)
$$

For $y = \tilde{\theta}_k$ we have,

$$
\left\langle \tilde{z}_k + \beta(\tilde{y}_k - \tilde{\theta}_k), \tilde{y}_k - \tilde{\theta}_k \right\rangle \leq 0. \qquad (55)
$$

Note that this also implies

$$
\phi(\tilde{\theta}_k, \tilde{z}_k) \leq 0. \qquad (56)
$$

We also have,

$$
\tilde{z}_{k+1} - \tilde{z}_k = \tilde{z}_{k+1} - (1 - \eta_{k+1})\tilde{z}_k - \eta_{k+1}\tilde{z}_k
$$

$$
\begin{aligned}
&= z_{k+1} - (1-\eta_{k+1})z_k - \eta_{k+1}\tilde{z}_k + \tilde{\zeta}_{k+2} - (1-\eta_{k+1})\tilde{\zeta}_{k+1} \\
&= \eta_{k+1}\left(\nabla f(\theta_k) + e_{k+1} + \nu_{k+1} + \zeta_{k+1}\right) - \eta_{k+1}\tilde{z}_k + \tilde{\zeta}_{k+2} - (1-\eta_{k+1})\tilde{\zeta}_{k+1} \\
&= \eta_{k+1}\left(\nabla f(\theta_k) + e_{k+1} + \nu_{k+1}\right) + (\tilde{\zeta}_{k+1} - \tilde{\zeta}_{k+2}) - \eta_{k+1}\tilde{z}_k + \tilde{\zeta}_{k+2} - (1-\eta_{k+1})\tilde{\zeta}_{k+1} \\
&= \eta_{k+1}\left(\nabla f(\theta_k) + \tilde{\epsilon}_{k+1}\right) - \eta_{k+1}\tilde{z}_k,
\end{aligned}
$$

where, $\tilde{\epsilon}_k = e_k + \nu_k + \tilde{\zeta}_k$. Then, using (17) we have,

$$
\begin{aligned}
\left\langle \tilde{y}_k - \tilde{\theta}_k, \tilde{z}_{k+1} - \tilde{z}_k \right\rangle &= \left\langle \tilde{y}_k - \tilde{\theta}_k, \eta_{k+1}\left(\nabla f(\theta_k) + \tilde{\epsilon}_{k+1}\right) - \eta_{k+1}\tilde{z}_k \right\rangle \\
&= \left\langle \tilde{\theta}_{k+1} - \tilde{\theta}_k, \nabla f(\tilde{\theta}_k) \right\rangle + \left\langle \tilde{\theta}_{k+1} - \tilde{\theta}_k, \nabla f(\theta_k) - \nabla f(\tilde{\theta}_k) \right\rangle \\
&\quad + \left\langle \tilde{y}_k - \tilde{\theta}_k, \eta_{k+1}\tilde{\epsilon}_{k+1} \right\rangle - \left\langle \tilde{\theta}_{k+1} - \tilde{\theta}_k, \tilde{z}_k \right\rangle \\
&\geq f(\tilde{\theta}_{k+1}) - f(\tilde{\theta}_k) - \frac{L_G}{2}\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\|_2^2 - \frac{\beta}{2\eta_{k+1}}\left\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\right\|_2^2 \\
&\quad + \left\langle \tilde{y}_k - \tilde{\theta}_k, \eta_{k+1}\tilde{\epsilon}_{k+1} \right\rangle - \left\langle \tilde{\theta}_{k+1} - \tilde{\theta}_k, \tilde{z}_k \right\rangle . \qquad (57)
\end{aligned}
$$

Combining (53), (54), (55), and (57), using (17), and rearranging, we get,

$$
\begin{aligned}
&\phi(\tilde{\theta}_k, \tilde{z}_k) - \phi(\tilde{\theta}_{k+1}, \tilde{z}_{k+1}) \\
\leq& f(\tilde{\theta}_k) - f(\tilde{\theta}_{k+1}) + \frac{L_G}{2}\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\|_2^2 + \frac{\beta}{2\eta_{k+1}}\left\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\right\|_2^2 + \frac{\eta_{k+1}}{\beta}\left\|\nabla f(\theta_k) - \nabla f(\tilde{\theta}_k)\right\|_2^2 \\
&- \left\langle \tilde{y}_k - \tilde{\theta}_k, \eta_{k+1}\tilde{\epsilon}_{k+1} \right\rangle - \eta_{k+1}\beta\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 + \frac{L_\phi}{2}\left[\left\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\right\|_2^2 + \|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2\right] \\
&W(\tilde{\theta}_{k+1}, \tilde{z}_{k+1}) - W(\tilde{\theta}_k, \tilde{z}_k) \\
\leq& -\frac{\eta_{k+1}\beta}{2}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 + \frac{(L_G + L_\phi)\eta_{k+1}^2}{2}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 + \frac{L_\phi}{2}\|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2 + \frac{\eta_{k+1}L_G^2}{\beta}\left\|\theta_k - \tilde{\theta}_k\right\|_2^2 \\
&- \eta_{k+1}\left\langle \tilde{y}_k - \tilde{\theta}_k, \tilde{\epsilon}_{k+1} \right\rangle
\end{aligned}
$$

