Convergence of First-Order Methods for Constrained Nonconvex Optimization with Dependent Data

Ahmet Alacaoglu [*1] **Hanbaek Lyu** [*2]

Abstract

We focus on analyzing the classical stochastic projected gradient methods under a general dependent data sampling scheme for constrained smooth nonconvex optimization. We show the worst-case rate of convergence $\tilde{O}(t^{-1/4})$ and complexity $\tilde{O}(\varepsilon^{-4})$ for achieving an ε-near stationary point in terms of the norm of the gradient of Moreau envelope and gradient mapping. While classical convergence guarantee requires i.i.d. data sampling from the target distribution, we only require a mild mixing condition of the conditional distribution, which holds for a wide class of Markov chain sampling algorithms. This improves the existing complexity for the constrained smooth nonconvex optimization with dependent data from $\tilde{O}(\varepsilon^{-8})$ to $\tilde{O}(\varepsilon^{-4})$ with a significantly simpler analysis. We illustrate the generality of our approach by deriving convergence results with dependent data for stochastic proximal gradient methods, adaptive stochastic gradient algorithm AdaGrad and stochastic gradient algorithm with heavy ball momentum. As an application, we obtain first online nonnegative matrix factorization algorithms for dependent data based on stochastic projected gradient methods with adaptive step sizes and optimal rate of convergence.

1. Introduction

Consider the minimization of a function $f : \mathbb{R}^p \to \mathbb{R}$ given as an expectation:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \{f(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}\left[\ell(\boldsymbol{\theta}, \mathbf{x})\right]\}, \tag{1}$$

where π is a distribution on a sample space $\Omega \subseteq \mathbb{R}^d$ with a density function; $\ell\colon \Omega \times \boldsymbol{\Theta} \to \mathbb{R}$ a per-sample loss function and $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ a closed convex set with an efficiently computable projection

$$\text{proj}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2. \tag{2}$$

We assume that f is a smooth and possibly nonconvex function. Constrained nonconvex optimization with *dependent data* arise in many situations such as decentralized constrained optimization over networked systems, where the i.i.d. sampling requires significantly more communication than the dependent sampling (Johansson et al., 2007; 2010; Duchi et al., 2012). Other applications are policy evaluation in reinforcement learning where the Markovian data is naturally present since the underlying model is a Markov Decision Process (Bhandari et al., 2018), and online nonnegative matrix factorization and network denoising (Lyu et al., 2020).

1.1. Related Work and Summary of Contributions

It is well-known that obtaining optimal complexity with *single-sample* projected stochastic gradient descent (SGD) for constrained nonconvex problems is significantly more challenging than unconstrained nonconvex optimization (Ghadimi et al., 2016; Davis and Drusvyatskiy, 2019; Alacaoglu et al., 2021). This challenge has been recently overcome by (Davis and Drusvyatskiy, 2019) within the framework of weakly convex optimization, which resulted in optimal complexity results for projected/proximal SGD (PSGD). Later, this result is extended for algorithms such as SGD with heavy ball momentum (Mai and Johansson, 2020) or adaptive algorithms such as AMSGrad and AdaGrad (Alacaoglu et al., 2021). These guarantees require i.i.d. sampling from the underlying distribution π.

Optimization with non-i.i.d. data is studied in the convex and nonconvex cases with gradient/mirror descent in (Sun et al., 2018; Duchi et al., 2012; Nagaraj et al., 2020) and block coordinate descent in (Sun et al., 2020). SGD is also recently considered in (Wang et al., 2021) for convex problems. Another important work in this direction is (Karimi et al., 2019) that focused on unconstrained nonconvex case

[*]Equal contribution [1]Wisconsin Institute for Discovery, University of Wisconsin–Madison, WI, USA [2]Department of Mathematics, University of Wisconsin–Madison, WI, USA. Correspondence to: Hanbaek Lyu <hlyu@math.wisc.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

with a different assumption on the dependent data compared to previous works and relaxed the assumptions on the variance. For constrained and nonconvex problems, the work (Lyu et al., 2020) showed asymptotic guarantees of stochastic majorization-minimization (SMM)-type algorithms to stationary points of the expected loss function.

More recently, (Lyu, 2022) studied generalized SMM-type algorithms with dependent data for constrained problems and showed the complexity $\tilde{O}(\varepsilon^{-8})$ with standard assumptions (that we will clarify in the sequel) and $\tilde{O}(\varepsilon^{-4})$ when all the iterates of the algorithm lie in the interior of the constraint set, for obtaining ε-stationarity. We also remark that (Lyu, 2022) showed that for the 'empirical loss functions' (recursive average of sample losses), SMM-type algorithms need only $\tilde{O}(\varepsilon^{-4})$ iterations for making the stationarity gap under ε. Our present work does not consider empirical loss functions but focus on expected loss functions. See (Lyu, 2022) for more details.

Since the complexity $\tilde{O}(\varepsilon^{-8})$ is suboptimal for nonconvex expected loss minimization, the motivation of our work is to understand if this complexity is improvable or if it is inherent when we handle dependent data and constraints jointly. Our results conclude that the complexity is indeed improvable and show the near-optimal complexity $\tilde{O}(\varepsilon^{-4})$ for constrained nonconvex problems with dependent data. Unlike our result, previous work (Lyu, 2022) needed an additional assumption that the iterates lie in the interior of the constraint (which is difficult to satisfy in general for constrained problems) for the optimal complexity $\tilde{O}(\varepsilon^{-4})$. Moreover, to our knowledge, no convergence rate of *projected* SGD is known in the *constrained* nonconvex case with non-i.i.d. sampling. We also show the first rates for AdaGrad (Duchi et al., 2010) and SGD with heavy ball momentum (Mai and Johansson, 2020) for this setting. See Table 1 for a summary of the discussion above.

After the completion of our manuscript, we became aware of the recent concurrent work (Dorfman and Levy, 2022) that analyzed AdaGrad with multi level Monte Carlo gradient estimation for dependent data. This work focused on the *unconstrained* nonconvex setting whereas our main focus is the more general class of *constrained* nonconvex problems. Hence we believe the two results complement each other.

We also note that slightly stronger versions of Assumption 2.1 are required even for unconstrained nonconvex optimization with dependent data, see (Sun et al., 2018; Dorfman and Levy, 2022). It is well-known that this assumption is difficult to satisfy in the unconstrained setting, but it is more realistic with the presence of constraints. Because of this reason, our results incorporating the constraints and projections in the algorithm provides a more realistic problem setup. While our results would recover those in (Sun et al., 2018) when specialized to the unconstrained case, due to Assumption 2.1, this unconstrained setting would be less realistic as argued above. Because of this, and for other motivating applications, the main focus of this paper is obtaining optimal complexity results for constrained nonconvex problems.

1.2. Contribution

We consider convergence of stochastic first-order methods, including proximal and projected stochastic gradient descent (SGD), projected SGD with momentum, and stochastic adaptive gradient descent (AdaGrad-norm). These are all classical nonconvex optimization algorithms that have been used extensively in various optimization and machine learning tasks. Our main focus is to establish optimal convergence rate for such stochastic first-order methods under very general data sampling scheme, including functions of Markov chains, state-dependent Markov chains, and more general stochastic processes with fast enough mixing of multi-step conditional distribution.

To summarize our results, consider the following simple first-order method:

Step 1. Sample $\mathbf{x}_{t+1}$ from a distribution conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_t$; ($\triangleright$ possibly non-i.i.d. samples)

Step 2. Compute a stochastic gradient $G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})$ (see Assumption 2.1 for Def.) and $\boldsymbol{\theta}_{t+1} \leftarrow \text{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\right)$, where the step size α_t is chosen so that either (1) non-summable and square-summable; or (2) according to AdaGrad-norm: $\alpha_t^{-2} = \alpha_{t-1}^{-2} + \|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 \alpha^{-2}$ for $\alpha > 0$.

An important point here is that *we do not require the new training point $\mathbf{x}_{t+1}$ to be distributed according to the stationary distribution π, nor to be independent on all the previous samples $\mathbf{x}_1, \ldots, \mathbf{x}_t$*. For instance, we allow one to sample $\mathbf{x}_{t+1}$ according to an underlying Markov chain, so that each step of sampling is computationally very efficient but the distribution $\mathbf{x}_{t+1}$ conditional on $\mathbf{x}_t$ could be far from π. This may induce bias in estimating the stochastic gradient $G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})$.

Suppose f is ρ-smooth; $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is convex, closed; and the training samples $\mathbf{x}_t$ are a function of some underlying Markov chain mixing sufficiently fast (see Section 2). Under some mild assumptions used in the literature (Sun et al., 2018; Lyu, 2022; Bhandari et al., 2018), we establish the following convergence results for a wide range of stochastic first-order methods under non-i.i.d. data setting:

- We show that any convergent subsequence of $(\boldsymbol{\theta}_t)_{t\geq 0}$ converges to a stationary point of (1) almost surely. The rate of convergence for finding stationary points is

$\tilde{O}(T^{-1/4})$ (measured using gradient mapping norm). (Thm. 3.1, 3.9, 3.3)

- The same result as above holds when $(\boldsymbol{\theta}_t)_{t\geq 0}$ are generated by using stochastic heavy ball (see Alg. 3) and projected SGD with state-dependent Markov chain (see Thm. 3.8.

This is the same rate of convergence as in the i.i.d. case, up to log-factors, which was obtained in (Davis and Drusvyatskiy, 2019) in terms of gradient mapping as shown in Thm. 3.9. Hence our analysis shows that the convergence of the algorithm and the order of the rate of convergence are not affected by such statistical bias in sampling training data, which was described earlier in this subsection. Furthermore, our result improves the rate of convergence of stochastic algorithms for constrained nonconvex expected loss minimization with dependent data (Lyu, 2022), see Thm. 3.9 for the details. Moreover, we extend our analysis to obtain similar results for such projected SGD algorithms as adaptive gradient algorithm AdaGrad (see Algorithm 2 and Theorem 3.3) and SGD with heavy ball momentum (see Algorithm 3 and Theorem 3.4).

1.3. Notations

We fix $p \in \mathbb{N}$ to be the dimension of the ambient Euclidean space $\mathbb{R}^p$ equipped with the inner product $\langle \cdot, \cdot \rangle$ that also induces the Euclidean norm $\|\cdot\|$. For each $\varepsilon > 0$, let $B_\varepsilon := \{x \in \mathbb{R}^p \,|\, \|x\| \leq \varepsilon\}$ denote the ε-ball centered at the origin. We also use the distance function defined as $\operatorname{dist}(\boldsymbol{\theta}, \boldsymbol{\Theta}) = \min_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ and the σ-algebra $\mathcal{F}_{t-k} = \sigma(\mathbf{x}_1, \ldots, \mathbf{x}_{t-k})$. We denote $f \colon \boldsymbol{\Theta} \to \mathbb{R}$ to be a generic objective function for which we introduce the precise assumptions in Section 2, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is closed and convex. Let $\iota_{\boldsymbol{\Theta}}$ denote the indicator function of the set $\boldsymbol{\Theta}$, where $\iota_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\iota_{\boldsymbol{\Theta}}(\theta) = +\infty$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$. Note that

$$\operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \{\varphi(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \iota_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\}. \quad (3)$$

1.4. Preliminaries on Stationarity Measures

Since we do not expect the first-order optimality conditions to be satisfied exactly in a finite number of iterations in practice, we wish to estimate the worst-case number of iterations required to achieve an ε-*approximate* solution and the corresponding scaling with ε. To this end, we can relax the first-order optimality conditions as follows: For each $\varepsilon > 0$, we say $\boldsymbol{\theta}^*$ is an ε-*stationary point* (or ε-approximate stationary point) for f over $\boldsymbol{\Theta}$ if and only if $\operatorname{dist}(\mathbf{0}, \partial\varphi(\boldsymbol{\theta}^*)) \leq \varepsilon$. We say a point $\boldsymbol{\theta}^*$ is *approximately near stationary* for f over $\boldsymbol{\Theta}$ if there exists some point $\hat{\boldsymbol{\theta}}$ near $\boldsymbol{\theta}$ that is approximately stationary for f over $\boldsymbol{\Theta}$. We will make this notion precise through the following discussion.

One of the central notions in the recent influential work by Davis and Drusvyatskiy (2019) in analyzing convergence rates of first-order methods for constrained nonconvex problems is the *Moreau envelope*, which is a smooth approximation of an objective function that is closely related to proximal mapping. For a constant $\lambda > 0$, we define the *Moreau envelope* φ_λ of φ defined in (3) as

$$\varphi_\lambda(\boldsymbol{\theta}) := \min_{\boldsymbol{\theta}' \in \mathbb{R}^p} \left(\varphi(\boldsymbol{\theta}') + \frac{1}{2\lambda} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \right). \quad (4)$$

If f is ρ-weakly convex and if $\lambda < \rho^{-1}$, then the minimum in the right hand side is uniquely achieved at a point $\hat{\boldsymbol{\theta}}$, which we call the *proximal point* of $\boldsymbol{\theta}$. Accordingly, we define the proximal map

$$\begin{aligned} \hat{\boldsymbol{\theta}} &:= \operatorname{prox}_{\lambda\varphi}(\boldsymbol{\theta}) \\ &:= \operatorname*{arg\,min}_{\boldsymbol{\theta}' \in \mathbb{R}^p} \left(\varphi(\boldsymbol{\theta}') + \frac{1}{2\lambda} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2 \right) \end{aligned} \quad (5)$$

Also in this case, the Moreau envelope φ_λ is C^1 with gradient given by (see (Davis and Drusvyatskiy, 2019))

$$\nabla\varphi_\lambda(\boldsymbol{\theta}) = \lambda^{-1}(\boldsymbol{\theta} - \operatorname{prox}_{\lambda\varphi}(\boldsymbol{\theta})). \quad (6)$$

When $\boldsymbol{\theta}$ is a stationary point of φ, then its proximal point $\hat{\boldsymbol{\theta}}$ should agree with $\boldsymbol{\theta}$. Hence the gradient norm of the Moreau envelope φ_λ may provide an alternative measure of stationarity. Indeed, as shown in (Davis and Drusvyatskiy, 2019), it provides a measure of *near stationarity* in the sense that if $\|\nabla\varphi_\lambda(\boldsymbol{\theta})\|$ is small, then since the proximal point $\hat{\boldsymbol{\theta}}$ in (5) is within $\lambda\|\nabla\varphi_\lambda(\boldsymbol{\theta})\|$ from $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ approximately stationary in terms of $\operatorname{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}}))$:

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| \leq \lambda\|\nabla\varphi_\lambda(\boldsymbol{\theta})\|, \quad \varphi(\hat{\boldsymbol{\theta}}) \leq \varphi(\boldsymbol{\theta}), \quad (7)$$

$$\operatorname{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}})) \leq \|\nabla\varphi_\lambda(\boldsymbol{\theta})\|. \quad (8)$$

Note that the first and the last inequality above follows from the first-order optimality condition for $\hat{\boldsymbol{\theta}}$ together with (6) (see also Propositions B.2 and B.1 in Appendix B).

Hence, in the literature of weakly convex optimization, it is common to state the results in terms of the norm of the gradient of Moreau envelope (Davis and Drusvyatskiy, 2019; Drusvyatskiy and Paquette, 2019) which we will also adopt. When g is additionally smooth, a commonly adopted measure to state convergence results is *gradient mapping* which is defined as (Nesterov, 2013)

$$\begin{aligned} \|\mathcal{G}_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\| &= \hat{\rho} \left\| \boldsymbol{\theta}_t - \operatorname{proj}_{\boldsymbol{\Theta}} \left(\boldsymbol{\theta}_t - \frac{1}{\hat{\rho}} \nabla f(\boldsymbol{\theta}_t) \right) \right\| \\ &=: \hat{\rho} \|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t\|, \end{aligned} \quad (9)$$

for any $\lambda > 0$, where we also defined $\tilde{\boldsymbol{\theta}}_t$. The results (Davis and Drusvyatskiy, 2019) (Drusvyatskiy and Lewis, 2018)

	$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta})$ f: L-smooth	$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} f(\boldsymbol{\theta})$ f: L-smooth	Markovian data	Constrained
SMM (Lyu, 2022)	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-8})^{\dagger}$	✓	✓
SGD (Sun et al., 2018; Karimi et al., 2019)	$\tilde{O}(\varepsilon^{-4})$	–	✓	✗
Proj. SGD (Davis and Drusvyatskiy, 2019)	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-4})$	✗	✓
Proj. SGD-Sec. 3.1	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-4})$	✓	✓
AdaGrad-Sec. 3.2	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-4})$	✓	✓

Table 1. Complexity comparison for stochastic nonconvex optimization with non-i.i.d. data. Complexities in each column are the number of stochastic gradients to obtain: $\mathbb{E}\|\nabla f(\boldsymbol{\theta})\| \leq \varepsilon$ and $\mathbb{E}\left[\text{dist}(0, \partial\varphi(\boldsymbol{\theta}))\right] \leq \varepsilon$, respectively (where φ is defined in (3)). †This work showed the improved complexity $\tilde{O}(\varepsilon^{-4})$ under the additional assumption that the iterates of the algorithm are in the interior of $\boldsymbol{\Theta}$, which does not necessarily hold in the constrained case. We do not make such an assumption in this paper.

showed how to translate the guarantees on the gradient of the Moreau envelope to gradient mapping by proving that

$$\|\mathcal{G}_{1/2\hat{\rho}}(\boldsymbol{\theta})\| \leq \frac{3}{2}\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta})\|.$$

It is easy to show that a small gradient mapping implies that $\boldsymbol{\theta}_t$ is close to $\text{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_t - (1/\hat{\rho})\nabla f(\boldsymbol{\theta}_t)\right)$ which itself is approximately stationary in view of Sec. 1.4 which can be shown by using the definition of $\tilde{\boldsymbol{\theta}}_t$ and smoothness of f. Even though such an approximately stationary point can be computed in the deterministic case, computation of $\nabla f(\boldsymbol{\theta}_t)$ is not tractable in the stochastic case. However, as we show in Sec. 3.6, we can still output a point which is approximately stationary, in a tractable manner, with the claimed complexity results in our dependent data setting.

2. Stochastic Gradient Estimation

Denote as $\Delta_{[t-k,t]}$ the worst-case total variation distance between conditional distribution of $\mathbf{x}_t$ given $\mathbf{x}_1, \ldots, \mathbf{x}_{t-k} \in \Omega$ and the stationary distribution π. Namely,

$$\Delta_{[t-k,t]} := \sup_{\mathbf{x}_1,\ldots,\mathbf{x}_{t-k}} \|\pi_t(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{t-k}) - \pi\|_{TV}, \quad (10)$$

where $\pi_{t|t-k} = \pi_t(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{t-k})$ denotes the probability distribution of $\mathbf{x}_t$ conditional on the past points $\mathbf{x}_1, \ldots, \mathbf{x}_{t-k}$.

Most of our theoretical results (except Theorem 3.8 for state-dependent Markov chains, see Section 3.5) operate under the following three assumptions.

Assumption 2.1. The function f is C^1 smooth and has ρ-Lipschitz gradient and the set $\boldsymbol{\Theta}$ is closed and convex. There exists an open set U containing $\boldsymbol{\Theta}$ and a mapping $G : U \times \Omega \to \mathbb{R}^p$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}, \mathbf{x})\right] = \nabla f(\boldsymbol{\theta})$. Also $\boldsymbol{\theta} \mapsto G(\boldsymbol{\theta}, x)$ is L_1-Lipschitz for all x for some $L_1 > 0$.

Assumption 2.2. We can sample a sequence of points $(\mathbf{x}_t)_{t\geq 1}$ in Ω in a way that: (1) For each $x \geq 0$, $\Delta_{[t,t+N]}$ is non-increasing in $N \geq 0$; and (2) $\lim_{N\to\infty} \Delta_{[t,t+N]} = 0$ for all $t \geq 0$; and (3) there exists a sequence $k_t \in [0, t]$, $t \geq 1$ such that $\Delta_{[t-k_t,t]} \to 0$ and $\sum_{t=1}^{\infty} \alpha_t \Delta_{[t-k_t,t]} < \infty$, where $\alpha_t > 0$ denotes the stepsize in the first-order method.

Assumption 2.3. Assume either of the two: *(i)* There is $L \in (0, \infty)$ such that for each $t \geq 1$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}\left[\|G(\boldsymbol{\theta}, \mathbf{x}_{t+1})\| \,|\, \mathcal{F}_t\right] \leq L$ and the process $(\mathbf{x}_t)_{t\geq 0}$ is a function of some time-homogeneous Markov chain; or *(ii)* There is $L \in (0, \infty)$ such that $\|G(\boldsymbol{\theta}, \mathbf{x})\| \leq L$ for all $\boldsymbol{\theta}, x$.

Assumption 2.1 is about smoothness of the objective and stochastic gradient operator G. The former is standard in the literature of stochastic constrained first-order methods and the latter is also common when we additionally work with dependent data(see, e.g., (Davis and Drusvyatskiy, 2019; Sun et al., 2020; Lyu, 2022)).

Assumption 2.2 states that: (1) The N-step conditional distribution $\pi_{t+N|t}$ can only be closer to the stationary distribution π when N increases; (2) the N-step conditional data distribution $\pi_{t+N|t}$ converges to the stationary distribution π asymptotically; and (3) such convergence (mixing) occurs at a sufficiently fast rate. The sequence k_t plays a critical role in controlling dependence in data samples. The key idea is that, when analyzing quantities at time $t+1$, one conditions on a 'distant past' $t - k_t$ (instead of the present t) and approximates the multi-step conditional data distribution $\pi_{t+1|t-k_t}$ by the stationary distribution π. The error of such approximation in the total variation distance is bounded by $\Delta_{[t-k_t,t]}$. Assumption 2.2 requires that this quantity should be summable after being multiplied by the stepsize α_t.

There are two notable special cases that satisfy Assumption 2.2. First, Assumption 2.2 is trivially satisfied (with $k_t \equiv 0$) in the i.i.d. case since then $\pi_{s|t} \equiv \pi$ whenever $s > t$.

Second, suppose $\mathbf{x}_t$ is given by a function g of some underlying time-homogeneous Markov chain X_t with a sta-

tionary distribution π. In this case Assumption 2.2(1) holds by Scheffé's lemma (see, e.g., Lemma 2.1 in (Tsybakov, 2004)). (Here time-inhomogeneity is not necessary.) If X_t is irreducible and aperiodic on a finite state space, then Assumption 2.2(2) holds with $\Delta_{[t-k,t]} = O(\exp(-ck))$ for some constant $c > 0$ independent of t (Levin and Peres, 2017). So Assumption 2.2(3) is verified for any $k_t \geq C \log t$ for $C > 0$ large enough so that $\sum_{t\geq 1} \exp(-ck_t) < \infty$ and for any $\alpha_t = O(1)$. In the case when the underlying Markov chain X_t has countably infinite or uncountable state space, then a more general condition for geometric ergodicity is enough to imply Assumption 2.2 (see, e.g., (Levin and Peres, 2017; Meyn and Tweedie, 2012)). See (Lyu et al., 2020) and (Sun et al., 2018) for concrete applications and sampling methods that satisfy this assumption. This assumption is common in the literature (Bhandari et al., 2018; Lyu, 2022; Lyu et al., 2020; Sun et al., 2018; Nagaraj et al., 2020) and i.i.d. sampling is another special case.

We emphasize that Assumption 2.2 does not necessarily reduce to time-homogeneous and state-independent Markov chains. Our main focus is using Assumption 2.2 which is the main assumption on the data in most of the works we compare with. However, we also discuss another popular setting of modeling dependent data samples by *state-dependent Markov chain*. See 3.6-3.7 and Thm. 3.8.

Next, we discuss Assumption 2.3 on boundedness of stochastic gradients. In the i.i.d. case, it is standard to assume uniform boundedness of $\mathbb{E}_{\mathbf{x}\sim\pi}[\|G(\boldsymbol{\theta}, \mathbf{x})\|]$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (Davis and Drusvyatskiy, 2019; Davis et al., 2020). In the non-i.i.d. case, it has been customary to make stronger assumption of uniform boundedness of $G(\boldsymbol{\theta}, \mathbf{x})$ even in the unconstrained nonconvex case (Sun et al., 2018; Dorfman and Levy, 2022), which does not properly generalize the standard assumption in the i.i.d. case. This is mostly for controlling the error of multi-step conditional expectation of the stochastic gradient by its stationary expectation, which is the crucial issue in the non-i.i.d. case that is non-existent in the i.i.d. case.

In this work, we are able to analyze the non-i.i.d. setting under a much weaker condition in Assumption 2.3(i) that only assumes one-step conditional expectation of the norm of the stochastic gradient is bounded. Although for a technical reason we will also need to assume that the data samples $(\mathbf{x}_t)_{t\geq 0}$ are given as a function of some time-homogeneous Markov chain, Assumption 2.3(i) properly generalizes the standard assumptions in the i.i.d. case. In addition, We also analyze non-i.i.d. setting under uniformly bounded stochastic gradients but with more general data sampling setting (Assumption 2.3(ii)), including time-inhomogeneious and non-Markovian setting.

Now we state a key lemma that handles the bias due to dependent data and is algorithm independent. In the sequel, we will invoke this lemma for different algorithms such as SGD, AdaGrad or SGD with heavy ball momentum.

Lemma 2.4 (Key lemma)**.** *Let Assumptions 2.1, 2.2, 2.3 hold and $\boldsymbol{\theta}_t$ be generated according to Algorithm 1, 2 or 3. Fix $\hat{\rho} > \rho$ and denote $\hat{\boldsymbol{\theta}} = \mathrm{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta})$ and fix $1 \leq k \leq t$. Then*

$$\Big|\mathbb{E}\big[\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\rangle \,|\, \mathcal{F}_{t-k}\big] - \langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t, \mathbf{x})\right]\rangle\Big| \leq \frac{4L^2}{\hat{\rho}-\rho}\,\Delta_{[t-k,t]} + \frac{2L(L_1+\hat{\rho})}{\hat{\rho}-\rho}\,\mathbb{E}\left[\sum_{s=t-k}^{t-1} \alpha_s \|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| \,\middle|\, \mathcal{F}_{t-k}\right]. \tag{11}$$

This lemma borrows some ideas from (Lyu, 2022). The important difference is that, the result of the lemma makes it explicit the dependence on the step size and gradient norms to be applicable with AdaGrad. This is needed because the step size of AdaGrad does not have a specific decay schedule. The proof is given in Section C.

3. Convergence Rate Analysis

3.1. Projected SGD with Dependent Data

Now we state our first main result in this work, which extends the convergence result of projected SGD with i.i.d. samples in (Davis and Drusvyatskiy, 2019) to the general dependent sample setting. This result improves the existing complexity of stochastic algorithms from (Lyu, 2022) for solving constrained nonconvex stochastic optimization under dependent data, see Section 3.6 for details. We use the notion of *global convergence* with respect to arbitrary initialization below. The proof of this result is in Appendix D.

Algorithm 1 Projected Stochastic Gradient Algorithm (PSGD)

Input: Initialize $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$; $T > 0$; Stepsizes $(\alpha_t)_{t\geq 1}$
Sample τ from $\{1, \ldots, T\}$ independently of everything else where $\mathbb{P}(\tau = k) = \frac{\alpha_k}{\sum_{t=1}^{T} \alpha_t}$.
For $t = 1, 2, \ldots, T$ **do:**
  Sample $\mathbf{x}_{t+1}$ from $\pi_{t+1} = \pi_{t+1}(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_t)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \mathrm{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\right)$
End for
Return: $\boldsymbol{\theta}_T$ (Optionally, $\boldsymbol{\theta}_T^{\text{out}}$ as either $\boldsymbol{\theta}_\tau$ or $\arg\min_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T\}} \|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta})\|^2$.)

Theorem 3.1 (Projected stochastic gradient method)**.** *Let Assumptions 2.1-2.3 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1. Fix $\hat{\rho} > \rho$. Then the following hold:*

(i) *(Rate of convergence) For each $T \geq 1$,*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] = O\Big(\frac{1}{\sum_{k=1}^T \alpha_k}\Big(\sum_{t=1}^T \alpha_t^2 + \sum_{t=1}^T k_t\alpha_t\alpha_{t-k_t} + \sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t]}]\Big)\Big). \quad (12)$$

In particular, with $\alpha_t = \frac{c}{\sqrt{t}}$ for some $c > 0$ and under exponential mixing, we have that $\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|\right] \leq \varepsilon$ with $\tilde{O}\left(\varepsilon^{-4}\right)$ samples.

(ii) *(Global convergence) Further assume that $\sum_{t=0}^{\infty} k_t\alpha_t\alpha_{t-k_t} < \infty$. Then $\|\nabla\varphi_{1/\hat{\rho}}(\hat{\boldsymbol{\theta}}_t)\| \to 0$ as $t \to \infty$ almost surely. Furthermore, $\boldsymbol{\theta}_t$ converges to the set of all stationary points of f over $\boldsymbol{\Theta}$.*

If $(\mathbf{x}_t)_{t\geq 1}$ is exponentially mixing, then Theorem 3.1**(ii)** holds with $\alpha_t = t^{-1/2}(\log t)^{-1-\varepsilon}$ for any fixed $\varepsilon > 0$ and $k_t = O(\log t)$.

3.2. AdaGrad with Dependent Data

We next establish the convergence of AdaGrad with dependent data and constrained nonconvex optimization. We will use AdaGrad with scalar step sizes (see Alg. 2), which is also referred to as AdaGrad-norm (Ward et al., 2019; Levy, 2017; Streeter and McMahan, 2010).

Algorithm 2 AdaGrad-norm (Streeter and McMahan, 2010)

Input: Initialize $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$; $T > 0$; $(\alpha_t)_{t\geq 1}$; $v_0 > 0$; $\alpha > 0$
Optionally, sample τ from $\{1, \ldots, T\}$ independently of everything else where $\mathbb{P}(\tau = k) = \frac{1}{T}$.
For $t = 1, 2, \ldots, T$ **do:**
  Sample $\mathbf{x}_{t+1}$ from $\pi_{t+1} = \pi_{t+1}(\cdot \mid \mathbf{x}_1, \ldots, \mathbf{x}_t)$
  $v_t = v_{t-1} + \|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2$
  $\alpha_t = \frac{\alpha}{\sqrt{v_t}}$
  $\boldsymbol{\theta}_{t+1} \leftarrow \text{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\right)$
End for
Return: $\boldsymbol{\theta}_T$ (Optionally, $\boldsymbol{\theta}_T^{\text{out}}$ as either $\boldsymbol{\theta}_\tau$ or $\arg\min_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T\}} \|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta})\|^2$.)

For this section, we introduce an additional assumption on the boundedness of the objective values.

Assumption 3.2. (A4) There exists $C_\varphi \in (0, \infty)$ such that $|f(\boldsymbol{\theta})| \leq C_\varphi$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Compared to projected SGD, the step size of AdaGrad does not have a specific decay schedule, which makes it challenging to use the existing bias analyses for dependent data (for example the idea from (Lyu, 2022)) since they critically rely on knowing the precise decay rate of the step sizes.

To be able to apply such an analysis for adaptive algorithms, we use a generalized result in Lem. 2.4 and use the particular form of AdaGrad step size in Thm. 3.3 to achieve the optimal $\tilde{O}(\varepsilon^{-4})$ complexity. Full proof of the result is given in Appendix E.

Theorem 3.3 (AdaGrad-norm)**.** *Let Assumption 2.1-2.3 and Assumption 3.2 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 2. Fix $\hat{\rho} > \rho$ and a nondecreasing, diverging sequence $(k_t)_{t\geq 1}$. Then, for each $T \geq 1$,*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] = O\Big(\frac{k_T \log(TL^2)}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\Delta_{[t-k_t,t]}]\Big). \quad (13)$$

We note that unlike Thm. 3.1, for AdaGrad we only prove nonasymptotic complexity results and not asymptotic convergence statements for the output sequence of the algorithms. Even though asymptotic convergence of AdaGrad with i.i.d. data is proven in (Li and Orabona, 2019), the technique in that paper relies on using the inequality (128) multiplied with α_t. However, the specific form of (128) is important in our development to use Lem. 2.4 to handle the dependent data, since α_t brings additional stochastic dependencies. Even though we believe an appropriate modification of Lem. 2.4 can be possible, we do not pursue such generalization in the present work.

Since the step size in this case is nonincreasing, Assumption 2.2 reduces to $\sum_{t=1}^{\infty} \Delta_{[t-k,t]} < \infty$. This, for example, is satisfied for the exponential mixing case that is mentioned in Theorem 3.1 and considered in the previous work (Lyu, 2022; Sun et al., 2018; Bhandari et al., 2018).

3.3. Stochastic Heavy Ball with Dependent Data

Because of space limitations, we defer the formal description of SGD with heavy ball momentum to the appendix (Algorithm 3) and include a summary of the complexity result here. The extended theorem for this case, including the asymptotic convergence of the sequence and the proofs are given in Appendix F.

Theorem 3.4. *Let Assumption 2.1-2.3 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 3. Fix $\hat{\rho} \geq 2\rho$. Then, for any momentum parameter $\beta \in (0, 1]$ and $T \geq 1$:*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] = O\Big(\frac{1}{\beta^2\sum_{k=1}^T \alpha_k}\Big(\sum_{t=1}^T \alpha_t^2 + \sum_{t=1}^T k_t\alpha_t\alpha_{t-k_t} + \sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t]}]\Big)\Big). \quad (14)$$

Our analysis for the heavy ball method appears to be more flexible compared to (Mai and Johansson, 2020) even when restricted to the i.i.d. case. In this case, we allow variable step sizes $\alpha_t = \frac{\gamma}{\sqrt{t}}$ whereas (Mai and Johansson, 2020) requires constant step size $\alpha_t = \alpha = \frac{\gamma}{\sqrt{T}}$. We can also

use any momentum parameter $\beta \in (0, 1]$ whereas (Mai and Johansson, 2020) restricts to $\beta = \alpha$. This point is important since in practice β is used as a tuning parameter.

3.4. Proximal SGD with Dependent Data

In this section, we describe how our developments for stochastic gradient method extends to the proximal case, using the ideas from (Davis and Drusvyatskiy, 2019). In particular, the problem we solve in this section is

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \Big(\varphi(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\Big), \tag{15}$$

where f is as in (1) and $r\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a convex, proper, closed function. In this case, in step 1 of Algorithm 1, we use $\text{prox}_{\alpha_t r}$ instead of $\text{proj}_{\boldsymbol{\Theta}}$ to define $\boldsymbol{\theta}_{t+1}$. We include the following result combining the ideas from Lem. 2.4, Thm. 3.1 and (Davis and Drusvyatskiy, 2019) for proving convergence of proximal stochastic gradient algorithm with dependent data. Full details are given in Appendix G.

Theorem 3.5. *Let Assumption 2.1-2.3 hold, r be convex, proper, closed and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1 where we use* $\text{prox}_{\alpha_t r}$ *instead of* $\text{proj}_{\boldsymbol{\Theta}}$ *in step 1. Fix $\hat{\rho} > \rho$. For each $T \geq 1$,*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] = O\Bigg(\frac{1}{\sum_{k=1}^T \alpha_k}\Bigg(\sum_{t=1}^T \alpha_t^2 + \sum_{t=1}^T k_t\alpha_t\alpha_{t-k_t} + \sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t+1]}]\Bigg)\Bigg). \tag{16}$$

3.5. Projected SGD with state-dependent Markovian data

Next, we state an analogous result to Theorem 3.1 when the data samples $(\mathbf{x}_t)_{t\geq 0}$ form a *state-dependent Markov chain*. It extends the corresponding results in (Karimi et al., 2019; Tadić and Doucet, 2017) to the constrained case. One difference is that in the constrained case, we need a slightly stronger assumption on the norms of the gradients, see 2.3. The assumptions below were adapted from (Karimi et al., 2019) and (Tadić and Doucet, 2017).

Assumption 3.6. The sequence of data samples $(\mathbf{x}_t)_{t\geq 0}$ form a *state-dependent Markov chain controlled by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, denoted as $(X_t)_{t\geq 0}$. That is, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists a Markov kernel $P_{\boldsymbol{\theta}} : \Omega \to \Omega$ such that for any bounded measurable function H,

$$\mathbb{E}[H(X_{t+1})|\mathcal{F}_t] = P_{\boldsymbol{\theta}_t} H(X_t), \tag{17}$$

where $\mathcal{F}_t := \sigma(X_0, \boldsymbol{\theta}_0, X_1, \boldsymbol{\theta}_1, \ldots, X_t, \boldsymbol{\theta}_t)$.

Assumption 3.7. There is a Lipschitz continuous solution to the Poisson equation for $(X_t)_{t\geq 0}$. That is, there exists a measurable function $\hat{G}$ such that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}, x \in \Omega$,

$$\hat{G}(\boldsymbol{\theta}, x) - P_{\boldsymbol{\theta}}\hat{G}(\boldsymbol{\theta}, x) = G(\boldsymbol{\theta}, x) - \nabla f(\boldsymbol{\theta}), \tag{18}$$

where f denotes the objective function in (1) and $G(\boldsymbol{\theta}, x)$ is as in Assumption 2.1. Furthermore, There exists C_1, C_2, C_3 such that

$$\|\hat{G}(\boldsymbol{\theta}, x)\| \leq C_1, \quad \|P_{\boldsymbol{\theta}}\hat{G}(\boldsymbol{\theta}, x)\| \leq C_2, \tag{19}$$

$$\sup_x \|P_{\boldsymbol{\theta}}\hat{G}(\boldsymbol{\theta}, x) - P_{\boldsymbol{\theta}'}\hat{G}(\boldsymbol{\theta}', x)\| \leq C_3\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|. \tag{20}$$

Theorem 3.8 (Projected SGD with state-dependent MC data)**.** *Let Assumptions 2.1, 2.3, 3.6, 3.7 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1. A complexity result as in Theorem 3.1* ***(i)*** *still hold with possibly different constants. See Theorem K.1 for details.*

While Lemma 2.4 was the key to establish convergence of PSGD (Theorem 3.1) under the mixing condition in Assumption 2.2, a similar role is played by the solution of Poisson equation stated in Assumption 3.7 for the state-dependent case. The proof of Theorem 3.8 follows the same lines as Theorem 3.1 using a similar analysis as in (Karimi et al., 2019) for the bias and properties of the sequences $\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_t$. See Appendix K.

3.6. Complexity for Constrained Smooth Optimization with Dependent Data

We next compare our complexity with the one derived in (Lyu, 2022) for constrained smooth nonconvex optimization with dependent data which, to our knowledge, is the only complexity result for this setting. First, we introduce the next assumption to replace Assumption 2.1. We next show how to translate our result to a direct stationarity measure in view of Sec. A.1 to compare with the $\tilde{O}(\varepsilon^{-8})$ complexity result in (Lyu, 2022) for an equivalent stationarity measure (see Sec. A.1 for details). The proof of the result is given in Appendix H.

Theorem 3.9 (Sample complexity)**.** *Let Assumption 2.1-2.3 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by any of the Algorithms 1, 2, and 3. Fix $\hat{\rho} > \rho$, assume $\Delta_{[t-k_t,t+1]} = O(\lambda^{k_t})$ for $\lambda < 1$ and that $\boldsymbol{\Theta}$ is compact. Pick $\hat{t}$ randomly from $\{1, \ldots, T\}$ as in the respective theorems for the algoritms, let $\hat{N} = O(\varepsilon^{-2})$, and define*

$$\breve{\boldsymbol{\theta}}_{\hat{t}+1} = \text{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_{\hat{t}} - \frac{1}{\hat{N}}\sum_{i=1}^{\hat{N}} \nabla\ell(\boldsymbol{\theta}_{\hat{t}}, \mathbf{x}^{(i)})\right). \tag{21}$$

Then $\mathbb{E}\left[\text{dist}(\mathbf{0}, \partial(f + \iota_{\boldsymbol{\Theta}})(\breve{\boldsymbol{\theta}}_{\hat{t}+1}))\right] \leq \varepsilon$ with $\tilde{O}(\varepsilon^{-4})$ samples.