Summing both sides from $k = 0$ to $N-1$, and using (56), we get,

$$
\sum_{k=0}^{i} \frac{\eta_{k+1}\beta}{2}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 \leq W(\tilde{\theta}_0, \tilde{z}_0) + \sum_{k=0}^{N-1} r_{k+1},
$$

where

$$
r_{k+1} = \frac{(L_G + L_\phi)\eta_{k+1}^2}{2}\left\|\tilde{y}_k - \tilde{\theta}_k\right\|_2^2 + \frac{L_\phi}{2}\|\tilde{z}_{k+1} - \tilde{z}_k\|_2^2 + \eta_{k+1}\left\langle \tilde{\theta}_k - \tilde{y}_k, \tilde{\epsilon}_{k+1} \right\rangle + \frac{\eta_{k+1}L_G^2}{\beta}\left\|\theta_k - \tilde{\theta}_k\right\|_2^2.
$$

■

Proof. [Proof of Lemma A.6] The proof is similar to that of [GRW20, Proposition 1] with minor modifications. First, note that we no longer have $\mathbb{E}\left[(\tilde{\theta}_k - \tilde{y}_k)^\top \tilde{\epsilon}_k | \mathcal{F}_{k-1}\right] = 0$ since $\{\tilde{\epsilon}_k\}_k$ is no longer a martingale difference sequence. But we can show that the term is small enough, i.e., of the order of the stepsize. Indeed note that, using (41), we have

$$
\begin{aligned}
\mathbb{E}\left[(\tilde{\theta}_k - \tilde{y}_k)^\top \tilde{\epsilon}_k | \mathcal{F}_{k-1}\right] &= \mathbb{E}\left[(\tilde{\theta}_k - \tilde{y}_k)^\top (\nu_k + \tilde{\zeta}_k) | \mathcal{F}_{k-1}\right] \\
&\leq \sqrt{\mathbb{E}\left[\|\tilde{\theta}_k - \tilde{y}_k\|_2^2 | \mathcal{F}_{k-1}\right]}\sqrt{\mathbb{E}\left[\|\nu_k + \tilde{\zeta}_k\|_2^2 | \mathcal{F}_{k-1}\right]}
\end{aligned}
$$

$$\leq \sqrt{\mathbb{E}\left[\frac{2\|\tilde{z}_k\|_2^2}{\beta^2}|\mathcal{F}_{k-1}\right]}\eta_k$$
$$\leq \frac{2\sigma\eta_k}{\beta}. \tag{58}$$

Combining (58) with the proof of [GRW20, Proposition 1] we obtain the result in Lemma A.6. ∎

Lemma A.7. *Let Y_k^i be defined as in (46) and let η_j be of the form $(N+j)^{-a}$ where $1/2 < a < 1$. Then, for $k \leq i \leq N$, we have*

$$Y_{k-1}^i \leq \exp\left(-\frac{1}{8}\left((N+i)^{1-a} - (N+k)^{1-a}\right)\right)$$
$$\sum_{i=k}^{N} Y_{k-1}^i \eta_i = \mathcal{O}(1). \tag{59}$$