The assumption on $\Delta_{[t-k_t,t+1]}$ and hence the dependent sampling is consistent with the related works (Lyu, 2022; Sun et al., 2018).

Even though gradient of the Moreau envelope is a *near approximate stationarity* measure, in the specific case of this section, we show that we can output a point that is *approximately stationary* with respect to the direct stationarity measure in Prop. B.1(i) (also mentioned in Section 1.4). This permits a direct comparison with the previous result on constrained nonconvex optimization with dependent data (Lyu, 2022) and shows our improvement. Lemma H.1 in the appendix gives the necessary post-processing step for this, which is used in Theorem 3.9.

4. Application: Online Dictionary Learning

Consider the *online dictionary learning* (ODL) problem, which is stated as the stochastic program

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\subseteq\mathbb{R}^{p\times r}} \Big(f(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{X}\sim\pi}\left[\ell(\mathbf{X},\boldsymbol{\theta})\right] \Big) \text{ where } \ell(\mathbf{X},\boldsymbol{\theta}) := \inf_{H\in\boldsymbol{\Theta}'\subseteq\mathbb{R}^{p\times n}} d\left(\mathbf{X},\boldsymbol{\theta}H\right) + R(H) \tag{22}$$

where $d(\cdot,\cdot) : \mathbb{R}^{p\times n}\times\mathbb{R}^{p\times n}\to[0,\infty)$ is a multi-convex function that measures dissimilarity between two $p\times n$ matrices (e.g., the squared frobenius norm, KL-divergence), $R:\mathbb{R}^{p\times n}\to[0,\infty)$ denotes a convex regularizer for the code matrix H, and r is an integer parameter for the rank of the intended compressed representation of data matrix $\mathbf{X}$. In words, we seek to learn a single dictionary matrix $\boldsymbol{\theta}\in\mathbb{R}^{p\times r}$ within the constraint set $\boldsymbol{\Theta}$ (e.g., nonnegative matrices with bounded norm), which provides the best linear reconstruction (w.r.t. the d-metric) of an unknown random matrix X drawn from some distribution π. Here, we may put L_1-regularization on H in order to promote dictionary $\boldsymbol{\theta}$ that enable sparse representation of observed data.

The most extensively investigated instance of the above ODL problem is when d equals the squared Frobenius distance. In this case, Mairal et al. (Mairal et al., 2010) provided an online algorithm based on the framework of stochastic majorization-minimization (Mairal, 2013). A well-known result in (Mairal et al., 2010) states that the above algorithm converges almost surely to the set of stationary points of the expected loss function f in (22), provided the data matrices $(\mathbf{X}_t)_{t\geq1}$ are i.i.d. according to the stationary distribution π. Later Lyu, Needell, and Balzano (Lyu et al., 2020) generalized the analysis to the case where $(\mathbf{X}_t)$ are given by a function of some underlying Markov chain. Recently, Lyu (2022) provided the first convergence rate bound of the ODL algorithm in Mairal et al. (2010) of order $O((\log t)^{1+\varepsilon}/t^{1/4})$ for the empirical loss function and $O((\log t)^{1+\varepsilon}/t^{1/8})$ for the expected loss function for arbitrary $\varepsilon>0$.

Suppose we are given a sequence of data matrices $(\mathbf{X}_t)_{t\geq1}$ that follows π in some asymptotic sense. Under some mild assumptions, one can compute the subgradient of the loss function $\boldsymbol{\theta}\mapsto\ell(\mathbf{X}_t,\boldsymbol{\theta})$ in two steps and can perform a standard stochastic projected gradient descent:

$$\begin{cases} H_t \leftarrow \arg\min_{H\in\boldsymbol{\Theta}'} d\left(\mathbf{X},\boldsymbol{\theta}_{t-1}H\right) + \lambda\|H\|_1, \\ G(\boldsymbol{\theta}_{t-1},\mathbf{X}_t) = \nabla_{\boldsymbol{\theta}} d(\mathbf{X}_t,\boldsymbol{\theta}_{t-1}H_t), \\ \boldsymbol{\theta}_t \leftarrow \mathrm{Proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_{t-1} - \alpha_t G(\boldsymbol{\theta}_{t-1},\mathbf{X}_t)\right). \end{cases} \tag{23}$$

For instance, consider the following standard assumption on 'uniqueness of sparse coding problem':

Assumption 4.1. For each $\mathbf{X}$ and $\boldsymbol{\theta}$, $\inf_{H\in\boldsymbol{\Theta}'\subseteq\mathbb{R}^{p\times n}} d(\mathbf{X},\boldsymbol{\theta}H) + R(H)$ admits a unique solution in $\boldsymbol{\Theta}'\subseteq\mathbb{R}^{p\times n}$.

Note that Assumption 4.1 is trivially satisfied if $R(H)$ contains a regularization $\kappa_2\|H\|_F^2$ for some $\kappa_2>2$. Under Assumption 4.1, Danskin's theorem (Bertsekas, 1997) implies that the function $\boldsymbol{\theta}\mapsto\ell(\mathbf{X},\boldsymbol{\theta})$ is differentiable and satisfies $\nabla_{\boldsymbol{\theta}}\ell(\mathbf{X},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} d(\mathbf{X},\boldsymbol{\theta}\mathbf{H}^\star)$, where $\mathbf{H}^\star$ is the unique solution of $\inf_{H\in\boldsymbol{\Theta}'\subseteq\mathbb{R}^{p\times n}} d(\mathbf{X},\boldsymbol{\theta}H)+\lambda\|H\|_1$. Hence we may choose $G(\boldsymbol{\theta}_{t-1},\mathbf{X}_t) = \nabla_{\boldsymbol{\theta}} d(\mathbf{X}_t,\boldsymbol{\theta}\mathbf{H}_t)$ in (23).

Notice that (23) is a projected SGD algorithm for the ODL problem (22), which is a constrained nonconvex problem. Zhao et al. (Zhao et al., 2017) provided asymptotic analysis of this algorithm (especially for online nonnegative matrix factorization) for general dissimilarity metric d. For a wide class of dissimilarity metrics such as Csizár f-divergence, Bregman divergence, ℓ_1 and ℓ_2 metrics, and Huber loss, this work showed that when the data matrices are i.i.d. and the stepsizes α_t are non-summable ($\sum_{t=1}^\infty\alpha_t=\infty$) and square-summable ($\sum_{t=1}^\infty\alpha_t^2<\infty$), then the sequence of dictionary matrices $(\boldsymbol{\theta}_t)_{t\geq1}$ obtained by (23), regardless of initialization, converges almost surely to the set of stationary points of (22). The asymptotic analysis uses a rather involved technique inspired from dynamical systems literature and does not provide a rate of convergence. Moreover, such asymptotic guarantees has not been available to the more general Markovian data setting.

When the function $\boldsymbol{\theta}\mapsto\ell(\mathbf{X},\boldsymbol{\theta})$ for each $\mathbf{X}$ is ρ-weakly convex for some $\rho>0$, then the expected loss function in (22) is also ρ-weakly convex, so in this case a direct application of the main result in (Davis et al., 2020) would yield a rate of convergence $O((\log t)/t^{1/4})$ for (23) with i.i.d. data matrices X_t. Such hypothesis of weak convexity of the loss function is implied under smoothness of d (Assumption L.1). Then our main results extends the theoretical guarantees for (23) to more general setting when $(\mathbf{X}_t)$ are given as a function of some underlying Markov chain with exponential mixing, and also extends to other variants of PSGD such as the AdaGrad (Algorithm 2) and the stochastic heavy ball (Algorithm 3). The full statement of this result for ODL with stochastic first-order methods on non-i.i.d. data is stated in Corollary L.2 in Appendix L. *To our best knowledge, this is the first time that projected SGD with adaptive step sizes has been applied to ODL problems*

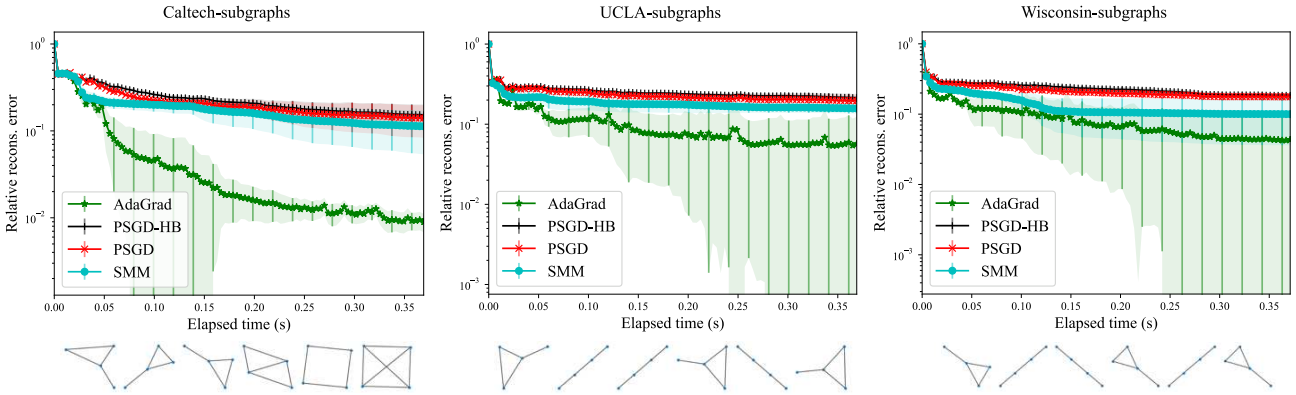

Figure 1. Plot of reconstruction error vs. elapsed time for four algorithms for online NMF: AdaGrad, PSGD-Heavy Ball, PSGD, and SMM. Data stream is a sequence of 4-node subgraph adjacency matrices sampled by an MCMC motif-sampling algorithm in (Lyu et al., 2023) from three college Facebook networks (Traud et al., 2012). Six consecutive Markovian samples of subgraphs are shown in each plot. Shaded region represents one standard deviation from ten runs.

with optimal complexity bounds for the general Markovian data case.

Numerical validation. We now illustrate the empirical performance of Alg. 1, 2, 3 to verify our theoretical findings for ODL with dependent data. However, we highlight that the main contribution of our paper is *theoretical*: obtaining the optimal complexity for constrained nonconvex problems with dependent data. Moreover, the algorithms we analyze, namely, projected SGD, SGD with momentum and AdaGrad are the default solvers in most of the libraries for machine learning/deep learning such as PyTorch/TensorFlow and their empirical success is well-established. Hence, the results here are not meant to be complete benchmarks, but rather empirical support for our theory.

For generating the samples, we used networks for 3 different schools (`Caltech36, UCLA26, Wisconsin87`) from the Facebook100 dataset (Traud et al., 2012), following a similar setup to (Lyu et al., 2020). We then used the Markov Chain Monte Carlo (MCMC) algorithm of (Lyu et al., 2023) to generate 300 correlated subgraphs from the networks. We then used the resulting matrix as a stream of Markovian data and stopped the algorithms once all the samples are used. For comparison, we used the stochastic majorization-minimization (SMM) algorithm from (Mairal et al., 2010; Lyu et al., 2020), which is the state-of-the-art algorithm for ODL problems.

In Fig. 1, we see convergence of all the algorithms with respect to the normalized reconstruction error, which is in line with our theoretical results. Moreover, we observe that AdaGrad converges significantly faster than other methods, especially for the sequence of subgraphs from `Caltech`. The difference in speed of convergence between all methods is marginal for the `UCLA` and `Wisconsin`. We suspect that this different behavior is realated to the fact that subgraphs in `Caltech` induced on random paths of $k = 4$ nodes are more likely to contain more edges than those from the other two (much sparser) networks.

5. Conclusion

In this paper, we have established convergence and complexity results of a wide range of classcial stochastic first-order methods (PSGD, AdaGrad, PSGD-Momentum) under general non-i.i.d. data sampling assumption. Our results show that if the dependence in data samples decays in the length of conditioned steps via MC mixing or Poisson equation, then standard rate of convergence in the i.i.d. case is extended to the more general non-i.i.d. case. Our analysis shows that independence between data samples is not really needed in analyzing stochastic first-order method. We also numerically verified our results on the problem of online dictionary learning from subgraph samples generated by an MCMC algorithm.

Acknowledgements

The authors are grateful to Stephen Wright for valuable discussions. The work of A. Alacaoglu was supported by NSF awards 2023239 and 2224213; and DOE ASCR Subcontract 8F-30039 from Argonne National Laboratory. The work of H. Lyu was partially supported by NSF grants DMS-2010035 and DMS-2206296.

References

Alacaoglu, A., Malitsky, Y., and Cevher, V. (2021). Convergence of adaptive algorithms for constrained weakly convex optimization. *Advances in Neural Information Processing Systems*, 34.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.

Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239.

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. (2020). Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154.

Dorfman, R. and Levy, K. Y. (2022). Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR.

Drusvyatskiy, D. and Lewis, A. S. (2018). Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948.

Drusvyatskiy, D. and Paquette, C. (2019). Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558.

Duchi, J., Hazan, E., and Singer, Y. (2010). Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. (2012). Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578.

Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M. (2015). Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE.

Ghadimi, S., Lan, G., and Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305.

Johansson, B., Rabi, M., and Johansson, M. (2007). A simple peer-to-peer algorithm for distributed optimization in sensor networks. In *2007 46th IEEE Conference on Decision and Control*, pages 4705–4710. IEEE.

Johansson, B., Rabi, M., and Johansson, M. (2010). A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Levy, K. (2017). Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30.

Li, X. and Orabona, F. (2019). On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR.

Lyu, H. (2020). Convergence of block coordinate descent with diminishing radius for nonconvex optimization. *arXiv preprint arXiv:2012.03503*.

Lyu, H. (2022). Convergence and complexity of stochastic block majorization-minimization. *arXiv preprint arXiv:2201.01652*.

Lyu, H., Memoli, F., and Sivakoff, D. (2023). Sampling random graph homomorphisms and applications to network data analysis. *Journal of Machine Learning Research*, 24:1–79.

Lyu, H., Needell, D., and Balzano, L. (2020). Online matrix factorization for markovian data and applications to network dictionary learning. *Journal of Machine Learning Research*, 21(251):1–49.

Mai, V. and Johansson, M. (2020). Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*, pages 6630–6639. PMLR.

Mairal, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. *Advances in Neural Information Processing Systems*, 26.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. (2020). Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33:16666–16676.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176.

Streeter, M. and McMahan, H. B. (2010). Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*.

Sun, T., Sun, Y., Xu, Y., and Yin, W. (2020). Markov chain block coordinate descent. *Computational Optimization and Applications*, 75(1):35–61.

Sun, T., Sun, Y., and Yin, W. (2018). On markov chain gradient descent. In *Advances in Neural Information Processing Systems*, pages 9896–9905.

Tadić, V. B. and Doucet, A. (2017). Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304.

Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180.

Tsybakov, A. B. (2004). Introduction to nonparametric estimation, 2009. *URL https://doi. org/10.1007/b13794. Revised and extended from the*, 9(10).

Wang, Y., Pan, B., Tu, W., Liu, P., Jiang, B., Gao, C., Lu, W., Jui, S., and Kong, L. (2021). Sample average approximation for stochastic optimization with dependent data: Performance guarantees and tractability. *arXiv preprint arXiv:2112.05368*.

Ward, R., Wu, X., and Bottou, L. (2019). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR.

Zhao, R., Tan, V., and Xu, H. (2017). Online nonnegative matrix factorization with general divergences. In *Artificial Intelligence and Statistics*, pages 37–45.

A. Background on stationarity measures

A.1. Direct stationarity measures

In this subsection, we introduce some notions on stationarity conditions and related quantities. A first-order necessary condition for $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ to be a first order stationary point of f over $\boldsymbol{\Theta}$ is that there exists a subgradient $v \in \partial f(\boldsymbol{\theta}^*)$ such that $-v$ belongs to the normal cone $N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) = \partial \iota_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, which is also equivalent to the variational inequality $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle v, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \geq 0$ due to the definition of the normal cone. Hence we introduce the following notion of first-order stationarity for constrained minimization problem:

$$\boldsymbol{\theta}^* \text{ is a } \textit{stationary point} \text{ of } f \text{ over } \boldsymbol{\Theta} \quad \overset{\text{def.}}{\iff} \quad \mathbf{0} \in v + N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) \quad \text{for some } v \in \partial f(\boldsymbol{\theta}^*) \tag{24}$$

$$\iff \quad \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle v, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \geq 0 \quad \text{for some } v \in \partial f(\boldsymbol{\theta}^*). \tag{25}$$

Note that if $\boldsymbol{\theta}^*$ is in the interior of $\boldsymbol{\Theta}$, then the above is equivalent to $\mathbf{0} \in \partial f(\boldsymbol{\theta}^*)$. Furthermore, if f is differentiable at $\boldsymbol{\theta}^*$, this is equivalent to $\nabla f(\boldsymbol{\theta}^*) = \mathbf{0}$, so $\boldsymbol{\theta}^*$ is a *critical point* of f.

In view of the preceding discussion and (3), we can also say that $\boldsymbol{\theta}$ is a stationary point of f over $\boldsymbol{\Theta}$ if and only if $\mathbf{0} \in \partial\varphi(\hat{\boldsymbol{\theta}})$). Accordingly, we may use $\operatorname{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}})) = 0$ as an equivalent notion of stationarity.

We relax the above first-order optimality conditions as follow: For each $\varepsilon > 0$,

$$\boldsymbol{\theta}^* \text{ is an } \varepsilon\textit{-stationary point} \text{ for } f \text{ over } \boldsymbol{\Theta} \quad \overset{\text{def.}}{\iff} \quad \operatorname{dist}(\mathbf{0}, \partial\varphi(\boldsymbol{\theta}^*)) \leq \varepsilon. \tag{26}$$

An alternative formulation of ε-stationarity would be using the 'stationarity gap'. Namely, we observe the following identity:

$$\operatorname{Gap}_{\boldsymbol{\Theta}}(f, \boldsymbol{\theta}^*) := \inf_{v\in\partial f(\boldsymbol{\theta}^*)} \left[- \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \right] = \operatorname{dist}(\mathbf{0}, \partial\varphi(\boldsymbol{\theta}^*)), \tag{27}$$

which is justified in Proposition B.1 in Appendix B. We call the quantity $\operatorname{Gap}_{\boldsymbol{\Theta}}(f, \boldsymbol{\theta}^*)$ above the *stationarity gap* at $\boldsymbol{\theta}^*$ for f over $\boldsymbol{\Theta}$. This measure of approximate stationarity was used in (Lyu, 2020; 2022), and it is also equivalent to a similar notion in (Nesterov, 2013). When $\boldsymbol{\theta}^*$ is in the interior of $\boldsymbol{\Theta}$ and if f is differentiable at $\boldsymbol{\theta}^*$, then (26) is equivalent to $\|\nabla f(\boldsymbol{\theta}^*)\| \leq \varepsilon$. In Proposition B.1, we provide an equivalent definition of ε-stationarity using the normal cone.

B. Preliminary Results

The next result illustrates the connection between the two stationarity measures given in (27) to compare with the existing result in (Lyu, 2022). Recall that the *normal cone* $N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ of $\boldsymbol{\Theta}$ at $\boldsymbol{\theta}^*$ is defined as

$$N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*) := \{u \in \mathbb{R}^p \,|\, \langle u, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0 \,\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}\}. \tag{28}$$

Note that the normal cone $N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ agrees with the subdifferential $\partial \iota_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$. When $\boldsymbol{\Theta}$ equals the whole space $\mathbb{R}^p$, then $N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \{\mathbf{0}\}$.

Proposition B.1. *For each $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, $v \in \partial f(\boldsymbol{\theta}^*)$, and $\varepsilon > 0$, following conditions are equivalent:*

(i) $\operatorname{dist}(\mathbf{0}, v + N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)) \leq \varepsilon$;

(ii) $- \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \leq \varepsilon.$

In particular, it holds that

$$\operatorname{dist}(\mathbf{0}, \partial f(\boldsymbol{\theta}^*) + N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)) = \inf_{v\in\partial f(\boldsymbol{\theta}^*)} \left[- \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \right]. \tag{29}$$

Proof. The last statement follows from the equivalence of **(i)** and **(ii)**. In order to show the equivalence, first suppose **(i)** holds. Then there exists $u \in N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$ and $w \in B_\varepsilon$ where B_ε is the ε-ball in ℓ_2 norm, such that $v + u + w = \mathbf{0}$. So $-v - w \in N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$, which is equivalent to

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v + w, \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \geq 0. \tag{30}$$

By Cauchy-Schwarz inequality, this implies

$$-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v,\, \frac{\boldsymbol{\theta}-\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|} \right\rangle \le \|w\| \le \varepsilon. \tag{31}$$

Conversely, suppose **(ii)** holds. We let $\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)$ denote the set of all feasible directions at $\boldsymbol{\theta}^*$ of norm bounded by 1, which consists of vectors of form $a(\boldsymbol{\theta}-\boldsymbol{\theta}^*)$ for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $a\in(0,\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^{-1}]$. Being the intersection of two convex sets, $\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)$ is convex. Then applying the minimax theorem (Sion, 1958) for the bilinear map $(x,u)\mapsto\langle v+\varepsilon u, x\rangle$ defined on the product of convex sets $\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)\times B_1$, observe that

$$\begin{aligned}
\sup_{u\in B_1}\inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)} \langle v+\varepsilon u,\, x\rangle &= \inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)}\sup_{u\in B_1} \langle v+\varepsilon u,\, x\rangle && (32)\\
&= \inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)} \left[\langle v,\, x\rangle + \sup_{u\in B_1}\langle \varepsilon u,\, x\rangle\right] && (33)\\
&= \inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)} \left[\langle v,\, x\rangle + \varepsilon\|x\|\right] && (34)\\
&= \inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)} \|x\|\left[\left\langle v,\, \frac{x}{\|x\|}\right\rangle + \varepsilon\right] \ge 0. && (35)
\end{aligned}$$

To see the last inequality, fix $x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)$. By definition, there exists some $\boldsymbol{\theta}_x\in\boldsymbol{\Theta}$ such that $x/\|x\| = \frac{\boldsymbol{\theta}_x-\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}_x-\boldsymbol{\theta}^*\|}$. Then by using **(ii)**,

$$\left\langle v,\, \frac{x}{\|x\|}\right\rangle + \varepsilon = \left\langle v,\, \frac{\boldsymbol{\theta}_x-\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}_x-\boldsymbol{\theta}^*\|}\right\rangle + \varepsilon \ge \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}} \left\langle v,\, \frac{\boldsymbol{\theta}-\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}\right\rangle + \varepsilon \ge -\varepsilon+\varepsilon \ge 0. \tag{36}$$

Attainment of the supremum at a u^* in (32) is guaranteed by strong duality, see (Bauschke and Combettes, 2011).

The above implies

$$\inf_{x\in\mathcal{D}_{\le 1}(\boldsymbol{\theta}^*)} \langle v+\varepsilon u^*,\, x\rangle \ge 0. \tag{37}$$

Thus we conclude that $-v-\varepsilon u^*\in N_{\boldsymbol{\Theta}}(\boldsymbol{\theta}^*)$. Then **(i)** holds since $\|u^*\|\le 1$. □

Proposition B.2. *Suppose f is ρ-weakly convex and $\lambda<\rho^{-1}$. Then for each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$,*

$$\sup_{v\in\partial f(\hat{\boldsymbol{\theta}})}\left[-\inf_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}\setminus\{\hat{\boldsymbol{\theta}}\}}\left\langle v(\hat{\boldsymbol{\theta}}),\, \frac{\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\|}\right\rangle\right] \le \lambda^{-1}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\| \le \lambda^{-2}\|\nabla\varphi_\lambda(\boldsymbol{\theta})\| \tag{38}$$

Proof. Recall that $\hat{\boldsymbol{\theta}}$ is the solution of a constrained optimization problem since $\varphi = f+\iota_{\boldsymbol{\Theta}}$ (5). Therefore, it satisfies the following first-order optimality condition: For some $v(\hat{\boldsymbol{\theta}})\in\partial f(\hat{\boldsymbol{\theta}})$,

$$\langle v(\hat{\boldsymbol{\theta}})+\lambda^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}),\, \boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\rangle \ge 0, \qquad \forall\boldsymbol{\theta}'\in\boldsymbol{\Theta}. \tag{39}$$

By rearranging and using Cauchy-Schwarz, this yields for all $\boldsymbol{\theta}'\in\boldsymbol{\Theta}$,

$$\langle v(\hat{\boldsymbol{\theta}}),\, \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}'\rangle \le \lambda^{-1}\langle\hat{\boldsymbol{\theta}}-\boldsymbol{\theta},\, \boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\rangle \le \lambda^{-1}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\cdot\|\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\|, \tag{40}$$

Now assume $\boldsymbol{\theta}'\neq\hat{\boldsymbol{\theta}}$. Dividing both sides by $\|\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\|$, we get

$$-\left\langle v(\hat{\boldsymbol{\theta}}),\, \frac{\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}}{\|\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}\|}\right\rangle \le \lambda^{-1}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\| = \lambda^{-2}\|\nabla\varphi_\lambda(\boldsymbol{\theta})\|. \tag{41}$$

Since this holds for all $v(\hat{\boldsymbol{\theta}})\in\partial f(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{\theta}'\in\boldsymbol{\Theta}\setminus\{\hat{\boldsymbol{\theta}}\}$, the assertion follows. □

The next two results will be used in Lem. 2.4 to control the bias due to dependent data.

Lemma B.3 ((Rockafellar and Wets, 2009)). *For any $\hat{\rho} \geq \rho$ and ρ-weakly convex function φ, it follows that $\boldsymbol{\theta} \mapsto \operatorname{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta})$ is $\frac{\hat{\rho}}{\hat{\rho}-\rho}$-Lipschitz.*

Lemma B.4 ((Alacaoglu et al., 2021)). *Let $\hat{\rho} \geq \rho$. Then for any $v \in \partial f(\boldsymbol{\theta})$,*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{2\|v\|}{\hat{\rho} - \rho}.$$

The following lemma is used in converting various finite total variation results into rate of convergence or asymptotic convergence results.

Lemma B.5 (Lem. A.5 in (Mairal, 2013)). *Let $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be sequences of nonnegative real numbers such that $\sum_{n=0}^{\infty} a_n b_n < \infty$. Then the following hold.*

(i) $\min_{1\leq k\leq n} b_k \leq \dfrac{\sum_{k=0}^{\infty} a_k b_k}{\sum_{k=1}^{n} a_k} = O\left(\left(\sum_{k=1}^{n} a_k\right)^{-1}\right)$.

(ii) *Further assume $\sum_{n=0}^{\infty} a_n = \infty$ and $|b_{n+1} - b_n| = O(a_n)$. Then $\lim_{n\to\infty} b_n = 0$.*

Proof. **(i)** follows from noting that

$$\left(\sum_{k=1}^{n} a_k\right) \min_{1\leq k\leq n} b_k \leq \sum_{k=1}^{n} a_k b_k \leq \sum_{k=1}^{\infty} a_k b_k < \infty. \tag{42}$$

The proof of **(ii)** is omitted and can be found in (Mairal, 2013). □

The next lemma is commonly used for adaptive gradient algorithms. For example, Lem. A.1 in (Levy, 2017) or Lem. 12 in (Duchi et al., 2010).

Lemma B.6 (Lem. 12 in (Duchi et al., 2010), Lem. A.1 in (Levy, 2017)). *For nonnegative real numbers a_i for $i \geq 1$, we have for any $v_0 > 0$*

$$\sum_{i=1}^{n} \frac{a_i}{v_0 + \sum_{j=1}^{i} a_j} \leq \log\left(1 + \frac{\sum_{i=1}^{n} a_i}{v_0}\right) \quad \text{and} \quad \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}.$$

The following uniform concentration lemma for vector-valued parameterized observables is due to (Lyu, 2022).

Lemma B.7 (Lem 7.1 in (Lyu, 2022)). *Fix compact subsets $\mathcal{X} \subseteq \mathbb{R}^q$, $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ and a bounded Borel measurable function $\psi : \mathcal{X} \times \boldsymbol{\Theta} \to \mathbb{R}^r$. Let $(\mathbf{x}_n)_{n\geq 1}$ denote a sequence of points in $\mathcal{X}$ such that $\mathbf{x}_n = \varphi(X_n)$ for $n \geq 1$, where $(X_n)_{n\geq 1}$ is a Markov chain on a state space Ω and $\varphi : \Omega \to \mathcal{X}$ is a measurable function. Assume the following:*

(a1) *The Markov chain $(X_n)_{n\geq 1}$ mixes exponentially fast to its unique stationary distribution and the stochastic process $(\mathbf{x}_n)_{n\geq 1}$ on $\mathcal{X}$ has a unique stationary distribution π.*

Suppose $w_n \in (0, 1]$, $n \geq 1$ are non-increasing and satisfy $w_n^{-1} - w_{n-1}^{-1} \leq 1$ for all $n \geq 1$. Define functions $\bar{\psi}(\cdot) := \mathbb{E}_{\mathbf{x}\sim\pi}[\psi(\mathbf{x}, \cdot)]$ and $\bar{\psi}_n : \boldsymbol{\Theta} \to \mathbb{R}^r$ recursively as $\bar{\psi}_0 \equiv \mathbf{0}$ and

$$\bar{\psi}_n(\cdot) = (1 - w_n)\bar{\psi}_{n-1}(\cdot) + w_n \psi(\mathbf{x}_n, \cdot). \tag{43}$$

Then fthere exists a constant $C > 0$ such that for all $n \geq 1$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\bar{\psi}(\boldsymbol{\theta}) - \mathbb{E}[\bar{\psi}_n(\boldsymbol{\theta})]\right\| \leq C w_n, \quad \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta})\right\|\right] \leq C w_n \sqrt{n}. \tag{44}$$

Furthermore, if $w_n\sqrt{n} = O(1/(\log n)^{1+\varepsilon})$ for some $\varepsilon > 0$, then $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\bar{\psi}(\boldsymbol{\theta}) - \bar{\psi}_n(\boldsymbol{\theta})\right\| \to 0$ as $t \to \infty$ almost surely.

C. Proof for Sections 2

In this section, we prove the key lemma (Lemma 2.4) we stated in the main text. For the reader's convenience, we restate the key lemma here:

Lemma C.1 (Lemma 2.4 in the main text)**.** *Let Assumptions 2.1, 2.2, 2.3 hold and $\boldsymbol{\theta}_t$ be generated according to Algorithm 1, 2 or 3. Fix $t \geq 0$, $k = k_t \in [0, t]$ as in Assumption 2.2, $\hat{\rho} > \rho$ and denote $\hat{\boldsymbol{\theta}} = \mathrm{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta})$. Then*

$$\left| \mathbb{E}\left[\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1}) \rangle \,|\, \mathcal{F}_{t-k} \right] - \langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t, \mathbf{x}) \right] \rangle \right| \tag{45}$$

$$\leq \frac{4L^2}{\hat{\rho} - \rho}\, \Delta_{[t-k,t]} + \frac{2L(L_1 + \hat{\rho})}{\hat{\rho} - \rho}\, \mathbb{E}\left[\sum_{s=t-k}^{t-1} \alpha_s \| G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1}) \| \,\middle|\, \mathcal{F}_{t-k} \right]. \tag{46}$$

For the proofs in this section, we use the following notations. Let $\pi_{t+1|t-k} = \pi_{t+1|t-k}(\cdot \,|\, \mathcal{F}_{t-k})$ denote the distribution of $\mathbf{x}_{t+1}$ conditional on the information $\mathcal{F}_{t-k} = \sigma(\mathbf{x}_1, \ldots, \mathbf{x}_{t-k})$. Also, $\mathbb{E}_{\mathbf{x}\sim\mu}$ will denote the expectation only with respect to the random variable $\mathbf{x}$ distributed as μ, leaving out any other random variable fixed.

The following proposition is an important ingredient for the proof of Lemma 2.4. It allows us to compare a multi-step conditional expectation of the stochastic gradient to its stationary expectation.

Proposition C.2. *Let Assumptions 2.1, 2.2, 2.3 hold and $\boldsymbol{\theta}_t$ be generated according to Algorithm 1, 2 or 3. Suppose $\lim_{N\to\infty} \|\pi_{t+N|t} - \pi\|_{TV} = 0$ for all $t \geq 0$. Fix $t \geq 0$ and $k \in [0, t]$. Then*

$$\| \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] - \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] \| \leq \begin{cases} 2L\Delta_{[t-k,t+1]} & \text{if Assumption 2.3(i) holds;} \\ 2L\Delta_{[t-k,t]} & \text{if Assumption 2.3(ii) holds.} \end{cases} \tag{47}$$

Proof. Recall that by Scheffé's lemma, if two probability measures μ and ν on the same probability space have densities α and β with respect to a reference measure dm, then $\|\mu - \nu\|_{TV} = \frac{1}{2}\int |\alpha - \beta|\, dm$ (see, e.g., Lemma 2.1 in (Tsybakov, 2004)). For each integer $m \geq 0$, let $\pi'_{t+m|t-k}$ denote the density functions of $\pi_{t+m|t-k}$ with respect to the Lebesgue measure, which we denote by $d\xi$.

We first prove the statement under Assumption 2.3(ii). In this case, $\|G(\boldsymbol{\theta}, \mathbf{x})\|$ is assumed to be uniformly bounded by $L < \infty$ but we do not impose any additional assumption on the data samples $(\mathbf{x}_t)_{t\geq 0}$ besides the asymptotic mixing condition $\lim_{N\to\infty} \|\pi_{t+N|t} - \pi\|_{TV} = 0$ for all $t \geq 0$.

Fix an integer $N \geq 1$. Noting that $\boldsymbol{\theta}_{t-k}$ is deterministic with respect to $\mathcal{F}_{t-k}$, we have

$$\| \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] \| \tag{48}$$

$$= \| \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] - \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+N}) \,|\, \mathcal{F}_{t-k}] \| \tag{49}$$

$$= \| \mathbb{E}_{\mathbf{x}\sim\pi_{t|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N-1|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] \| \tag{50}$$

$$\leq \int_\Omega \|G(\boldsymbol{\theta}_{t-k}, \mathbf{x})\| |\pi'_{t+1|t-k}(\mathbf{x}) - \pi'_{t+N|t-k}(\mathbf{x})|\, d\xi \tag{51}$$

$$\leq 2L \|\pi_{t+1|t-k} - \pi_{t+N|t-k}\|_{TV}, \tag{52}$$

where we have used Scheffé's lemma for the last equality. By a similar argument,

$$\| \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] \| \leq \int_\Omega \|G(\boldsymbol{\theta}_{t-k}, \mathbf{x})\| |\pi'(\mathbf{x}) - \pi'_{t+N|t-k}(\mathbf{x})|\, d\xi \tag{53}$$

$$\leq 2L \|\pi - \pi_{t+N|t-k}\|_{TV}. \tag{54}$$

By triangle inequality, it then follows

$$\| \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] \| \leq 2L \left(\|\pi - \pi_{t+N|t-k}\|_{TV} + \|\pi_{t+1|t-k} - \pi_{t+N|t-k}\|_{TV} \right). \tag{55}$$

Now by the hypothesis that $\lim_{N\to\infty} \|\pi_{t+N|t-k} - \pi\|_{TV} = 0$, the last expression above converges to $2L\|\pi_{t+1|t-k} - \pi\|_{TV}$ as $N \to \infty$. Since the left hand side above does not depend on N, this shows the claim (47).

Next, we prove the statement under Assumption 2.3(i). In this case, we only assume the one-step conditional expectation of the norm of the stochastic gradient is bounded:

$$\mathbb{E}[\|G(\boldsymbol{\theta}, \mathbf{x}_{t+1}\| \,|\, \mathcal{F}_t] \leq L \quad \text{for all } t \geq 0, \tag{56}$$

which is much weaker than the uniform boundedness of $\|G\|$ in Assumption 2.3(ii). In order to handle a technical difficulty in this general setting, we will need to assume that the data samples $(\mathbf{x}_t)_{t\geq 0}$ is a function of some time-homogeneous Markov chain. That is, there exists a time-homogeneous Markov chain $(X_t)_{t\geq 0}$ on some state space $\mathfrak{X}$ and a function $w : \mathfrak{X} \to \Omega$ such that $\mathbf{x}_t = w(X_t)$ for all $t \geq 0$. By the time-homogeneity of the chain $(X_t)_{t\geq 0}$, there exists a Markov transition kernel P such that

$$\mathbb{P}(X_{t+1} = b \,|\, X_t = a) \equiv P(a, b) \qquad \text{for all } t \geq 0 \text{ and } a, b \in \mathfrak{X}. \tag{57}$$

We will proceed similarly as before. The key technical detail to avoid using uniform boundedness of G is to rewrite expectations of G by the expectation of a one-step conditional expectation of G. Then a similar argument as before will work only with the assumption that the one-step conditional expectation of G is bounded. We give the details of this sketched approach below.

Fix an integer $N \geq 1$. Since the conditional expectation $\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+m}) \,|\, \mathcal{F}_{t-k}]$ is deterministic with respect to $\mathcal{F}_{t-k}$, we can write

$$\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] = \mathbb{E}[\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_t] \,|\, \mathcal{F}_{t-k}]. \tag{58}$$

Similarly, write

$$\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+N}) \,|\, \mathcal{F}_{t-k}] = \mathbb{E}[\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+N}) \,|\, \mathcal{F}_{t+N-1}] \,|\, \mathcal{F}_{t-k}]. \tag{59}$$

Now for given $s \geq 0$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\mathbf{x} \in \Omega$, define

$$\tilde{G}_s(\boldsymbol{\theta}, \mathbf{x}) := \mathbb{E}[G(\boldsymbol{\theta}, \mathbf{x}_{s+1}) \,|\, \mathbf{x}_s = \mathbf{x}]. \tag{60}$$

The only randomness being integrated out in the expectation in the above definition is the random data sample $\mathbf{x}_{s+1}$ conditional on the data sample $\mathbf{x}_s$ a step before being $\mathbf{x}$. The measure used in the integral is the one-step conditional distribution $\pi_{s+1|s}$. By the time-homogeneity assumption, the distribution $\pi_{s+1|s}$ does not depend on s. It follows that the function $\tilde{G}_s$ above does not depend on s. Therefore, we will omit the subscript s in $\tilde{G}_s$. Note that by Jensen's inequality and (56),

$$\|\tilde{G}(\boldsymbol{\theta}, \mathbf{x})\| \leq \mathbb{E}[\|G(\boldsymbol{\theta}, \mathbf{x}_{s+1})\| \,|\, \mathbf{x}_s = \mathbf{x}] \leq L. \tag{61}$$

Using (58) and (59), we have

$$\begin{aligned}
&\|\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})]\| && (62)\\
&\quad = \|\mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}] - \mathbb{E}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+N}) \,|\, \mathcal{F}_{t-k}]\| && (63)\\
&\quad = \|\mathbb{E}_{\mathbf{x}\sim\pi_{t|t-k}}[\tilde{G}_t(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N-1|t-k}}[\tilde{G}_{t+N-1}(\boldsymbol{\theta}_{t-k}, \mathbf{x})]\| && (64)\\
&\quad = \|\mathbb{E}_{\mathbf{x}\sim\pi_{t|t-k}}[\tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N-1|t-k}}[\tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})]\| && (65)\\
&\quad = \left\| \int_\Omega \tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})(\pi'_{t|t-k}(\mathbf{x}) - \pi'_{t+N-1|t-k}(\mathbf{x}))\, d\xi \right\| && (66)\\
&\quad \leq \int_\Omega \|\tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})\| |\pi'_{t|t-k}(\mathbf{x}) - \pi'_{t+N-1|t-k}(\mathbf{x})|\, d\xi && (67)\\
&\quad \leq 2L\|\pi_{t|t-k} - \pi_{t+N-1|t-k}\|_{TV}, && (68)
\end{aligned}$$

where we have used Scheffé's lemma and the fact that $\tilde{G}_s$ does not depend on s. By a similar argument and noting that $\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] = \mathbb{E}_{\mathbf{x}_{t-k}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t-k+1})] = \mathbb{E}_{\mathbf{x}_{t-k}\sim\pi}[\mathbb{E}\left[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t-k+1})\right] \,|\, \mathbf{x}_{t-k}]$,

$$\begin{aligned}
&\|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t-k}}[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})]\| && (69)\\
&\quad = \|\mathbb{E}_{\mathbf{x}\sim\pi}[\tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N-1|t-k}}[\tilde{G}(\boldsymbol{\theta}_{t-k}, \mathbf{x})]\| && (70)\\
&\quad \leq 2L\|\pi - \pi_{t+N-1|t-k}\|_{TV}. && (71)
\end{aligned}$$

By triangle inequality, it then follows

$$\|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_{t-k},\mathbf{x})] - \mathbb{E}[G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\,|\,\mathcal{F}_{t-k}]\| \le 2L\left(\|\pi-\pi_{t+N-1|t-k}\|_{TV} + \|\pi_{t|t-k}-\pi_{t+N-1|t-k}\|_{TV}\right). \quad (72)$$

Now by the hypothesis $\lim_{N\to\infty}\|\pi_{t+N-1|t-k}-\pi\|_{TV}=0$, so the last expression above converges to $2L\|\pi_{t|t-k}-\pi\|_{TV}$ as $N\to\infty$. Since the left hand side above does not depend on N, this shows (47). □

We now prove Lemma 2.4.