Proof. First, using the fact that $1 - x \leq \exp(-x)$, we obtain

$$Y_{k-1}^i = \prod_{j=k}^{i}\left(1 - \frac{\eta_j}{4}\right) \leq \prod_{j=k}^{i}\exp\left(-\frac{\eta_j}{4}\right) = \exp\left(-\sum_{j=k}^{i}\frac{(N+j)^{-a}}{4}\right)$$
$$\leq \exp\left(-\int_k^i \frac{(N+j)^{-a}}{8}\right)dj$$
$$\leq \exp\left(-\frac{1}{8}\left((N+i)^{1-a} - (N+k)^{1-a}\right)\right),$$

which proves the first claim. Next, we have

$$\sum_{i=k}^{N} Y_{k-1}^i \eta_i \leq \sum_{i=k}^{N}\exp\left(-\frac{1}{8}\left((N+i)^{1-a} - (N+k)^{1-a}\right)\right)(N+i)^{-a}$$
$$= \exp\left(\frac{(N+k)^{1-a}}{8}\right)\sum_{i=k+N}^{2N}\exp\left(-\frac{i^{1-a}}{8}\right)i^{-a}$$
$$\leq \exp\left(\frac{(N+k)^{1-a}}{8}\right)\int_{k+N}^{2N}\exp\left(-\frac{(i-1)^{1-a}}{8}\right)(i-1)^{-a}di$$
$$= \frac{\exp\left(\frac{(N+k)^{1-a}}{8}\right)}{1-a}\int_{(k+N-1)^{1-a}}^{(2N-1)^{1-a}}\exp(-u)\,du$$
$$\leq \frac{\exp\left(\frac{(N+k)^{1-a}}{8} - \frac{(N+k-1)^{1-a}}{8}\right)}{1-a}$$
$$\leq \frac{e}{1-a},$$

which proves the second claim. ∎

B Proofs for the State-independent Case

Before proving the proof we present some notations. Recall that,

$$\phi(\theta, z) = \min_{y\in\Theta}\left(\langle z, y-\theta\rangle + \frac{\beta}{2}\|y-\theta\|_2^2\right),$$

and

$$y'_k = \underset{y\in\Theta}{\operatorname{argmin}} \left\{ \langle z_k, y - \theta_k\rangle + \frac{\beta}{2}\|y-\theta_k\|_2^2 \right\}.$$

For a given θ, and z, we introduce the following notation for convenience.

$$H(y) = \langle z, y-\theta\rangle + \frac{\beta}{2}\|y-\theta\|_2^2$$

Noting that $H(y)$ is β-strongly convex, we have,

$$\frac{\beta}{2}\|y_k - y'_k\|_2^2 \le H(y_k) - H(y'_k).$$

We choose the parameters of Algorithm 2 such that

$$H(y_k) - H(y'_k) \le \delta_k^2.$$

The specific choice of δ_k will be set later. Let us define the following merit function.

$$W(\theta, z) = (f(\theta) - f^*) - \phi(\theta, z) + \alpha_1 \|\nabla f(\theta) - z\|_2^2, \quad \alpha_1 > 0, \tag{60}$$

where $f^* > -\infty$ is a uniform lower bound on the function f. We also need the following result from [AMP05] on mixing properties of the data under Assumption 2.4 (a).

Lemma B.1. *[AMP05] Let Assumption 2.4 (a) be true. Then, for any $\theta \in \Theta$, the chain $\{x_k\}_k$ is exponentially mixing in the sense of Definition 26.*

Proof. [Proof of Theorem 2.2] First we establish recursion relations on the three components of $W(\theta, z)$: $(f(\theta) - f^*)$, $\phi(\theta, z)$, and $\alpha_1\|\nabla f(\theta) - z\|_2^2$. Using (17), Assumption 2.1, Young's inequality,

$$\begin{aligned}
f(\theta_{k+1}) - f(\theta_k) \le& \nabla f(\theta_k)^\top(\theta_{k+1} - \theta_k) + \frac{L_G}{2}\|\theta_{k+1} - \theta_k\|_2^2 \\
=& \eta_{k+1}\nabla f(\theta_k)^\top(y'_k - \theta_k) + \eta_{k+1}(\nabla f(\theta_k) - z_k)^\top(y_k - y'_k) \\
&+ \eta_{k+1}(z_k + \beta(y'_k - \theta))^\top(y_k - y'_k) - \eta_{k+1}\beta\langle y'_k - \theta_k, y_k - y'_k\rangle + \frac{L_G\mathcal{D}_\Theta^2\eta_{k+1}^2}{2} \\
\le& \eta_{k+1}\left(H(y_k) - H(y'_k) - \frac{\beta}{2}\|y_k - y'_k\|_2^2\right) + \frac{\eta_{k+1}\beta}{16}\|y'_k - \theta_k\|_2^2 + 4\eta_{k+1}\beta\|y_k - y'_k\|_2^2 \\
&+ \frac{L_G\mathcal{D}_\Theta^2\eta_{k+1}^2}{2} + \frac{\eta_{k+1}\beta}{16}\|\nabla f(\theta_k) - z_k\|_2^2 + \frac{4\eta_{k+1}}{\beta}\|y_k - y'_k\|_2^2 + \eta_{k+1}\nabla f(\theta_k)^\top(y'_k - \theta_k) \\
\le& \eta_{k+1}\left(H(y_k) - H(y'_k)\right) + \frac{\eta_{k+1}\beta}{16}\|y'_k - \theta_k\|_2^2 + 4\eta_{k+1}\beta\|y_k - y'_k\|_2^2 + \frac{L_G\mathcal{D}_\Theta^2\eta_{k+1}^2}{2} \\
&+ \frac{\eta_{k+1}\beta}{16}\|\nabla f(\theta_k) - z_k\|_2^2 + \frac{4\eta_{k+1}}{\beta}\|y_k - y'_k\|_2^2 + \eta_{k+1}\nabla f(\theta_k)^\top(y'_k - \theta_k).
\end{aligned} \tag{61}$$