Proof of Lemma 2.4. Denote $V(\mathbf{x},\boldsymbol{\theta}) := \langle\hat{\boldsymbol{\theta}}-\boldsymbol{\theta},\, G(\boldsymbol{\theta},\mathbf{x})\rangle$. Note that $\mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x},\boldsymbol{\theta}_t)\right] = \langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t,\mathbf{x})\right]\rangle$. Observe that by triangle inequality

$$\begin{aligned}
&|\mathbb{E}\left[V(\mathbf{x}_{t+1},\boldsymbol{\theta}_t)\,|\,\mathcal{F}_{t-k}\right] - \mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x},\boldsymbol{\theta}_t)\right]| && (73)\\
&\quad\le |\mathbb{E}\left[V(\mathbf{x}_{t+1},\boldsymbol{\theta}_t) - V(\mathbf{x}_{t+1},\boldsymbol{\theta}_{t-k})\,|\,\mathcal{F}_{t-k}\right]| && (74)\\
&\qquad + |\mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x},\boldsymbol{\theta}_{t-k}) - V(\mathbf{x},\boldsymbol{\theta}_t)\right]| && (75)\\
&\qquad\quad + |\mathbb{E}\left[V(\mathbf{x}_{t+1},\boldsymbol{\theta}_{t-k})\,|\,\mathcal{F}_{t-k}\right] - \mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x},\boldsymbol{\theta}_{t-k})\right]|\,. && (76)
\end{aligned}$$

We will bound the three terms in the right in order.

In order to bound the first term in the right hand side above, we first write

$$\begin{aligned}
V(\mathbf{x}_{t+1},\boldsymbol{\theta}_t) - V(\mathbf{x}_{t+1},\boldsymbol{\theta}_{t-k}) &= \langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t, G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle - \langle\hat{\boldsymbol{\theta}}_{t-k}-\boldsymbol{\theta}_{t-k}, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle\\
&= \langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t, G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1}) - G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle + \langle\hat{\boldsymbol{\theta}}_t-\hat{\boldsymbol{\theta}}_{t-k}, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle\\
&\qquad + \langle\boldsymbol{\theta}_{t-k}-\boldsymbol{\theta}_t, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle.
\end{aligned}$$

By applying iterated expectation twice, we get

$$\begin{aligned}
&\mathbb{E}\left[\langle\boldsymbol{\theta}_{t-k}-\boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle\,|\,\mathcal{F}_{t-k}\right]\\
&\quad= \mathbb{E}_{\boldsymbol{\theta}_t}\left[\mathbb{E}\left[\langle\boldsymbol{\theta}_{t-k}-\boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle\,|\,\boldsymbol{\theta}_t,\,\mathcal{F}_{t-k}\right]\,|\,\mathcal{F}_{t-k}\right]\\
&\quad= \mathbb{E}_{\boldsymbol{\theta}_t}\left[\langle\boldsymbol{\theta}_{t-k}-\boldsymbol{\theta}_t,\, \mathbb{E}\left[G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\,|\,\boldsymbol{\theta}_t,\,\mathcal{F}_{t-k}\right]\rangle\,|\,\mathcal{F}_{t-k}\right]\\
&\quad= \mathbb{E}\left[\langle\boldsymbol{\theta}_{t-k}-\boldsymbol{\theta}_t,\, \mathbb{E}\left[G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\,|\,\boldsymbol{\theta}_t,\,\mathcal{F}_{t-k}\right]\rangle\,|\,\mathcal{F}_{t-k}\right]. \qquad (77)
\end{aligned}$$

We can rewrite the conditional expectation $\mathbb{E}[\langle\hat{\boldsymbol{\theta}}_{t-k}-\hat{\boldsymbol{\theta}}_t,\, G(\boldsymbol{\theta}_{t-k},\mathbf{x}_{t+1})\rangle\,|\,\mathcal{F}_{t-k}]$ similarly as above.

Next, we will observe that

$$\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\| \le \frac{2}{\hat{\rho}-\rho}\|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x})]\| \le \frac{2L}{\hat{\rho}-\rho}. \quad (78)$$

The first inequality above is due to Lemma B.4. Under Assumption 2.3(ii), where since $\|G\|\le L$, the second inequality follows by using Jensen's inequality. In case of Assumption 2.3(i), we need a bit more careful argument. For each $N\ge 1$,

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\| &\le \frac{2}{\hat{\rho}-\rho}\|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x})]\| && (79)\\
&\le \frac{2}{\hat{\rho}-\rho}\left(\|\mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t}}[G(\boldsymbol{\theta}_t,\mathbf{x})]\| + \|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t}}[G(\boldsymbol{\theta}_t,\mathbf{x})]\|\right) && (80)\\
&\le \frac{2}{\hat{\rho}-\rho}\left(\mathbb{E}[\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+N})\|\,|\,\mathcal{F}_t] + \|\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\pi_{t+N|t}}[G(\boldsymbol{\theta}_t,\mathbf{x})]\|\right). && (81)
\end{aligned}$$

Note that $\mathbb{E}[\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+N})\|\,|\,\mathcal{F}_t]\le L$ by Assumption 2.3(ii) and iterated expectation. Furthermore, the second term in the last expression above vanishes as $N\to\infty$ by Proposition C.2 and Assumption 2.2. Therefore we can conclude (78) under Assumption 2.1(ii) as well.

Now by using Cauchy-Schwarz inequality, L_1-Lipschitz continuity of $\boldsymbol{\theta} \mapsto G(\mathbf{x}, \boldsymbol{\theta})$ (see Assumption 2.1), Lemma B.3 and Assumption 2.3, we obtain

$$|\mathbb{E}[V(\mathbf{x}_{t+1}, \boldsymbol{\theta}_t) - V(\mathbf{x}_{t+1}, \boldsymbol{\theta}_{t-k}) \,|\, \mathcal{F}_{t-k}]| \leq \frac{2LL_1 + 2\hat{\rho}L}{\hat{\rho} - \rho} \mathbb{E}\left[\|\boldsymbol{\theta}_{t-k} - \boldsymbol{\theta}_t\| \,|\, \mathcal{F}_{t-k}\right] \tag{82}$$

$$\leq \frac{2LL_1 + 2\hat{\rho}L}{\hat{\rho} - \rho} \mathbb{E}\left[\sum_{s=t-k}^{t-1} \alpha_s \|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| \,\middle|\, \mathcal{F}_{t-k},\right], \tag{83}$$

where for the last inequality we have used $\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s-1}\| \leq \|\alpha_{s-1} G(\boldsymbol{\theta}_{s-1}, \mathbf{x}_s)\|$ for $s \geq 1$ along with triangle inequality.

A similar argument shows

$$|\mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x}, \boldsymbol{\theta}_t) - V(\mathbf{x}, \boldsymbol{\theta}_{t-k})\right]| \leq \frac{2LL_1 + 2\hat{\rho}L}{\hat{\rho} - \rho} \mathbb{E}\left[\sum_{s=t-k}^{t-1} \alpha_s \|G(\boldsymbol{\theta}_{s-1}, \mathbf{x}_s)\| \,|\, \mathcal{F}_{t-k}\right]. \tag{84}$$

We continue by estimating the last term on the RHS of (76).

Proceeding by Cauchy-Schwarz inequality, and using that $\boldsymbol{\theta}_{t-k}, \hat{\boldsymbol{\theta}}_{t-k}$ are deterministic with respect to $\mathcal{F}_{t-k}$, we get

$$|\mathbb{E}\left[V(\mathbf{x}_{t+1}, \boldsymbol{\theta}_{t-k}) \,|\, \mathcal{F}_{t-k}\right] - \mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x}, \boldsymbol{\theta}_{t-k})\right]| \tag{85}$$

$$= \left|\mathbb{E}\left[\langle \hat{\boldsymbol{\theta}}_{t-k} - \boldsymbol{\theta}_{t-k},\, G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1})\rangle \,|\, \mathcal{F}_{t-k}\right] - \mathbb{E}_{\mathbf{x}\sim\pi}\left[V(\mathbf{x}, \boldsymbol{\theta}_{t-k})\right]\right| \tag{86}$$

$$= \left|\langle \hat{\boldsymbol{\theta}}_{t-k} - \boldsymbol{\theta}_{t-k},\, \mathbb{E}\left[G(\boldsymbol{\theta}_{t-k}, \mathbf{x}_{t+1}) \,|\, \mathcal{F}_{t-k}\right]\rangle - \langle \hat{\boldsymbol{\theta}}_{t-k} - \boldsymbol{\theta}_{t-k},\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_{t-k}, \mathbf{x})\right]\rangle\right| \tag{87}$$

$$\leq 2L\|\hat{\boldsymbol{\theta}}_{t-k} - \boldsymbol{\theta}_{t-k}\|\, \Delta_{[t-k,t]}\} \tag{88}$$

$$\leq \frac{4L^2}{\hat{\rho} - \rho} \Delta_{[t-k,t]}, \tag{89}$$

where the last step follows from Proposition C.2. Combining (83), (84), (89) with (76) then shows the assertion. □

D. Proof for Section 3.1

Theorem D.1 (Theorem 3.1 in the main text)**.** *Let Assumptions 2.1-2.3 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1. Fix $\hat{\rho} > \rho$. Then the following hold:*

(i) *(Rate of convergence) For each $T \geq 1$,*

$$E\left[\|\nabla \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] \tag{90}$$

$$\leq \frac{\hat{\rho}^2 L^2}{\hat{\rho} - \rho} \frac{1}{\sum_{k=1}^T \alpha_k} \left[\frac{\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_1) - \inf \varphi_{1/\hat{\rho}}}{\hat{\rho} L^2} + \frac{1}{2}\sum_{t=1}^T \alpha_t^2 + \frac{2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \sum_{t=1}^T k_t \alpha_t \alpha_{t-k_t} + \frac{4}{\hat{\rho} - \rho} \sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t]}]\right]. \tag{91}$$

In particular, with $\alpha_t = \frac{c}{\sqrt{t}}$ for some $c > 0$ and under exponential mixing, we have that $\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|\right] \leq \varepsilon$ with $\tilde{O}\left(\varepsilon^{-4}\right)$ samples.

(ii) *(Global convergence) Further assume that $\sum_{t=0}^{\infty} k_t \alpha_t \alpha_{t-k_t} < \infty$. Then $\|\nabla \varphi_{1/\hat{\rho}}(\hat{\boldsymbol{\theta}}_t)\| \to 0$ as $t \to \infty$ almost surely. Furthermore, $\boldsymbol{\theta}_t$ converges to the set of all stationary points of f over $\boldsymbol{\Theta}$.*

Proof. Recall the definition of $\varphi_{1/\hat{\rho}}$ from (4). We start as in (Davis and Drusvyatskiy, 2019) with the difference of conditoning on $\mathcal{F}_{t-k}$ instead of the latest iterate, and use Lemma 2.4 to handle the additional bias due to dependent sampling.

Denote $\hat{\boldsymbol{\theta}}_t = \text{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta}_t)$ for $t \geq 1$ and fix $k \in \{0, \ldots, t\}$. Observe that

$$\mathbb{E}\Big[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1}) \,\Big|\, \mathcal{F}_{t-k}\Big] \leq \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_{t+1} - \hat{\boldsymbol{\theta}}_t\|^2 \,\middle|\, \mathcal{F}_{t-k}\right] \tag{92}$$

$$= \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) \,\middle|\, \mathcal{F}_{t-k}\right] + \frac{\hat{\rho}}{2}\mathbb{E}\left[\Big\|\text{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})) - \text{proj}_{\boldsymbol{\Theta}}(\hat{\boldsymbol{\theta}}_t)\Big\|^2 \,\middle|\, \mathcal{F}_{t-k}\right] \tag{93}$$

$$\leq \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) \,\middle|\, \mathcal{F}_{t-k}\right] + \frac{\hat{\rho}}{2}\mathbb{E}\left[\Big\|(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t) - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\Big\|^2 \,\middle|\, \mathcal{F}_{t-k}\right] \tag{94}$$

$$\leq \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2 \,\middle|\, \mathcal{F}_{t-k}\right] + \hat{\rho}\alpha_t\mathbb{E}\left[\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\rangle \,\middle|\, \mathcal{F}_{t-k}\right] + \frac{\alpha_t^2\hat{\rho}L^2}{2} \tag{95}$$

$$\leq \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t) \,\middle|\, \mathcal{F}_{t-k}\right] + \hat{\rho}\alpha_t\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t, \mathbf{x})\right]\rangle + \frac{\alpha_t^2\hat{\rho}L^2}{2} \tag{96}$$

$$+ \hat{\rho}\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\,\mathbb{E}\left[\Delta_{[t-k,t]}\right] + \frac{2L(L_1+\hat{\rho})}{\hat{\rho}-\rho}\alpha_{t-k}\sum_{s=t-k}^{t-1}\mathbb{E}\left[\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| \,|\, \mathcal{F}_{t-k}\right]\right). \tag{97}$$

Namely, the first and the last inequalities use the definition of Moreau envelope $\varphi_{1/\hat{\rho}}$ and $\hat{\boldsymbol{\theta}}_t \in \boldsymbol{\Theta}$, the second inequality uses 1-Lipschitzness of the projection operator, and the last inequality uses Lemma 2.4 and that α_s is non-increasing in s. Note that using iterated expectation, Assumption 2.3, and the fact that $\boldsymbol{\theta}_s$ is deterministic with respect to $\mathcal{F}_s$, for each $t-k \leq s \leq t-1$, we get

$$\mathbb{E}\left[\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| \,|\, \mathcal{F}_{t-k}\right] = \mathbb{E}\left[\mathbb{E}\left[\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| \,|\, \mathcal{F}_s\right] \,|\, \mathcal{F}_{t-k}\right] \leq L. \tag{98}$$

Hence the summation in the last term above is bounded above by kL. Then by using Assumption 2.1 and the weak convexity of g, we have

$$\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, \mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t, \mathbf{x})\right]\rangle \leq f(\hat{\boldsymbol{\theta}}_t) - f(\boldsymbol{\theta}_t) + \frac{\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2. \tag{99}$$

By using this estimate in (97) and then integrating out $\mathcal{F}_{t-k}$, we get

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\right] - \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\right] \leq \hat{\rho}\alpha_t\mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) - f(\boldsymbol{\theta}_t) + \frac{\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2\right] + \frac{\alpha_t^2\hat{\rho}L^2}{2} \tag{100}$$

$$+ \hat{\rho}\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\,\mathbb{E}[\Delta_{[t-k,t]}] + k\frac{2L^2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\alpha_{t-k}\right). \tag{101}$$

Now we chose $k = k_t \to \infty$ as $t \to \infty$. Summing over $t = 1, \ldots, T$ results in

$$\hat{\rho}\sum_{t=1}^{T}\alpha_t\mathbb{E}\Big[f(\boldsymbol{\theta}_t) - f(\hat{\boldsymbol{\theta}}_t) - \frac{\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2\Big] \leq \left(\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_1) - \inf\varphi_{1/\hat{\rho}}\right) + \frac{\hat{\rho}L^2}{2}\sum_{t=1}^{T}\alpha_t^2 \tag{102}$$

$$+ \frac{4\hat{\rho}L^2}{\hat{\rho}-\rho}\sum_{t=1}^{T}\alpha_t\mathbb{E}[\Delta_{[t-k_t,t]}] + \frac{2L^2\hat{\rho}(L_1+\hat{\rho})}{\hat{\rho}-\rho}\sum_{t=1}^{T}k_t\alpha_t\alpha_{t-k_t}. \tag{103}$$

Next, we use the fact that the function $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta}) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2$ is strongly convex with parameter $(\hat{\rho}-\rho)/2$ that is minimized at $\hat{\boldsymbol{\theta}}_t$ to get

$$f(\boldsymbol{\theta}_t) - f(\hat{\boldsymbol{\theta}}_t) - \frac{\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2 = \left(f(\boldsymbol{\theta}_t) + \frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\|^2\right) - \left(f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\|^2\right) + \frac{\hat{\rho}-\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2 \tag{104}$$

$$\geq (\hat{\rho}-\rho)\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2 \tag{105}$$

$$= \frac{\hat{\rho}-\rho}{\hat{\rho}^2}\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2. \tag{106}$$

where the second to the last equality uses (6). Combining with (103), this implies

$$\frac{\hat{\rho}-\rho}{\hat{\rho}}\sum_{t=1}^{T}\alpha_t\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right] \le \left(\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_1)-\inf\varphi_{1/\hat{\rho}}\right)+\frac{\hat{\rho}L^2}{2}\sum_{t=1}^{T}\alpha_t^2 \tag{107}$$

$$+\frac{4\hat{\rho}L^2}{\hat{\rho}-\rho}\sum_{t=1}^{T}\alpha_t\mathbb{E}[\Delta_{[t-k_t,t]}]+\frac{2L^2\hat{\rho}(L_1+\hat{\rho})}{\hat{\rho}-\rho}\sum_{t=1}^{T}k_t\alpha_t\alpha_{t-k_t}. \tag{108}$$

This shows the assertion when $\boldsymbol{\theta}_T^{\text{out}}=\boldsymbol{\theta}_\tau$. If $\boldsymbol{\theta}_T^{\text{out}}\in\arg\min_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_T\}}\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta})\|^2$, the assertion follows from (108) and Lemma B.5 in Appendix B.

For the second part of (i), we plug in the value of α_t and $k_t=O(\log t)$, $\Delta_{[t-k_t,t]}=O(\lambda^{k_t})$ for $\lambda\in(0,1)$ under the exponential mixing assumption.

Next, we show **(ii)**. We will first show that $\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|\to 0$ almost surely as $t\to\infty$. Under the hypothesis, by (108), we have

$$\sum_{t=1}^{\infty}\alpha_t\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right]<\infty. \tag{109}$$

By Fubini's theorem, this implies

$$\sum_{t=1}^{\infty}\alpha_t\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2<\infty\quad\text{almost surely.} \tag{110}$$

We will then use Lemma B.5 **(ii)** to deduce that $\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|\to 0$ almost surely as $t\to\infty$. To this end, it suffices to verify

$$\left|\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\|^2-\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right|=O(\alpha_t). \tag{111}$$

Indeed, by using (6) and Lemma B.3 in Appendix B,

$$\frac{1}{\hat{\rho}}\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})-\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|\le\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|+\|\text{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta}_{t+1})-\text{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta}_t)\| \tag{112}$$

$$\le\frac{2\hat{\rho}-\rho}{\hat{\rho}-\rho}\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\| \tag{113}$$

$$=\frac{2\hat{\rho}-\rho}{\hat{\rho}-\rho}\|\text{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t-\alpha_tG(\boldsymbol{\theta}_t,\mathbf{x}_{t+1}))-\text{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t)\| \tag{114}$$

$$\le\alpha_t\frac{2\hat{\rho}-\rho}{\hat{\rho}-\rho}L, \tag{115}$$

where the last inequality uses Assumption 2.3. This estimate and Lemma B.4 imply

$$\left|\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\|^2-\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right| \tag{116}$$

$$\le\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})-\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|\left(\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\|+\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|\right) \tag{117}$$

$$\le\alpha_t\frac{2\hat{\rho}-\rho}{\hat{\rho}-\rho}\frac{4L^2}{\hat{\rho}-\rho}. \tag{118}$$

Hence (111) follows, as desired.