Using (53),

$$\begin{aligned}
&\phi(\theta_k, z_k) - \phi(\theta_{k+1}, z_{k+1}) \\
\le& \langle z_k + \beta(y'_k - \theta_k), \theta_{k+1} - \theta_k\rangle - \langle y'_k - \theta_k, z_{k+1} - z_k\rangle + \frac{L_\phi}{2}\left[\|\theta_{k+1} - \theta_k\|_2^2 + \|z_{k+1} - z_k\|_2^2\right] \\
\le& \eta_{k+1}\langle z_k + \beta(y'_k - \theta_k), y'_k - \theta_k\rangle + \eta_{k+1}\langle z_k + \beta(y'_k - \theta_k), y_k - y'_k\rangle - \langle y'_k - \theta_k, z_{k+1} - z_k\rangle
\end{aligned}$$

$$
\begin{aligned}
&+\frac{L_\phi}{2}\left[\|\theta_{k+1}-\theta_k\|_2^2+\|z_{k+1}-z_k\|_2^2\right]\\
\le&\eta_{k+1}\left(H(y_k)-H(y_k')-\frac{\beta}{2}\left\|y_k-y_k'\right\|_2^2\right)-\eta_{k+1}\left\langle y_k'-\theta_k,\nabla F(\theta_k,x_{k+1})\right\rangle\\
&+\eta_{k+1}\left\langle y_k'-\theta_k,z_k\right\rangle+\frac{L_\phi}{2}\left[\|\theta_{k+1}-\theta_k\|_2^2+\|z_{k+1}-z_k\|_2^2\right]\\
\le&-\eta_{k+1}\beta\left\|y_k'-\theta_k\right\|_2^2+\eta_{k+1}\left(H(y_k)-H(y_k')\right)-\eta_{k+1}\left\langle y_k'-\theta_k,\nabla f(\theta_k)\right\rangle\\
&-\eta_{k+1}\left\langle y_k'-\theta_k,\xi_{k+1}(\theta_k,x_{k+1})\right\rangle+\frac{L_\phi}{2}\left[\|\theta_{k+1}-\theta_k\|_2^2+\|z_{k+1}-z_k\|_2^2\right].
\end{aligned}\tag{62}
$$

Recall Γ_i defined in (44). Then

$$
\begin{aligned}
\nabla f(\theta_i)-z_i=&\frac{\Gamma_i}{\Gamma_{i-1}}\left(\nabla f(\theta_{i-1})-z_{i-1}\right)+\left(\nabla f(\theta_i)-\nabla f(\theta_{i-1})\right)-\eta_i\left(\tilde\epsilon_i+\nabla f(\theta_{i-1})-\nabla f(\theta_{i-1})\right)\\
=&(1-\eta_i)\left(\nabla f(\theta_{i-1})-z_{i-1}\right)+\frac{\eta_i}{\eta_i}\left(\nabla f(\theta_i)-\nabla f(\theta_{i-1})\right)-\eta_i\xi_i.
\end{aligned}
$$