Finally, assume f is continuously differentiable. Choose a subsequence t_k such that $\hat{\boldsymbol{\theta}}_t$ converges to some limit point $\hat{\boldsymbol{\theta}}_\infty$. We will argue that $\boldsymbol{\theta}_t\to\hat{\boldsymbol{\theta}}_\infty$ almost surely as $t\to\infty$ and $\hat{\boldsymbol{\theta}}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$. By (7) and the first part of **(ii)**, it holds that $\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|+\text{dist}(\mathbf{0},\partial\varphi(\hat{\boldsymbol{\theta}}_t))\to 0$ almost surely as $t\to\infty$. By triangle inequality $\|\hat{\boldsymbol{\theta}}_\infty-\boldsymbol{\theta}_t\|\le\|\hat{\boldsymbol{\theta}}_\infty-\hat{\boldsymbol{\theta}}_t\|+\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|$, this implies $\hat{\boldsymbol{\theta}}_t\to\hat{\boldsymbol{\theta}}_\infty$.

Next, fix arbitrary $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \{\hat{\boldsymbol{\theta}}_\infty\}$. Since $\hat{\boldsymbol{\theta}}_t \to \hat{\boldsymbol{\theta}}_\infty \neq \boldsymbol{\theta}$, it holds that $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}_t$ for all sufficiently large t. Note that

$$\left|\left\langle \nabla f(\hat{\boldsymbol{\theta}}_\infty),\, \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty\|}\right\rangle - \left\langle \nabla f(\hat{\boldsymbol{\theta}}_t),\, \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t\|}\right\rangle\right| \leq \|\nabla f(\hat{\boldsymbol{\theta}}_\infty) - \nabla f(\hat{\boldsymbol{\theta}}_t)\| + \left|\left\langle \nabla f(\hat{\boldsymbol{\theta}}_\infty),\, \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty\|} - \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t\|}\right\rangle\right|. \quad (119)$$

The last term tends to zero since $\hat{\boldsymbol{\theta}}_t \to \hat{\boldsymbol{\theta}}_\infty$ and the function $\boldsymbol{\theta}' \mapsto \frac{\boldsymbol{\theta}-\boldsymbol{\theta}'}{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}$ is continuous whenever $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$. Also, since ∇f is continuous and $\hat{\boldsymbol{\theta}}_t \to \hat{\boldsymbol{\theta}}_\infty$, the first term also tends to zero as $t \to \infty$. Then by using the relation (27), we get

$$\left\langle \nabla f(\hat{\boldsymbol{\theta}}_\infty),\, \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_\infty\|}\right\rangle \geq \left\langle \nabla f(\hat{\boldsymbol{\theta}}_t),\, \frac{\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t}{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t\|}\right\rangle - o(1) \geq -\mathrm{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}}_t)) - o(1) \quad (120)$$

for all sufficiently large $t \geq 1$. by the first part of **(ii)** and (7), we have $\mathrm{dist}(\mathbf{0}, \partial\varphi(\hat{\boldsymbol{\theta}}_t)) \to 0$ as $t \to \infty$. But since the left hand side does not depend on t, it implies that the left hand side above is nonnegative. As $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \{\hat{\boldsymbol{\theta}}_\infty\}$ is arbitrary, we conclude that $\hat{\boldsymbol{\theta}}_\infty$ is a stationary point of f over $\boldsymbol{\Theta}$. □

E. Proof for Section 3.2

Theorem E.1 (Theorem 3.3 in the main text)**.** *Let Assumption 2.1-2.3 and Assumption 3.2 hold and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 2. Fix $\hat{\rho} > \rho$ and a nondecreasing, diverging sequence $(k_t)_{t\geq 1}$. Then, for each $T \geq 1$,*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\mathrm{out}})\|^2\right] \leq \frac{\hat{\rho}^2 L}{T(\hat{\rho}-\rho)}\left(\frac{C_\varphi\sqrt{v_0+TL^2}}{\alpha\hat{\rho}L} + \sqrt{T}\right. \quad (121)$$

$$\left. + \frac{2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\left(\sqrt{T}k_T + \frac{\sqrt{T}k_T\alpha^2}{2}\log(1+v_0^{-1}TL^2)\right) + \frac{2L}{\hat{\rho}-\rho}\sum_{t=1}^{T}\mathbb{E}[\Delta_{[t-k_t,t+1]}]\right) \quad (122)$$

$$= O\left(\frac{k_T\log(TL^2)}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\Delta_{[t-k_t,t+1]}]\right). \quad (123)$$

Proof of Theorem 3.3. We proceed as the proof of Thm. 3.1, but with the difference that α_t is random and depends on the history of observed stochastic gradients, with $G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})$ being the last stochastic gradient that α_t depends on.

We estimate as in the first chain of inequalities in the proof of Thm. 3.1 with α_t dividing both sides and by omitting the expectation because of the randomness of α_t. In particular, we have

$$\frac{1}{\alpha_t}\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1}) \leq \frac{1}{\alpha_t}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_{t+1}-\hat{\boldsymbol{\theta}}_t\|^2\right] \quad (124)$$

$$= \frac{1}{\alpha_t}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\left\|\mathrm{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})) - \mathrm{proj}_{\boldsymbol{\Theta}}(\hat{\boldsymbol{\theta}}_t)\right\|^2\right] \quad (125)$$

$$\leq \frac{1}{\alpha_t}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\left\|(\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t) - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\right\|^2\right] \quad (126)$$

$$\leq \frac{1}{\alpha_t}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2\right] + \hat{\rho}\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle + \frac{\alpha_t\hat{\rho}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2}{2} \quad (127)$$

$$= \frac{1}{\alpha_t}\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t) + \hat{\rho}\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle + \frac{\alpha_t\hat{\rho}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2}{2}. \quad (128)$$

Proceeding as in the proof of Theorem 3.1, namely, by taking expectation conditional on $\mathcal{F}_{t-k}$, using Lemma 2.4, using (99),

and then integrating $\mathcal{F}_{t-k}$ out, we obtain

$$\mathbb{E}\left[\frac{1}{\alpha_t}\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\right] - \mathbb{E}\left[\frac{1}{\alpha_t}\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\right] \leq \hat{\rho}\mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) - f(\boldsymbol{\theta}_t) + \frac{\rho}{2}\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\|^2\right] + \mathbb{E}\left[\frac{\alpha_t\hat{\rho}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2}{2}\right] + \hat{\rho}\mathbb{E}\left[\frac{4L^2}{\hat{\rho}-\rho}\Delta_{[t-k,t]} + \frac{2L(L_1+\hat{\rho})}{\hat{\rho}-\rho}\sum_{s=t-k}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\|.\right]. \quad (129)$$

The only difference from before is that while bounding $\|\boldsymbol{\theta}_s - \boldsymbol{\theta}_{s-1}\|$ in Lem. 2.4 we did not use the worst case bound for $\|G(\boldsymbol{\theta}_s, x_{s+1})\|$ as in Assumption 2.1.

We use (106) on this inequality with $k = k_t$ where k_t is nondecreasing, sum for $t \in \{1, 2, \ldots, T\}$ and rearrange to get

$$\frac{\hat{\rho}-\rho}{\hat{\rho}}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right] \leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t) - \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})}{\alpha_t}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\frac{\alpha_t\hat{\rho}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2}{2}\right] + \sum_{t=1}^{T}\hat{\rho}\mathbb{E}\left[\frac{4L^2}{\hat{\rho}-\rho}\Delta_{[t-k_t,t]} + \frac{2L(L_1+\hat{\rho})}{\hat{\rho}-\rho}\sum_{s=t-k_t}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\|\right]. \quad (130)$$

We continue to upper bound the terms on the RHS of this inequality. We use Lem. B.6 to bound

$$\sum_{t=1}^{T}\alpha_t\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 = \sum_{t=1}^{T}\frac{\alpha}{\sqrt{v_0 + \sum_{j=1}^{t}\|G(\boldsymbol{\theta}_j, \mathbf{x}_{j+1})\|^2}}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 \leq 2\sqrt{\sum_{t=1}^{T}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2}, \quad (131)$$

where we also used that $v_0 > 0$. By taking expectation, and using Jensen's inequality, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2\right] \leq \mathbb{E}\left[2\sqrt{\sum_{t=1}^{T}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2}\right] \leq 2\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2\right]} \leq 2\sqrt{T}L. \quad (132)$$

We next use Assumption 3.2 to obtain

$$\sum_{t=1}^{T}\left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right)\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t) \leq \sum_{t=1}^{T}\left|\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right||\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)| \leq C_\varphi\sum_{t=1}^{T}\left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}\right) \leq C_\varphi\frac{\sqrt{v_0 + TL^2}}{\alpha}. \quad (133)$$

since $\frac{1}{\alpha_t} = \frac{\sqrt{v_0 + \sum_{j=1}^{t}\|G(\boldsymbol{\theta}_j, \mathbf{x}_{j+1})\|^2}}{\alpha}$ is monotonically nondecreasing in t.

It remains to estimate the last term on (130) which is the main additional error term that is due to dependent data. For convenience, let us define $\alpha_s\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\| = 0$ for $s \leq 0$. Then we have

$$\sum_{t=1}^{T}\sum_{s=t-k_t}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\| \leq \sum_{t=1}^{T}\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\|,$$

where $\alpha_s = \frac{\alpha}{\sqrt{v_0 + \sum_{j=1}^{s}\|G(\boldsymbol{\theta}_j, \mathbf{x}_{j+1})\|^2}}$ and the inequality used that k_t is nondecreasing.

By Young's inequality, we can upper bound this term as

$$\sum_{t=1}^{T}\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\| = \sum_{t=1}^{T}\left(\frac{(k_T)^{1/2}}{t^{1/4}}\right)\left(\frac{t^{1/4}}{(k_T)^{1/2}}\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, x_{s+1})\|\right) \quad (134)$$

$$\leq \sum_{t=1}^{T}\frac{k_T+1}{2\sqrt{t}} + \sum_{t=1}^{T}\frac{\sqrt{t}}{2k_T}\left(\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\|\right)^2 \quad (135)$$

$$\leq \sqrt{T}k_T + \sum_{t=1}^{T}\frac{\sqrt{t}}{2k_T}\left(\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s, \mathbf{x}_{s+1})\|\right)^2. \quad (136)$$

We continue estimating the last term on RHS. Using the inequality $(\sum_{i=1}^m a_i)^2 \le m\sum_{i=1}^m a_i^2$ that follows from Cauchy-Schwarz, we get

$$\begin{aligned}\sum_{t=1}^{T}\frac{\sqrt{t}}{2k_T}\left(\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s,\mathbf{x}_{s+1})\|\right)^2 &\le \sum_{t=1}^{T}\frac{\sqrt{t}}{2}\sum_{s=t-k_T}^{t-1}\alpha_s^2\|G(\boldsymbol{\theta}_s,\mathbf{x}_{s+1})\|^2\\ &\le \frac{\sqrt{T}}{2}\sum_{t=1}^{T}\sum_{s=t-k_T}^{t-1}\alpha_s^2\|G(\boldsymbol{\theta}_s,\mathbf{x}_{s+1})\|^2\\ &= \frac{\sqrt{T}}{2}\sum_{s=1}^{k_T}\sum_{t=1}^{T-s}\alpha_t^2\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2, \end{aligned} \tag{137}$$

since for any (c_s), we have $\sum_{t=1}^{T}\sum_{s=t-k_T}^{t-1} c_s = (c_{1-k_T}+c_{2-k_T}+\cdots+c_0)+(c_{2-k_T}+c_{3-k_T}+\cdots+c_1)+\cdots+(c_{T-k_T}+c_{T-k_T+1}+\cdots+c_{T-1}) = (c_{1-k_T}+c_{2-k_T}+\cdots+c_{T-k_T})+(c_{2-k_T}+c_{3-k_T}+\cdots+c_{T-k_T+1})+\cdots+(c_0+c_1+\cdots+c_{T-1}) = \sum_{s=1}^{k_T}\sum_{t=1-s}^{T-s} c_t$. Since in our case $c_t = 0$ for $t<1$, we have also that $\sum_{s=1}^{k_T}\sum_{t=1-s}^{T-s} c_t = \sum_{s=1}^{k_T}\sum_{t=1}^{T-s} c_t$.

We now have that the rightmost summation in (137) is of the form in the first inequality in Lem. B.6. We continue from (137) by using the definition of α_t

$$\begin{aligned}\sum_{t=1}^{T}\frac{\sqrt{t}}{2k_T}\Big(\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s,\mathbf{x}_{s+1})\|\Big)^2 &\le \frac{\sqrt{T}}{2}\sum_{s=1}^{k_T}\sum_{t=1}^{T-s}\alpha_t^2\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\\ &= \frac{\sqrt{T}}{2}\sum_{s=1}^{k_T}\sum_{t=1}^{T-s}\frac{\alpha^2}{v_0+\sum_{i=1}^{t}\|G(\boldsymbol{\theta}_i,\mathbf{x}_{i+1})\|^2}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\\ &\le \frac{\sqrt{T}\alpha^2}{2}\sum_{s=1}^{k_T}\log\left(1+v_0^{-1}\sum_{t=1}^{T-s}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\right)\\ &\le \frac{\sqrt{T}k_T\alpha^2}{2}\log\left(1+v_0^{-1}\sum_{t=1}^{T}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\right),\end{aligned}$$

where the third line applies the second inequality in Lem. B.6. Using this estimation on (136) gives us

$$\sum_{t=1}^{T}\sum_{s=t-k_T}^{t-1}\alpha_s\|G(\boldsymbol{\theta}_s,x_{s+1})\| \le \sqrt{T}k_T + \frac{\sqrt{T}k_T\alpha^2}{2}\log\left(1+v_0^{-1}\sum_{t=1}^{T}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\right). \tag{138}$$

Collecting (131), (133) and (138) on (130) results in the bound

$$\begin{aligned}\frac{\hat\rho-\rho}{\hat\rho}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\varphi_{1/\hat\rho}(\boldsymbol{\theta}_t)\|^2\right] \le \frac{\sqrt{v_0+TL^2}C_\varphi}{\alpha} + \hat\rho L\sqrt{T} + \sum_{t=1}^{T}\hat\rho\mathbb{E}\left[\frac{4L^2}{\hat\rho-\rho}\Delta_{[t-k,t]}\right]\\ + 2L\hat\rho\frac{L_1+\hat\rho}{\hat\rho-\rho}\left(\sqrt{T}k_T + \frac{\sqrt{T}k_T\alpha^2}{2}\log\left(1+v_0^{-1}\sum_{t=1}^{T}\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\|^2\right)\right).\end{aligned} \tag{139}$$

We divide both sides by T to conclude. □

F. Stochastic Heavy Ball with Dependent Data

In this section, we focus on stochastic heavy ball method (Algorithm 3), a popular SGD method with momentum, which dates back to (Polyak, 1964). This method is analyzed for convex optimization in (Ghadimi et al., 2015) and for constrained and stochastic nonconvex optimization with i.i.d. data in (Mai and Johansson, 2020). Some features of our analysis simplify and relax some conditions from the analysis in (Mai and Johansson, 2020) even with i.i.d. data, see Lem. F.1 and Remark F.3 for the details.

Algorithm 3 Stochastic heavy ball (momentum SGD)

Input: Initialize $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$; $T > 0$; $(\alpha_t)_{t\geq 1}$; $\beta > 0$; $z_1 > 0$
Optionally, sample τ from $\{1, \dots, T\}$ independently of everything else where $\mathbb{P}(\tau = k) = \frac{\alpha_k}{\sum_{t=1}^T \alpha_t}$.
For $t = 1, 2, \dots, T$ **do:**
    Sample $\mathbf{x}_{t+1}$ from the conditional distribution $\pi_{t+1} = \pi_{t+1}(\cdot \,|\, \mathbf{x}_1, \dots, \mathbf{x}_t)$
    $\boldsymbol{\theta}_{t+1} \leftarrow \text{proj}_{\boldsymbol{\Theta}} \left(\boldsymbol{\theta}_t - \alpha_t z_t\right)$
    $z_{t+1} = \beta G(\boldsymbol{\theta}_{t+1}, \mathbf{x}_{t+1}) + \frac{1-\beta}{\alpha_{t+1}} \left(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\right)$
End for
Return: $\boldsymbol{\theta}_T$ (Optionally, return $\boldsymbol{\theta}_\tau$)

We start with a lemma showing a bound on the norm of the sequence (z_k). We use this lemma to simplify some of the estimations in (Mai and Johansson, 2020) that analyzed the algorithm in the i.i.d. case.

Lemma F.1. *Let (z_t) be defined as Alg. 3 and let Assumption 2.3 hold. Then, we have*

$$\|z_{t+1}\|^2 \leq \beta L + (1-\beta)(\alpha_t/\alpha_{t+1})^2 \|z_t\|^2 \quad \textit{for all } t \geq 1 \tag{140}$$

and

$$\sum_{t=1}^T \beta \alpha_t^2 \|z_t\|^2 \leq \alpha_1^2 \|z_1\|^2 + \beta L^2 \sum_{t=1}^T \alpha_{t+1}^2.$$

Proof. By the definition of z_t and convexity of $\|\cdot\|^2$, we have

$$\|z_{t+1}\|^2 \leq \beta \|G(\boldsymbol{\theta}_{t+1}, \mathbf{x}_{t+1})\|^2 + \frac{1-\beta}{\alpha_{t+1}^2} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|^2 \tag{141}$$

$$\leq \beta \|G(\boldsymbol{\theta}_{t+1}, \mathbf{x}_{t+1})\|^2 + \frac{(1-\beta)\alpha_t^2}{\alpha_{t+1}^2} \|z_t\|^2, \tag{142}$$

where the second inequality used that $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}$ and that $\text{proj}_{\boldsymbol{\Theta}}$ is nonexpansive. Using Assumption 2.3 and dividing both sides by α_{t+1}^2 gives the first inequality in the assertion. Also, by multiplying both sides of the inequality by α_{t+1}^2, we have

$$\alpha_{t+1}^2 \|z_{t+1}\|^2 \leq \beta \alpha_{t+1}^2 \|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 + (1-\beta)\alpha_t^2 \|z_t\|^2. \tag{143}$$

By using Assumption 2.3 in (143), then rearranging, multiplying both sides by t^δ, and summing (143) give

$$\sum_{t=1}^T \beta t^\delta \alpha_t^2 \|z_t\|^2 \leq -T^\delta \alpha_{T+1}^2 \|z_{T+1}\|^2 + \alpha_1^2 \|z_1\|^2 + \beta L^2 \sum_{t=1}^T t^\delta \alpha_{t+1}^2.$$

Removing the nonpositive term on the RHS gives the result. □

Theorem F.2 (extended version of Theorem 3.4 in the main text)**.** *Let Assumption 2.1-Assumption 2.3 hold. Let $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 3. Fix $\hat{\rho} \geq 2\rho$. Then, for any $\beta \in (0, 1]$,*

(i) *For each $T \geq 1$:*

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_T^{\text{out}})\|^2\right] \leq \frac{\hat{\rho}}{\sum_{t=1}^T \alpha_t} \Bigg(\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_0) - \inf \varphi_{1/\hat{\rho}} + \frac{(1+\beta(1-\beta))L^2}{2\beta^2} \sum_{t=1}^T \alpha_t^2$$
$$+ \frac{1-\beta}{2\beta^2} \alpha_1 \|z_1\|^2 + \frac{4L^2}{\hat{\rho}-\rho} \sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t]}] + \frac{2L^2(L_1+\hat{\rho})}{\hat{\rho}-\rho} \sum_{t=1}^T k_t \alpha_t \alpha_{t-k} \Bigg). \tag{144}$$

(ii) *(Global convergence) Further assume that $\alpha_t/\alpha_{t+1} \to 1$ as $t \to \infty$ and $\sum_{t=1}^\infty k_t \alpha_t \alpha_{t-k_t} < \infty$. Then $\|\nabla \varphi_{1/\hat{\rho}}(\hat{\boldsymbol{\theta}}_t)\| \to 0$ as $t \to \infty$ almost surely. Furthermore, $\boldsymbol{\theta}_t$ converges to the set of all stationary points of f over $\boldsymbol{\Theta}$.*

Remark F.3. Our analysis is more flexible compared to (Mai and Johansson, 2020) even when restricted to the i.i.d. case. In this case, we allow variable step sizes $\alpha_t = \frac{\gamma}{\sqrt{t}}$ whereas (Mai and Johansson, 2020) requires constant step size $\alpha_t = \alpha = \frac{\gamma}{\sqrt{T}}$. We can also use any $\beta \in (0, 1]$ whereas (Mai and Johansson, 2020) restricts to $\beta = \alpha$. This point is important since in practice β is used as a tuning parameter.

Proof. We proceed as the proof of Thm. 3.1. However, following the existing analyses for SHB (Ghadimi et al., 2015; Mai and Johansson, 2020) we use the following iterate $\bar{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t + \frac{1-\beta}{\beta}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$ and also $\hat{\boldsymbol{\theta}}_t = \operatorname{prox}_{\varphi/\hat{\rho}}(\bar{\boldsymbol{\theta}}_t)$. The useful property of $\bar{\boldsymbol{\theta}}_t$ exploited in (Mai and Johansson, 2020) with constant step sizes (and also in (Ghadimi et al., 2015) in the unconstrained setting), is that

$$\|\bar{\boldsymbol{\theta}}_{t+1} - \hat{\boldsymbol{\theta}}_t\|^2 = \left\|\boldsymbol{\theta}_{t+1} + \frac{1-\beta}{\beta}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) - \hat{\boldsymbol{\theta}}_t\right\|^2 = \frac{1}{\beta^2}\|\boldsymbol{\theta}_{t+1} - [(1-\beta)\boldsymbol{\theta}_t + \beta\hat{\boldsymbol{\theta}}_t]\|^2 \tag{145}$$

$$\leq \frac{1}{\beta^2}\|\boldsymbol{\theta}_t - \alpha_t z_t - [(1-\beta)\boldsymbol{\theta}_t + \beta\hat{\boldsymbol{\theta}}_t]\|^2 \tag{146}$$

$$= \|\bar{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_t)\|^2, \tag{147}$$

where the inequality used that $\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}, \hat{\boldsymbol{\theta}}_t$ and their convex combinations are feasible points and the projection is nonexpansive. The last step is by simple rearrangement and using the definition of z_t.

On the first chain of inequalities in Thm. 3.1, we evaluate $\varphi_{1/\hat{\rho}}$ at $\bar{\boldsymbol{\theta}}_{t+1}$ instead of $\boldsymbol{\theta}_{t+1}$ and then use the inequality in (147) to deduce

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{t+1}) \,\middle|\, \mathcal{F}_{t-k}\right] \leq \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\bar{\boldsymbol{\theta}}_{t+1} - \hat{\boldsymbol{\theta}}_t\|^2 \,\middle|\, \mathcal{F}_{t-k}\right] \tag{148}$$

$$\leq \mathbb{E}\left[f(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\bar{\boldsymbol{\theta}}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_t) - \hat{\boldsymbol{\theta}}_t\|^2 \,\middle|\, \mathcal{F}_{t-k}\right]. \tag{149}$$

We expand the square to obtain

$$\left\|\bar{\boldsymbol{\theta}}_t - \alpha_t G(\boldsymbol{\theta}_t, \mathbf{x}_t) - \hat{\boldsymbol{\theta}}_t\right\|^2 = \|\bar{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_t\|^2 - 2\alpha_t\langle\bar{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_t, G(\boldsymbol{\theta}_t, \mathbf{x}_t)\rangle + \alpha_t^2\|G(\boldsymbol{\theta}_t, \mathbf{x}_t)\|^2. \tag{150}$$

By using the last estimate on (149) and using the definition of $\varphi_{1/\hat{\rho}}$, $\bar{\boldsymbol{\theta}}_t$ along with Assumption 2.3 gives

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{t+1}) \,\middle|\, \mathcal{F}_{t-k}\right] \leq \mathbb{E}\Bigg[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_t) - \hat{\rho}\alpha_t\langle\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t, G(\boldsymbol{\theta}_t, \mathbf{x}_t)\rangle$$
$$- \frac{\hat{\rho}\alpha_t(1-\beta)}{\beta}\langle\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}, G(\boldsymbol{\theta}_t, \mathbf{x}_t)\rangle + \frac{\hat{\rho}\alpha_t^2 L^2}{2} \,\Bigg|\, \mathcal{F}_{t-k}\Bigg]. \tag{151}$$

We estimate the third term on RHS by Young's inequality, the nonexpansiveness of the projection and Assumption 2.3

$$-\frac{\hat{\rho}\alpha_t(1-\beta)}{\beta}\langle\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}, G(\boldsymbol{\theta}_t, \mathbf{x}_t)\rangle \leq \frac{\hat{\rho}(1-\beta)}{2\beta}\left(\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|^2 + \alpha_t^2\|G(\boldsymbol{\theta}_t, \mathbf{x}_t)\|^2\right) \tag{152}$$

$$\leq \frac{\hat{\rho}(1-\beta)}{2\beta}\left(\alpha_{t-1}^2\|z_{t-1}\|^2 + \alpha_t^2 L^2\right). \tag{153}$$

We insert this estimate back to (151) and use Lem. 2.4 as in the proof of Thm. 3.1 to obtain

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{t+1})\,\middle|\,\mathcal{F}_{t-k}\right] \le \mathbb{E}\Big[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{t+1}) - \hat{\rho}\alpha_t\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, G(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle + \frac{\hat{\rho}(1-\beta)}{2\beta}\alpha_{t-1}^2\|z_t\|^2 \tag{154}$$

$$+\frac{\hat{\rho}(2-\beta)\alpha_t^2L^2}{2\beta}\,\Big|\,\mathcal{F}_{t-k}\Big]$$

$$\le \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\,\middle|\,\mathcal{F}_{t-k}\right] - \hat{\rho}\alpha_t\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t,\,\mathbb{E}_{\mathbf{x}\sim\pi}\left[G(\boldsymbol{\theta}_t,\mathbf{x})\right]\rangle + \frac{\hat{\rho}(2-\beta)\alpha_t^2L^2}{2\beta}$$

$$+\hat{\rho}\alpha_t\left(\frac{2L^2}{\hat{\rho}-\rho}\,\Delta_{[t-k,t]} + k\frac{2L^2L_1+\hat{\rho}L^2}{\hat{\rho}-\rho}\alpha_{t-k}\right) \tag{155}$$

$$+\frac{\hat{\rho}(1-\beta)}{2\beta}\alpha_{t-1}^2\mathbb{E}\left[\|z_{t-1}\|^2\,\middle|\,\mathcal{F}_{t-k}\right]. \tag{156}$$

We now estimate the second term on the RHS

$$\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,x)]\rangle \ge f(\boldsymbol{\theta}_t) - f(\hat{\boldsymbol{\theta}}_t) - \frac{\rho}{2}\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 \tag{157}$$

$$= \left(f(\boldsymbol{\theta}_t)+\frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_t-\bar{\boldsymbol{\theta}}_t\|^2\right) - \left(f(\hat{\boldsymbol{\theta}}_t)+\frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t-\bar{\boldsymbol{\theta}}_t\|^2\right) - \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_t-\bar{\boldsymbol{\theta}}_t\|^2 \tag{158}$$

$$+\frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t-\bar{\boldsymbol{\theta}}_t\|^2 - \frac{\rho}{2}\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 \tag{159}$$

$$\ge \frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t-\bar{\boldsymbol{\theta}}_t\|^2 - \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_t-\bar{\boldsymbol{\theta}}_t\|^2 \ge \frac{\hat{\rho}}{2}\|\hat{\boldsymbol{\theta}}_t-\bar{\boldsymbol{\theta}}_t\|^2 - \frac{\hat{\rho}(1-\beta)^2\alpha_{t-1}^2}{2\beta^2}\|z_{t-1}\|^2. \tag{160}$$

where the first inequality is due to ρ-weak convexity of f, and the second inequality is by $\hat{\rho}-\rho$-strong convexity of $f(\cdot)+\frac{\hat{\rho}}{2}\|\cdot-\bar{\boldsymbol{\theta}}_t\|^2$ with the optimizer $\hat{\boldsymbol{\theta}}_t$ and $\hat{\rho}\ge 2\rho$. The third inequality is by nonexpansiveness of the projection and the definition of $\bar{\boldsymbol{\theta}}_t$.

We use (160) on (156), insert $k=k_t$, integrate out $\mathcal{F}_{t-k}$ and sum to get

$$\sum_{t=1}^T\hat{\rho}^2\alpha_t\mathbb{E}\left[\|\bar{\boldsymbol{\theta}}_t-\hat{\boldsymbol{\theta}}_t\|^2\right] \le -\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{T+1})\right] + \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_1) + \sum_{t=1}^T\frac{\hat{\rho}(2-\beta)\alpha_t^2L^2}{2\beta}$$

$$+\sum_{t=1}^T\frac{\hat{\rho}(1-\beta)^2\alpha_{t-1}^2}{2\beta^2}\mathbb{E}\|z_{t-1}\|^2 + \sum_{t=1}^T\hat{\rho}\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\,\mathbb{E}[\Delta_{[t-k_t,t]}] + k_t\frac{2L^2L_1+\hat{\rho}L^2}{\hat{\rho}-\rho}\alpha_{t-k_t}\right)$$

$$+\sum_{t=1}^T\frac{\hat{\rho}(1-\beta)}{2\beta}\alpha_{t-1}^2\mathbb{E}\left[\|z_{t-1}\|^2\right]. \tag{161}$$

Using Lem. F.1 for the terms involving $\|z_t\|^2$ and using $\|\nabla\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_t)\| = \hat{\rho}\|\hat{\boldsymbol{\theta}}_t-\bar{\boldsymbol{\theta}}_t\|$ finishes the proof of **(i)** after simple arrangements.

Next, we show **(ii)**. The argument for the second part is identical to that of Theorem 3.4 **(ii)**. The argument for the first part is also similar to that of Theorem 3.1 **(ii)** with a minor modification. Namely, from (161) and the hypothesis,

$$\sum_{t=1}^T\hat{\rho}^2\alpha_t\mathbb{E}\left[\|\bar{\boldsymbol{\theta}}_t-\hat{\boldsymbol{\theta}}_t\|^2\right] < \infty. \tag{162}$$

Using Fubini's theorem and (6), this implies

$$\sum_{t=1}^T\alpha_t\|\nabla\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_t)\|^2 < \infty \qquad \text{almost surely.} \tag{163}$$

Hence by Lemma B.5, it suffices to show that

$$\left|\|\nabla\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_{t+1})\|^2 - \|\nabla\varphi_{1/\hat{\rho}}(\bar{\boldsymbol{\theta}}_t)\|^2\right| = O(\alpha_t). \tag{164}$$

Proceeding as in the proof of Theorem 3.1 **(ii)**, the above follows if $\|z_t\|$ is uniformly bounded.

It remains to show that $\|z_t\|$ is uniformly bounded. For this, it suffices to show that $\|z_t\|^2 \leq 2L$ for all sufficiently large $t \geq 1$. We deduce this from Lemma F.1. If $\beta = 1$, the lemma implies $\|z_t\|^2 \leq L$ for all $t \geq 1$, so we may assume $\beta < 1$. Proceeding by an induction on t, suppose this bound holds for z_t. Then by Lemma F.1, we have

$$\|z_{t+1}\|^2 \leq \beta L + 2(1-\beta)(\alpha_t/\alpha_{t+1})^2 L. \tag{165}$$

Since $\beta < 1$ and $\alpha_t/\alpha_{t+1} \to 1$ as $t \to \infty$, there exists $t_0 > 0$ such that for all $t > t_0$, $(1-\beta)(\alpha_t/\alpha_{t+1})^2 < 1 - \beta/2$. Therefore, for all $t > t_0$,

$$\|z_{t+1}\|^2 \leq \beta L + (1-\beta/2)(2L) = 2L. \tag{166}$$

This shows the assertion. □

G. Proximal SGD with Dependent Data

In this section, we describe how our developments for stochastic gradient method extends to the proximal case, using the ideas from (Davis and Drusvyatskiy, 2019). In particular, the problem we solve in this section is

$$\boldsymbol{\theta}^* \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\varphi(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + r(\boldsymbol{\theta})\right), \quad f(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\pi}\left[\ell(\boldsymbol{\theta}, \mathbf{x})\right], \tag{167}$$

where $r\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a convex, proper, closed function. In this case, in step 1 of Algorithm 1, we use $\operatorname{prox}_{\alpha_t r}$ instead of $\operatorname{proj}_{\boldsymbol{\Theta}}$ to define $\boldsymbol{\theta}_{t+1}$.

Recall also that

$$\hat{\boldsymbol{\theta}}_t = \operatorname{prox}_{\varphi/\hat{\rho}}(\boldsymbol{\theta}_t).$$

In the projected case, when $r(\boldsymbol{\theta})$ is the indicator function of the set $\boldsymbol{\Theta}$, we had that $\hat{\boldsymbol{\theta}}_t \in \boldsymbol{\Theta}$. This was used, for example, in (93) to use nonexpansiveness for bounding $\|\boldsymbol{\theta}_{t+1} - \hat{\boldsymbol{\theta}}_t\|^2$ since $\boldsymbol{\theta}_{t+1} = \operatorname{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t - \alpha_t g_t)$. In this case, for the same step, one needs an intermediate result derived by (Davis and Drusvyatskiy, 2019).

Lemma G.1. *(Davis and Drusvyatskiy, 2019) Given the definition of $\hat{\boldsymbol{\theta}}_t$, we have for $t \geq 0$*

$$\hat{\boldsymbol{\theta}}_t = \operatorname{prox}_{\alpha_t r}(\alpha_t\hat{\rho}\boldsymbol{\theta}_t - \alpha_t\hat{v}_t + (1-\alpha_t\hat{\rho})\hat{\boldsymbol{\theta}}_t),$$

where $\hat{v}_t \in \partial f(\hat{\boldsymbol{\theta}}_t)$.

We include the following result combining the ideas from Lem. 2.4, Thm. 3.1 and (Davis and Drusvyatskiy, 2019) for proving convergence of proximal stochastic gradient algorithm with dependent data.

Theorem G.2. *[Theorem 3.5 in the main text] Let Assumption 2.1-2.3 hold, r be convex, proper, closed and $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1 where we use $\operatorname{prox}_{\alpha_t r}$ instead of $\operatorname{proj}_{\boldsymbol{\Theta}}$ in step 1. Fix $\hat{\rho} > \rho$. For each $T \geq 1$,*

$$\mathbb{E}\left[\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] \tag{168}$$

$$\leq \frac{\hat{\rho}^2 L^2}{\hat{\rho}-\rho}\frac{1}{\sum_{k=1}^T \alpha_k}\Big[\frac{\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_0) - \inf\varphi_{1/\hat{\rho}}}{\hat{\rho}L^2} + 2\sum_{t=1}^T \alpha_t^2 + \frac{2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\sum_{t=1}^T k_t\alpha_t\alpha_{t-k_t} \tag{169}$$

$$+ \frac{4}{\hat{\rho}-\rho}\sum_{t=1}^T \alpha_t \mathbb{E}[\Delta_{[t-k_t,t]}]\Big]. \tag{170}$$

Proof. We start the same as Thm. 3.1 and note by the definition of $\hat{\boldsymbol{\theta}}_{t+1}$

$$\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1}) \leq \varphi(\hat{\boldsymbol{\theta}}_t) + \frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_{t+1} - \hat{\boldsymbol{\theta}}_t\|^2. \tag{171}$$

We next estimate $\frac{\hat{\rho}}{2}\|\boldsymbol{\theta}_{t+1}-\hat{\boldsymbol{\theta}}_t\|^2$ similar to (Davis and Drusvyatskiy, 2019) by using 1-Lipschitzness of $\operatorname{prox}_{\alpha_t g}$. Let $\delta = 1-\alpha_t\hat{\rho}$ and estimate

$$\|\boldsymbol{\theta}_{t+1}-\hat{\boldsymbol{\theta}}_t\|^2 = \|\operatorname{prox}_{\alpha_t r}(\boldsymbol{\theta}_t-\alpha_t g_t) - \operatorname{prox}_{\alpha_t r}(\alpha_t\hat{\rho}\boldsymbol{\theta}_t - \alpha_t\hat{v}_t + \delta\hat{\boldsymbol{\theta}}_t)\|^2 \tag{172}$$

$$\leq \delta^2\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 - 2\delta\alpha_t\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, G(\boldsymbol{\theta}_t,\mathbf{x}_t)-\hat{v}_t\rangle + \alpha_t^2\|G(\boldsymbol{\theta}_t,\mathbf{x}_t)-\hat{v}_t\|^2, \tag{173}$$

where we skipped some intermediate steps, which are already in (Davis and Drusvyatskiy, 2019). We note that by Lem. 2.4, we have

$$-2\delta\alpha_t\mathbb{E}\left[\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, G(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle\,\middle|\,\mathcal{F}_{t-k}\right] = -2\delta\alpha_t\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}_t)]\rangle + \\ + 2\delta\alpha_t\left(\frac{2L^2}{\hat{\rho}-\rho}\Delta_{[t-k,t]} + k\frac{2L^2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\alpha_{t-k}\right). \tag{174}$$

We take the conditional expectation of (171) and use (173) with (174) to derive

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\,\middle|\,\mathcal{F}_{t-k}\right] \leq \mathbb{E}\left[\varphi(\hat{\boldsymbol{\theta}}_t)\,\middle|\,\mathcal{F}_{t-k}\right] + \delta^2\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 \\ - 2\delta\alpha_t\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}_t)]\rangle - 2\delta\alpha_t\mathbb{E}\left[\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, -\hat{v}_t\rangle\,\middle|\,\mathcal{F}_{t-k}\right] \\ + \alpha_t^2\mathbb{E}\left[\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-\hat{v}_t\|^2\,\middle|\,\mathcal{F}_{t-k}\right] + 2\delta\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\Delta_{[t-k,t]} + k\frac{2L^2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\alpha_{t-k}\right) \tag{175}$$

We integrate out $\mathcal{F}_{t-k}$ to obtain

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\right] \leq \mathbb{E}\left[\varphi(\hat{\boldsymbol{\theta}}_t)\right] + \delta^2\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 - 2\delta\alpha_t\mathbb{E}\left[\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}_t)]-\hat{v}_t\rangle\right] \\ + \alpha_t^2\mathbb{E}\left[\|G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-\hat{v}_t\|^2\right] + 2\delta\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\mathbb{E}[\Delta_{[t-k,t]}] + 4k\frac{2L^2L_1+\hat{\rho}L^2}{\hat{\rho}-\rho}\alpha_{t-k}\right). \tag{176}$$

Next, we use that the subdifferential of ρ-weakly convex g is ρ-hypomonotone (see (Davis and Drusvyatskiy, 2019)) and $\mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x})] \in \partial f(\boldsymbol{\theta}_t)$ and $\hat{v}_t \in \partial f(\hat{\boldsymbol{\theta}}_t)$ to derive

$$\langle\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t, \mathbb{E}_{\mathbf{x}\sim\pi}[G(\boldsymbol{\theta}_t,\mathbf{x}_t)]-\hat{v}_t\rangle \geq -\rho\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2. \tag{177}$$