Using Jensen's inequality,

$$
\begin{aligned}
\|\nabla f(\theta_i)-z_i\|_2^2\le&(1-\eta_i)\|\nabla f(\theta_{i-1})-z_{i-1}\|_2^2+\frac{1}{\eta_i}\|\nabla f(\theta_i)-\nabla f(\theta_{i-1})\|_2^2+\eta_i^2\|\xi_i\|_2^2\\
&-2\eta_i\left\langle\xi_i,(1-\eta_i)\left(\nabla f(\theta_{i-1})-z_{i-1}\right)+\left(\nabla f(\theta_i)-\nabla f(\theta_{i-1})\right)\right\rangle\\
\le&(1-\eta_i)\|\nabla f(\theta_{i-1})-z_{i-1}\|_2^2+2L_G^2\eta_i\left\|y_{i-1}'-\theta_{i-1}\right\|_2^2+2L_G^2\eta_i\left\|y_{i-1}-y_{i-1}'\right\|_2^2\\
&+\eta_i^2\|\xi_i\|_2^2-2\eta_i\left\langle\xi_i,(1-\eta_i)\left(\nabla f(\theta_{i-1})-z_{i-1}\right)+\left(\nabla f(\theta_i)-\nabla f(\theta_{i-1})\right)\right\rangle.
\end{aligned}\tag{63}
$$

Combining (61), (62), and (63) we have,

$$
\begin{aligned}
&W(\theta_{k+1},z_{k+1})-W(\theta_k,z_k)\\
=&f(\theta_{k+1})-f(\theta_k)-\phi(\theta_{k+1},z_{k+1})+\phi(\theta_k,z_k)+\alpha_1\|\nabla f(\theta_{k+1})-z_{k+1}\|_2^2-\alpha_1\|\nabla f(\theta_k)-z_k\|_2^2\\
\le&2\eta_{k+1}\left(H(y_k)-H(y_k')\right)-\frac{15\alpha_1\eta_{k+1}}{16}\|\nabla f(\theta_k)-z_k\|_2^2-\left(\frac{15\beta\eta_{k+1}}{16}-2\alpha_1L_G^2\eta_{k+1}\right)\left\|y_k'-\theta_k\right\|_2^2\\
&+\eta_{k+1}\left(4\beta+4/\alpha_1+2L_G^2\alpha_1\right)\|y_k-y_k'\|_2^2\\
&+\eta_{k+1}^2\left(\frac{L_GD_\Theta^2}{2}+\frac{L_\phi D_\Theta^2}{2}+\|z_k-\nabla F(\theta_k,x_{k+1})\|_2^2+\|\xi_{k+1}(\theta_k,x_{k+1})\|_2^2+2\|\xi_{k+1}(\theta_k,x_{k+1})\|_2\|\nabla f(\theta_k)-z_k\|_2\right)\\
&-\eta_{k+1}\left\langle y_k'-\theta_k,\xi_{k+1}(\theta_k,x_{k+1})\right\rangle-2\eta_{k+1}\left\langle\xi_{k+1}(\theta_k,x_{k+1}),\nabla f(\theta_{k+1})-z_k\right\rangle.
\end{aligned}
$$

Rearranging, and choosing $\alpha_1=\beta/(32L_G^2)$ we get,

$$
\begin{aligned}
&\frac{14\beta\eta_{k+1}}{16}\left\|y_k'-\theta_k\right\|_2^2+\frac{15\beta\eta_{k+1}}{512L_G^2}\|\nabla f(\theta_k)-z_k\|_2^2\le W(\theta_k,z_k)-W(\theta_{k+1},z_{k+1})\\
&+\left(4/\beta+4\beta+4/\alpha_1+2L_G^2\alpha_1\right)\eta_{k+1}\delta_k^2+\eta_{k+1}^2U_k-\eta_{k+1}S_k-\eta_{k+1}Q_k,
\end{aligned}\tag{64}
$$

where

$$
\begin{aligned}
U_k=&\frac{L_GD_\Theta^2}{2}+\frac{L_\phi D_\Theta^2}{2}+\|z_k-\nabla F(\theta_k,x_{k+1})\|_2^2+\|\xi_{k+1}(\theta_k,x_{k+1})\|_2^2\\
&+2\|\xi_{k+1}(\theta_k,x_{k+1})\|_2\|\nabla f(\theta_k)-z_k\|_2,\\
S_k=&\left\langle y_k'-\theta_k,\xi_{k+1}(\theta_k,x_{k+1})\right\rangle,
\end{aligned}
$$

$$Q_k = 2\left\langle \xi_{k+1}(\theta_k, x_{k+1}), \nabla f(\theta_{k+1}) - z_k \right\rangle.$$

Taking expectation on both sides and summing from $k=0$ to $k=N$, we get,

$$\sum_{k=0}^{N} \mathbb{E}\left[\frac{14\beta\eta_{k+1}}{16}\left\|y'_k - \theta_k\right\|_2^2 + \frac{15\beta\eta_{k+1}}{512L_G^2}\left\|\nabla f(\theta_k) - z_k\right\|_2^2\right] \leq W(\theta_0, z_0)$$
$$+\sum_{k=0}^{N}\left(4/\beta + 4\beta + 4/\alpha_1 + 2L_G^2\alpha_1\right)\eta_{k+1}\delta_k^2 + \sum_{k=0}^{N}\eta_{k+1}^2\mathbb{E}\left[U_k\right] - \sum_{k=0}^{N}\eta_{k+1}(\mathbb{E}\left[S_k\right] + \mathbb{E}\left[Q_k\right]), \tag{65}$$

Bound on $\mathbb{E}\left[U_k\right]$: Similar to (41), we have $\mathbb{E}\left[\|z_k\|_2\right] \leq \sigma$. Using Lipschitz continuity of $f(\cdot)$, as explained in Section 2, we have $\nabla f(\theta_k) \leq L$. Combining these with Assumption 2.3, we have,

$$\mathbb{E}\left[U_k\right] = \mathcal{O}(1) \tag{66}$$

Bound on $\mathbb{E}\left[S_k\right]$: Using the fact that $\mathbb{E}_{x\sim\pi}\left[\left\langle y'_{k-l} - \theta_{k-l}, \xi_{k+1}(\theta_{k-l}, x)\right\rangle\right] = 0$, for $l \in \{1, \cdots, k-1\}$, we have

$$\begin{aligned}
\mathbb{E}\left[S_k|\mathcal{F}_{k-l}\right] =& \mathbb{E}\left[\left\langle y'_k - \theta_k, \xi_{k+1}(\theta_k, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] - \mathbb{E}\left[\left\langle y'_k - \theta_k, \xi_{k+1}(\theta_{k-l}, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] \\
&+ \mathbb{E}\left[\left\langle y'_k - \theta_k, \xi_{k+1}(\theta_{k-l}, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] - \mathbb{E}\left[\left\langle y'_{k-l} - \theta_{k-l}, \xi_{k+1}(\theta_{k-l}, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] \\
&+ \mathbb{E}\left[\left\langle y'_{k-l} - \theta_{k-l}, \xi_{k+1}(\theta_{k-l}, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] - \mathbb{E}_{x\sim\pi}\left[\left\langle y'_{k-l} - \theta_{k-l}, \xi_{k+1}(\theta_{k-l}, x)\right\rangle\right] \\
=& \mathbb{E}\left[\left\langle y'_k - \theta_k, \sum_{i=k-l+1}^{k}\left(\xi_{k+1}(\theta_i, x_{k+1}) - \xi_{k+1}(\theta_{i-1}, x_{k+1})\right)\right\rangle|\mathcal{F}_{k-l}\right] \\
&+ \mathbb{E}\left[\left\langle \sum_{i=k-l+1}^{k}(y'_i - \theta_i - y'_{i-1} + \theta_{i-1}), \xi_{k+1}(\theta_{k-l}, x_{k+1})\right\rangle|\mathcal{F}_{k-l}\right] \\
&+ \left\langle y'_{k-l} - \theta_{k-l}, \mathbb{E}\left[\xi_{k+1}(\theta_{k-l}, x_{k+1})|\mathcal{F}_{k-l}\right]\right\rangle - \left\langle y'_{k-l} - \theta_{k-l}, \mathbb{E}_{x\sim\pi}\left[\xi_{k+1}(\theta_{k-l}, x)\right]\right\rangle \\
=& \mathbb{E}\left[\left\|y'_k - \theta_k\right\|_2 \sum_{i=k-l+1}^{k}\eta_i\left\|y_{i-1} - \theta_{i-1}\right\|_2|\mathcal{F}_{k-l}\right] \\
&+ \mathbb{E}\left[\sum_{i=k-l+1}^{k}\left(\left\|z_i - z_{i-1}\right\|_2/\beta + 2\left\|\theta_i - \theta_{i-1}\right\|_2\right)\left\|\xi_{k+1}(\theta_{k-l}, x_{k+1})\right\|_2|\mathcal{F}_{k-l}\right] \\
&+ \left\|y'_{k-l} - \theta_{k-l}\right\|_2\left\|\mathbb{E}\left[\xi_{k+1}(\theta_{k-l}, x_{k+1})|\mathcal{F}_{k-l}\right] - \mathbb{E}_{x\sim\pi}\left[\xi_{k+1}(\theta_{k-l}, x)\right]\right\|_2.
\end{aligned} \tag{67}$$