We combine (177) with $\|\hat{v}_t\|^2 \leq L^2$ (see (Davis and Drusvyatskiy, 2019)) on (176) to derive

$$\mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\right] \leq \mathbb{E}\left[\varphi_{1/\varphi}(\boldsymbol{\theta}_t)\right] - \hat{\rho}(\hat{\rho}-\rho)\alpha_t\mathbb{E}\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2 + 4\alpha_t^2L^2 \tag{178}$$

$$+ 2\delta\alpha_t\left(\frac{4L^2}{\hat{\rho}-\rho}\mathbb{E}[\Delta_{[t-k,t]}] + k\frac{2L^2(L_1+\hat{\rho})}{\hat{\rho}-\rho}\alpha_{t-k}\right). \tag{179}$$

We sum the inequality and argue similarly as in the proof of Theorem 3.1 to finish the proof. □

H. Proofs for Section 3.6

Lemma H.1. *Let Assumptions 2.1, 2.2, 2.3 hold, $\boldsymbol{\Theta}$ be compact, and $\Delta_{[t-k_t,t]} = O(\lambda^{k_t})$ for $\lambda < 1$. Let an algorithm output $\boldsymbol{\theta}_t$ (for example, a randomly selected iterate) such that $\mathbb{E}\|\boldsymbol{\theta}_t - \operatorname{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t - \nabla f(\boldsymbol{\theta}_t))\| \leq \varepsilon$ with $\tilde{O}(\varepsilon^{-4})$ queries to $\nabla\ell(\boldsymbol{\theta},\mathbf{x})$. Then, for $\breve{\boldsymbol{\theta}}_{t+1} = \operatorname{proj}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}_t - \tilde{\nabla}f(\boldsymbol{\theta}_t))\right)$ with $\tilde{\nabla}f(\boldsymbol{\theta}_t) = \frac{1}{\hat{N}}\sum_{i=1}^{\hat{N}}\nabla\ell(\boldsymbol{\theta}_t,\mathbf{x}^{(i)})$ with $\hat{N} = O(\varepsilon^{-2})$ samples, we have that*

$$\mathbb{E}\left[\operatorname{dist}(\mathbf{0}, \partial(f+\iota_{\boldsymbol{\Theta}})(\breve{\boldsymbol{\theta}}_{t+1}))\right] \leq \varepsilon \quad \textit{with} \quad \tilde{O}(\varepsilon^{-4}) \quad \textit{samples.}$$

Proof of Lemma H.1. By the definition of $\breve{\boldsymbol{\theta}}_{t+1}$, we have that

$$\boldsymbol{\theta}_t - \tilde{\nabla}f(\boldsymbol{\theta}_t) - \breve{\boldsymbol{\theta}}_{t+1} \in \partial\iota_{\boldsymbol{\Theta}}(\breve{\boldsymbol{\theta}}_{t+1}).$$

As a result, we have

$$\mathbb{E}\left[\operatorname{dist}(\mathbf{0}, \partial(f+\iota_{\boldsymbol{\Theta}})(\breve{\boldsymbol{\theta}}_{t+1}))\right] = \mathbb{E}\left[\min_{v\in\partial\iota_{\boldsymbol{\Theta}}(\breve{\boldsymbol{\theta}}_{t+1})} \|\nabla f(\breve{\boldsymbol{\theta}}_{t+1}) + v\|\right] \\ \leq \mathbb{E}\|\nabla f(\breve{\boldsymbol{\theta}}_{t+1}) - \breve{\boldsymbol{\theta}}_{t+1} + \boldsymbol{\theta}_t - \tilde{\nabla} f(\boldsymbol{\theta}_t)\|.$$

For convenience, let $\tilde{\boldsymbol{\theta}}_{t+1} = \operatorname{proj}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}_t - \nabla f(\boldsymbol{\theta}_t))$. We continue estimating the last inequality by using this definition, triangle inequality, nonexpansiveness of $\operatorname{proj}_{\boldsymbol{\Theta}}$, and ρ-smoothness of f

$$\begin{aligned}
&\mathbb{E}\Big[\operatorname{dist}(\mathbf{0}, \partial(f+\iota_{\boldsymbol{\Theta}})(\breve{\boldsymbol{\theta}}_{t+1}))\Big] \\
&\qquad \leq \mathbb{E}\left[\|\boldsymbol{\theta}_t - \breve{\boldsymbol{\theta}}_{t+1}\| + \|\nabla f(\breve{\boldsymbol{\theta}}_{t+1}) - \nabla f(\boldsymbol{\theta}_t)\| + \|\tilde{\nabla} f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|\right] \\
&\qquad \leq \mathbb{E}\left[(1+\rho)\|\boldsymbol{\theta}_t - \breve{\boldsymbol{\theta}}_{t+1}\| + \|\tilde{\nabla} f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|\right] \\
&\qquad \leq \mathbb{E}\left[(1+\rho)\left(\|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_{t+1}\| + \|\tilde{\boldsymbol{\theta}}_{t+1} - \breve{\boldsymbol{\theta}}_{t+1}\|\right) + \|\tilde{\nabla} f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|\right] \\
&\qquad \leq \mathbb{E}\left[(1+\rho)\|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_{t+1}\| + (2+\rho)\|\tilde{\nabla} f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|\right].
\end{aligned}$$

By the assumption in the lemma, recall that we have $\|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_{t+1}\| \leq \varepsilon$, therefore we have to estimate the last term in the last inequality. We use Lem. 7.1 in (Lyu, 2022) (see also Lemma B.7) with $\psi = \nabla\ell$ to get $\mathbb{E}\|\tilde{\nabla} f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\| = O(\hat{N}^{-1/2})$ with $\hat{N}$ samples and finish the proof. □

Proof of Theorem 3.9. When g is smooth, we can use the results in Sec. 2.2 in (Davis and Drusvyatskiy, 2019) to show that for any $\boldsymbol{\theta}$,

$$\|\mathcal{G}_{1/2\hat{\rho}}(\boldsymbol{\theta})\| \leq \frac{3}{2}\|\nabla\varphi_{1/\hat{\rho}}(\boldsymbol{\theta})\|.$$

This establishes that the upper bound of Thm. 3.1 also upper bounds the norm of the gradient mapping $\|\mathcal{G}_{1/2\hat{\rho}}(\boldsymbol{\theta})\|$. By invoking Thm. 3.1 with a randomly selected iterate, this establishes the bound required for Lem. H.1 and then applying Lem. H.1 gives the result. □

I. Proof and discussions for Section 4

Proof of Corollary L.2. Follows immediately from Theorems 3.1, 3.3, 3.4 and 3.9. For the last statement for squared Frobenius loss, see (Mairal et al., 2010) for verifying Assumption 4.1 and Assumption L.1 and recall that Assumption L.1 implies Assumption 2.1. □

J. Details about the experimental setup

For our experimental setup, we implemented the SGD based algorithms we have in this paper. The implementation of SMM uses one step of dictionary learning update given in (Mairal et al., 2010) with the special step size therein. We did not tune SMM further since the algorithm is well-established and specialized for ODL tasks, since the work of (Mairal et al., 2010).

For projected SGD and projected SGD with momentum, we used a step size of the form

$$\alpha_t = \frac{c}{\sqrt{t+1}},$$

and tuned $c \in [0.01, 1]$. In (Mairal et al., 2010) and (Zhao et al., 2017), the authors noted that using a step size $\alpha_t = \frac{c_1}{c_2 t + c_3}$ and tuning c_1, c_2, c_3 for SGD seemed to work well. We did not choose this rule in order not to tune three different parameters and since, as we show with our analysis, the best complexity is attained with a scaling of $\frac{1}{\sqrt{t}}$ for the step size. Consistent with (Mairal et al., 2010; Zhao et al., 2017), we also observed further tuning with such a rule enhances the empirical performance of SGD-based methods. However we refrain from such a specialized tuning, since our goal is not to provide an

exhaustive practical benchmark, but to enhance the theoretical understanding of algorithms whose practical merit is already well-established in a wide variety of tasks (SGD, SGD with momentum and AdaGrad).

For AdaGrad, we picked the step size

$$\alpha_t = \frac{c}{\sqrt{\sum_{i=1}^{t} \|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 + \varepsilon}},$$

with $\varepsilon = 10^{-8}$ and $c \in [0.1, 1]$ is tuned.

K. Convergence of PSGD in the state-dependent case

Theorem K.1 (extended version of Theorem 3.8 in the main text)**.** *Let Assumptions 2.1, 3.6, 3.7, and 2.3 hold. Let $(\boldsymbol{\theta}_t)_{t\geq 1}$ be a sequence generated by Algorithm 1. Fix $\hat{\rho} > \rho$. Then we have for each $T \geq 1$ that*

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|^2\right] \leq \frac{\hat{\rho}^2 L^2}{\hat{\rho} - \rho} \frac{1}{\sum_{k=1}^{T} \alpha_k} \left[\frac{\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_1) - \inf \varphi_{1/\hat{\rho}}}{\hat{\rho} L^2} + \frac{1}{2}\sum_{t=1}^{T} \alpha_t^2\right] \tag{180}$$

$$+ \frac{\hat{\rho}}{\sum_{k=1}^{T} \alpha_k} \left[\frac{\alpha_1}{2}\left(\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|^2 + C_1^2\right) + \frac{\alpha_T}{2}\frac{4LC_2}{\hat{\rho} - \rho} + \sum_{t=2}^{T} \frac{2L^2 C_3 \alpha_t \alpha_{t-1}}{\hat{\rho} - \rho}\right. \tag{181}$$

$$\left. + \sum_{t=2}^{T} C_2 \alpha_t \left(\alpha_{t-1} L + \frac{\hat{\rho}}{\hat{\rho} - \rho}\alpha_{t-1} L\right) + \sum_{t=2}^{T} |\alpha_{t-1} - \alpha_t| \frac{2LC_2}{\hat{\rho} - \rho}\right]. \tag{182}$$

In particular, with $\alpha_t = \frac{c}{\sqrt{t}}$ for some $c > 0$, we have that

$$\mathbb{E}\left[\|\nabla \varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_T^{\text{out}})\|\right] \leq \varepsilon \text{ with } \tilde{O}\left(\varepsilon^{-4}\right) \text{ samples.}$$

Remark K.2. Even though our main focus is operating under Assumption 2.1 which is the main assumption on the data used in most of the other works we compare with (Lyu, 2022), we also give this theorem for completeness. This theorem operates under another assumption depending on the solution of Poisson equation and is used in (Karimi et al., 2019; Tadić and Doucet, 2017). By using these techniques, we show that we can extend the guarantees in these papers to the constrained case. One difference is that in the constrained case, we need a slightly stronger assumption on the norms of the gradients, see Assumption 2.3.

Proof. We will follow the proof of Theorem 3.1 until (95) which is where the main error term due to non-i.d.d. data appears. We rewrite this inequality for convenience, after taking total expectation and summing the inequality for $t \geq 1$

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_{t+1})\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right] + \sum_{t=1}^{T} \hat{\rho}\alpha_t \mathbb{E}\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\rangle + \sum_{t=1}^{T} \frac{\alpha_t^2 \hat{\rho}}{2}\mathbb{E}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2 \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\varphi_{1/\hat{\rho}}(\boldsymbol{\theta}_t)\|^2\right] + \sum_{t=1}^{T} \hat{\rho}\alpha_t \mathbb{E}\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1}) - \nabla f(\boldsymbol{\theta}_t)\rangle \\
&\quad + \sum_{t=1}^{T} \hat{\rho}\alpha_t \mathbb{E}\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t,\, \nabla f(\boldsymbol{\theta}_t)\rangle + \sum_{t=1}^{T} \frac{\alpha_t^2 \hat{\rho}}{2}\mathbb{E}\|G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\|^2
\end{aligned} \tag{183}$$

We have to then bound for the second term on the right-hand side:

$$\left|\mathbb{E}\hat{\rho}\sum_{t=1}^{T} \alpha_t \langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t, \nabla f(\boldsymbol{\theta}_t) - G(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\rangle\right|. \tag{184}$$

We can then simply follow the same strategy as (Karimi et al., 2019) to obtain the result. For clarity, we write down these steps explicitly in the rest of this proof.

In particular, by (18), we have

$$\hat{\rho}\sum_{t=1}^{T} \alpha_t \langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t, \nabla f(\boldsymbol{\theta}_t) - G(\boldsymbol{\theta}_t, x_{t+1})\rangle = -\hat{\rho}\sum_{t=1}^{T} \alpha_t \langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t, \hat{G}(\boldsymbol{\theta}_t, \mathbf{x}_{t+1}) - P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t, \mathbf{x}_{t+1})\rangle.$$

Separating the inner product on the right-hand side to two parts and shifting indices give us

$$-\sum_{t=1}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle=-\alpha_1\langle\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1,\hat{G}(\boldsymbol{\theta}_1,\mathbf{x}_2)\rangle-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle$$
$$-\alpha_T\langle\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T,-P_{\boldsymbol{\theta}_T}\hat{G}(\boldsymbol{\theta}_T,\mathbf{x}_{T+1})\rangle-\sum_{t=2}^{T}\alpha_{t-1}\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1},-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle.$$

To bound the two sums in the right-hand side, we add and subtract $\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle$ to get

$$\begin{aligned}
&-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle-\sum_{t=2}^{T}\alpha_{t-1}\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1},-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle\\
&\quad=-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle\\
&\qquad-\sum_{t=2}^{T}\alpha_{t-1}\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1},-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle\\
&\quad=-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle-\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle\\
&-\sum_{t=2}^{T}\alpha_t\langle(\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1})-(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t),-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle-\sum_{t=2}^{T}(\alpha_{t-1}-\alpha_t)\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1},-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle.
\end{aligned}$$

Plugging back to (184), we get

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\nabla f(\boldsymbol{\theta}_t)-G(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\rangle\le\mathbb{E}\left[-\alpha_1\langle\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1,\hat{G}(\boldsymbol{\theta}_1,\mathbf{x}_2)\rangle-\alpha_T\langle\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T,-P_{\boldsymbol{\theta}_T}\hat{G}(\boldsymbol{\theta}_T,\mathbf{x}_{T+1})\rangle\right]\\
-\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle\\
-\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle\\
-\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle(\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1})-(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t),-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle\\
-\mathbb{E}\sum_{t=2}^{T}(\alpha_{t-1}-\alpha_t)\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1},-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle.
\end{aligned}\tag{185}$$

We bound the right-hand side in order. First

$$\mathbb{E}\left[-\alpha_1\langle\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1,\hat{G}(\boldsymbol{\theta}_1,\mathbf{x}_2)\rangle-\alpha_T\langle\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T,-P_{\boldsymbol{\theta}_T}\hat{G}(\boldsymbol{\theta}_T,\mathbf{x}_{T+1})\rangle\right]\le\frac{\alpha_1}{2}\left(\|\boldsymbol{\theta}_1-\hat{\boldsymbol{\theta}}_1\|^2+C_1^2\right)+\frac{2\alpha_T LC_2}{\hat{\rho}-\rho},$$

by Assumption 2.3, Assumption 3.7 and Lemma B.4.

Second, we use the tower rule and $\mathcal{F}_t$ measurability of $\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t$ where $\mathcal{F}_t:=\sigma(X_0,\boldsymbol{\theta}_0,X_1,\boldsymbol{\theta}_1,\ldots,X_t,\boldsymbol{\theta}_t)$, with Assumption 3.6 (used with $H(X_t)=G(\boldsymbol{\theta}_t,\mathbf{x}_t)$) to get

$$\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle=\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t,\mathbb{E}[\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_{t+1})\mid\mathcal{F}_t]-P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)\rangle=0.$$

Third,

$$-\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t, P_{\boldsymbol{\theta}_t}\hat{G}(\boldsymbol{\theta}_t,\mathbf{x}_t)-P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle \le \mathbb{E}\sum_{t=2}^{T}C_3\alpha_t\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|\|\boldsymbol{\theta}_t-\boldsymbol{\theta}_{t-1}\| \le \mathbb{E}\sum_{t=2}^{T}\frac{2L^2C_3\alpha_t\alpha_{t-1}}{\hat{\rho}-\rho},$$

where the first step used Assumption 3.7 and the last step used the definition of $\boldsymbol{\theta}_t$, nonexpansiveness of projection, Assumption 2.3 and Lemma B.4.

Fourth, we have

$$-\mathbb{E}\sum_{t=2}^{T}\alpha_t\langle(\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1})-(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t), -P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle \le \mathbb{E}\sum_{t=2}^{T}C_2\alpha_t\left(\|\boldsymbol{\theta}_t-\boldsymbol{\theta}_{t-1}\|+\|\hat{\boldsymbol{\theta}}_t-\hat{\boldsymbol{\theta}}_{t-1}\|\right) \le \sum_{t=2}^{T}C_2\alpha_t\left(\alpha_{t-1}L+\frac{\hat{\rho}}{\hat{\rho}-\rho}\alpha_{t-1}L\right).$$

where the first step used Assumption 3.7, and triangle inequality, and the last step used the definition of $\boldsymbol{\theta}_t$, nonexpansiveness of projection, Assumption 2.3 and Lemma B.3.

Fifth, by using Lemma B.4 and Assumption 3.7, we have

$$-\mathbb{E}\sum_{t=2}^{T}(\alpha_{t-1}-\alpha_t)\langle\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta}_{t-1}, -P_{\boldsymbol{\theta}_{t-1}}\hat{G}(\boldsymbol{\theta}_{t-1},\mathbf{x}_t)\rangle \le \sum_{t=2}^{T}|\alpha_{t-1}-\alpha_t|\frac{2LC_2}{\hat{\rho}-\rho}.$$

Plugging these five estimations to (185) bounds the error term in (184). Then plugging this to (183), we finish the proof after following the same steps as Theorem 3.1. □

L. Convergence of Online Dictionary Learning with first-order methods

Assumption L.1. For each $\mathbf{X}$ and $\boldsymbol{\theta}$, the function $\boldsymbol{\theta}\mapsto\ell(\mathbf{X},\boldsymbol{\theta})=\inf_{H\in\boldsymbol{\Theta}'}\left(d(\mathbf{X},\boldsymbol{\theta}H)+R(H)\right)$ is L-smooth for some $L>0$.

In (Mairal et al., 2010), it was shown that both Assumption 4.1 and Assumption L.1 are verified when d satisfies

$$d(\mathbf{X},\boldsymbol{\theta}H)=\|\mathbf{X}-\boldsymbol{\theta}H\|_F^2+\kappa_2\|H\|_F^2+\lambda\|H\|_1, \tag{186}$$

where $\kappa_2>0$ and $\lambda\ge 0$. Then the following result is a direct consequence of our main results, Theorems 3.1, 3.9, 3.3, and 3.8.

Corollary L.2. *Consider (22) and assume Assumption 4.1. Suppose we have a sequence of data matrices $(\mathbf{X}_t)_{t\ge0}$ and let $(\boldsymbol{\theta}_t)_{t\ge1}$ be the sequence of dictionary matrices in $\boldsymbol{\Theta}\subseteq\mathbb{R}^{p\times r}$ obtained by either of the three algorithms: Projected SGD (Algorithm 1), AdaGrad (Algorithm 2), and stochastic heavy ball (Algorithm 3). Suppose*

- **(a1)** *$\boldsymbol{\Theta}$ is compact and the sequence of data matrices $(\mathbf{X}_t)_{t\ge0}$ satisfy the assumption Assumption 2.2 and has a compact support;*
- **(a2)** *For each $\mathbf{X}$, the function $\boldsymbol{\theta}\mapsto\ell(\mathbf{X},\boldsymbol{\theta})$ is ρ-smooth for some $\rho>0$ over $\boldsymbol{\Theta}$.*

Then in all cases, we sample $\hat{t}\in\{1,\dots,T\}$ and compute $\breve{\boldsymbol{\theta}}_{\hat{t}+1}$ as in Theorem 3.9 and have the complexity $\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0},\partial(f+\iota_{\boldsymbol{\Theta}})(\breve{\boldsymbol{\theta}}_{t+1})\right)\right]\le\varepsilon$ with $T=\tilde{O}(\varepsilon^{-4})$ samples. Furthermore, Projected SGD and SHB converges almost surely to the set of stationary point of the objective function for (22). In particular, the above results hold under Assumption 2.2 and when d is as in (186).