Using Assumption 2.1 one has,

$$\mathbb{E}\left[\left\|y'_k - \theta_k\right\|_2 \sum_{i=k-l+1}^{k}\eta_i\left\|y_{i-1} - \theta_{i-1}\right\|_2\right] = \mathcal{O}(l\eta_{k-l+1}). \tag{68}$$

Using Assumption 2.1, Assumption 2.3, $z_{k+1} - z_k = \eta_{k+1}(\nabla F(\theta_k, x_{k+1}) - z_k)$, and $\mathbb{E}\left[\|z_k\|_2\right] \leq \sigma$ one has that

$$\mathbb{E}\left[\sum_{i=k-l+1}^{k}\left(\left\|z_i - z_{i-1}\right\|_2/\beta + 2\left\|\theta_i - \theta_{i-1}\right\|_2\right)\left\|\xi_{k+1}(\theta_{k-l}, x_{k+1})\right\|_2\right] = \mathcal{O}\left(l\eta_{k-l+1}\right). \tag{69}$$

Using Assumption 2.1, Assumption 2.4, Lemma B.1, (26), and Lipschitz continuity of $f(\cdot)$, we have,

$$\left\|y'_{k-l} - \theta_{k-l}\right\|_2\left\|\mathbb{E}\left[\xi_{k+1}(\theta_{k-l}, x_{k+1})|\mathcal{F}_{k-l}\right] - \mathbb{E}_{x\sim\pi}\left[\xi_{k+1}(\theta_{k-l}, x)\right]\right\|_2 \leq \mathcal{O}(exp(-rl)), \tag{70}$$

where r is as in (26). Combining (68), (69), and (70) with (67) we get,

$$\mathbb{E}\left[S_k\right] = \mathcal{O}(l\eta_{k-l+1} + exp(-rl)). \tag{71}$$

Bound on $\mathbb{E}\left[Q_k\right]$: Following similar techniques used to establish bound on $\mathbb{E}\left[S_k\right]$, we have,

$$\mathbb{E}\left[Q_k\right] = \mathcal{O}(l\eta_{k-l+1} + exp(-rl)) \tag{72}$$

Combining (66), (71), and (72) with (64), choosing $t_k = \lceil\sqrt{k}\rceil$ to ensure $\delta_k^2 = \eta_{\|}$, setting $l = \lceil\frac{\log(1/\eta_{k-l+1})}{r}\rceil$, and choosing $\eta_k = (N+k)^{-a}$, $1/2 < a < 1$, we get,

$$\sum_{k=0}^{N}\mathbb{E}\left[\frac{14\beta\eta_{k+1}}{16}\left\|y_k' - \theta_k\right\|_2^2 + \frac{15\beta\eta_{k+1}}{512L_G^2}\left\|\nabla f(\theta_k) - z_k\right\|_2^2\right] \leq W(\theta_0, z_0) + \mathcal{O}\left(N^{1-2a}\log N\right), \tag{73}$$

Dividing both sides by $\sum_{k=0}^{N}\eta_k$, and choosing $a = 1/2$ we get,

$$\mathbb{E}\left[V(\theta_R, z_R)\right] = \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right).$$

■

Acknowledgements

AR was affiliated with the Department of Statistics, UC Davis while this work was completed and was partially supported by National Science Foundation (NSF) grant CCF-1934568. KB was partially supported by a seed grant from the Center for Data Science and Artificial Intelligence Research, UC Davis and NSF Grant-2053918. SG was partially supported by an NSERC Discovery Grant.

References

[ABRW12] Alekh Agarwal, Peter Bartlett, Pradeep Ravikumar, and Martin Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012. 1

[ACD+19] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019. 1

[AD12] Alekh Agarwal and John C Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2012. 2

[AL22] Ahmet Alacaoglu and Hanbaek Lyu. Convergence and complexity of stochastic subgradient methods with dependent data for nonconvex optimization. *arXiv preprint arXiv:2203.15797*, 2022. 5, 8, 11

[AMP05] Christophe Andrieu, Éric Moulines, and Pierre Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization*, 44(1):283–312, 2005. 2, 5, 8, 14, 24

[Bar92] Peter L Bartlett. Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 243–252, 1992. 1

[BG22] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022. 3

[BJMO12] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012. 1

[BMP12] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012. 5

[Bor09] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009. 5

[BRS18] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018. 5

[BS17] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017. 5

[CDP15] Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*, pages 280–296. PMLR, 2015. 1

[DAJJ12] John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012. 5

[DDDJ19] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. Learning from weakly dependent data under dobrushin's condition. In *Conference on Learning Theory*, pages 914–928. PMLR, 2019. 2

[DL22] Ron Dorfman and Kfir Y Levy. Adapting to mixing time in stochastic optimization with Markovian data. *arXiv preprint arXiv:2202.04428*, 2022. 5

[DMN+21] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with Markovian noise: Application to linear stochastic approximation and TD-learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021. 5

[DMPS18] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Springer, 2018. 2, 7

[DNPR20] Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Convergence rates of accelerated Markov gradient descent with applications in reinforcement learning. *arXiv preprint arXiv:2002.02873*, 2020. 5

[DX20] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *arXiv preprint arXiv:2011.11173*, 2020. 3, 5

[FR13] Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pages 1–39. Springer, 2013. 1

[GKS21] Dan Garber, Atara Kaplan, and Shoham Sabach. Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems. *Mathematical Programming*, 186(1):185–208, 2021. 5

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. 1

[GRW20] Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020. 2, 4, 6, 8, 15, 21, 22, 23

[GSK13] Yair Goldberg, Rui Song, and Michael R Kosorok. Adaptive q-learning. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 150–162. Institute of Mathematical Statistics, 2013. 1

[HJN15] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015. 5, 13

[HK12] Elad Hazan and Satyen Kale. Projection-free online learning. In *29th International Conference on Machine Learning, ICML 2012*, pages 521–528, 2012. 5

[HL16] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016. 5

[HMPW16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016. 1

[Jag13] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013. 5, 13, 15

[KM17] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017. 2

[KMMW19] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019. 1

[KMN+20] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020. 5

[KY03] Harold Kushner and George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003. 5

[Lan20] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020. 1

[Lia10] Faming Liang. Trajectory averaging for stochastic approximation MCMC algorithms. *The Annals of Statistics*, 38(5):2823–2856, 2010. 9, 14

[LJJ15] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015. 5

[LW22] Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022. 1, 2, 3, 12

[LZ16] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016. 1, 5

[MB11] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011. 1

[MDPZH20] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020. 1

[MHK20] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 2020. 5, 6

[Mig94] Athanasios Migdalas. A regularization of the Frank—Wolfe method and unification of certain nonlinear programming methods. *Mathematical Programming*, 65(1):331–345, 1994. 5

[MT12] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. 10

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. 3

[QLX18] Chao Qu, Yan Li, and Huan Xu. Non-convex conditional gradient sliding. In *International Conference on Machine Learning*, pages 4208–4217. PMLR, 2018. 5

[QW20] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020. 1

[RBE21] Abhishek Roy, Krishnakumar Balasubramanian, and Murat A Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. *Advances in Neural Information Processing Systems*, 34:8913–8926, 2021. 2

[RSPS16] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic Frank-Wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016. 5

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012. 1

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. 1

[SSXY20] Tao Sun, Yuejiao Sun, Yangyang Xu, and Wotao Yin. Markov chain block coordinate descent. *Computational Optimization and Applications*, 75(1):35–61, 2020. 5

[SSY18] Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018. 5

[SY19] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD-learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019. 5

[SZ13] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013. 1

[TD17] Vladislav B Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304, 2017. 5

[WPT+21] Yafei Wang, Bo Pan, Wei Tu, Peng Liu, Bei Jiang, Chao Gao, Wei Lu, Shangling Jui, and Linglong Kong. Sample average approximation for stochastic optimization with dependent data: Performance guarantees and tractability. *arXiv preprint arXiv:2112.05368*, 2021. 5

[XBG22] Tesi Xiao, Krishnakumar Balasubramanian, and Saeed Ghadimi. A projection-free algorithm for constrained stochastic multi-level composition optimization. *arXiv preprint arXiv:2202.04296*, 2022. 2, 5, 6, 7, 8

[XXLZ21] Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of Adam-type reinforcement learning algorithms under Markovian sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10460–10468, 2021. 5

[XZL19] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[YSC19] Alp Yurtsever, Suvrit Sra, and Volkan Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR, 2019. 5

[ZJM21] Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[ZSM+20] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One-sample Stochastic Frank-Wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020. 2, 5, 6, 8