CONVERGENCE AND COMPLEXITY OF STOCHASTIC SUBGRADIENT METHODS WITH DEPENDENT DATA FOR NONCONVEX OPTIMIZATION

AHMET ALACAOGLU AND HANBAEK LYU

ABSTRACT. We show that under a general dependent data sampling scheme, the classical stochastic projected and proximal subgradient methods for weakly convex functions have worst-case rate of convergence $\tilde{O}(n^{-1/4})$ and complexity $\tilde{O}(\varepsilon^{-4})$ for achieving an $\varepsilon$-near stationary point in terms of the norm of the gradient of Moreau envelope. 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 specific case of 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 adaptive stochastic subgradient algorithm AdaGrad and stochastic subgradient 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 with optimal rate of convergence guarantee.

1. INTRODUCTION

Consider the minimization of a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ given as an expected loss function:

$$\theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}_{x \sim \pi} [\ell(\theta, x)],$$

where $\pi$ is a probability distribution on a sample space $\Omega \subseteq \mathbb{R}^d$ with a density function, $\ell : \Omega \times \Theta \rightarrow \mathbb{R}$ a per-sample loss function, and $\Theta \subseteq \mathbb{R}^p$ a closed convex set with an efficiently computable projection

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

We assume $f$ is $\rho$-weakly convex, meaning that $\theta \rightarrow f(\theta) + \frac{\rho}{2} \|\theta\|^2$ is convex for some constant $\rho > 0$. The class of weakly convex functions includes convex functions and smooth functions with Lipschitz gradients. Constrained nonconvex optimization with dependent data arise in many situations such as decentralized optimization over networked systems, where the i.i.d. sampling requires significantly more communication than the dependent sampling [JRJ07, JRJ10]. Other applications are policy evaluation in reinforcement learning where the Markovian data is naturally present since the underlying model is a Markov Decision Process [BRS18] and online nonnegative matrix factorization [LNB20].

Related work. Optimization with non-i.i.d. data is studied in the convex and nonconvex cases with gradient descent in [SSY18], with block coordinate descent in [SSXY20]. Stochastic (sub)gradient descent (SGD) is also recently considered in [WPT+21] for convex problems. In the constrained nonconvex case, the work [LNB20] showed asymptotic guarantees of stochastic majorization-minimization (SMM)-type algorithms to stationary points of the expected loss function and the recent work [Lyu22] showed nonasymptotic guarantees. More recently, [Lyu22] studied a generalized SMM-type algorithms and showed the complexity $\tilde{O}(\varepsilon^{-8})$ in the general case and $\tilde{O}(\varepsilon^{-4})$ when all the iterates of the algorithm lie in the interior of the constraint set, for making the stationarity gap (see LHS of (3)) for the expected loss function less than $\varepsilon$. We also remark that [Lyu22] 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 $\varepsilon$. Our present work does not consider complexity with respect to the empirical loss functions. See [Lyu22] for more details.

To our knowledge, no convergence rate of projected SGD is known in the nonconvex case with non-i.i.d. sampling. Moreover, our result shows the optimal complexity $\tilde{O}(\varepsilon^{-4})$ for constrained problems.

arXiv:2203.15797v1 [math.OC] 29 Mar 2022
without assuming the iterates to be in the interior of the constraint set, improving over the previously known complexity $O(\varepsilon^{-8})$. We also show the first rates for AdaGrad and SGD with heavy ball momentum 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 [DL22] 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.

**Overview of main results.** In this subsection, we state a special case of one of our main results, Theorem 4.1, for a standard projected SGD algorithm with diminishing step sizes. Informally, a special instance of the algorithm proceeds by iterating the following two steps:

**Step 1.** Sample $x_{t+1}$ from a distribution conditional on $x_1, \ldots, x_t$;

**Step 2.** Compute a stochastic subgradient $G(\theta_t, x_{t+1})$ and $\theta_{t+1} \leftarrow \text{proj}_\Theta \left( \theta_t - \frac{\varepsilon}{\sqrt{t}} G(\theta_t, x_{t+1}) \right)$,

for $\varepsilon > 0$. An important point here is that we do not require the new training point $x_{t+1}$ to be distributed according to the stationary distribution $\pi$, nor to be independent on all the previous samples $x_1, \ldots, x_t$. For instance, we allow one to sample $x_{t+1}$ according to an underlying Markov chain, so that each step of sampling is computationally very efficient but the distribution $x_{t+1}$ conditional on $x_t$ could be far from $\pi$. This may induce bias in estimating the stochastic subgradient $G(\theta_t, x_{t+1})$.

Our analysis shows that the convergence of the algorithm and the order of the rate of convergence is not affected by such statistical bias in sampling training data. Informally speaking, we obtain the following result:

**Corollary 1.1** (Informal, see Alg. 1, Thm. 4.1 and 8.1). Suppose $f$ is $\rho$-smooth and $\Theta \subseteq \mathbb{R}^p$ is convex and closed. Furthermore, assume that the training samples $x_n$ are a function of some underlying exponentially mixing Markov chain. Then any convergent subsequence of $(\theta_t)_{t \geq 0}$ converges to a stationary point of $g$ over $\Theta$ almost surely. Moreover, for each $\tilde{\rho} > \rho$, the following hold for the gradient mapping:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left\| \theta_t - \text{proj}_\Theta \left( \theta_t - \frac{1}{\tilde{\rho}} \nabla f(\theta_t) \right) \right\| = \tilde{O}(t^{-1/4}).$$

Furthermore, if $\Theta$ is compact, then the algorithm produces a point $\hat{\theta}$ such that

$$\mathbb{E} \left[ \text{dist}(0, \partial(f + i_\Theta)(\hat{\theta})) \right] \leq \varepsilon \quad \text{with} \quad \tilde{O}(\varepsilon^{-4}) \quad \text{samples}.$$

Similar results hold also for adaptive gradient descent (see Alg. 2) and momentum gradient descent (see Alg. 3).

This is the same rate of convergence for i.i.d. data case obtained in [DD19] in terms of gradient mapping as shown in Thm. 8.1. This improves the rate of convergence of stochastic algorithms for constrained nonconvex expected loss minimization with dependent data [Lyu22], see Thm. 8.1 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 5.1) and SGD with heavy ball momentum (see Algorithm 3 and Theorem 6.2).

The crux of our approach is combining a more improved version of the Markovian bias analysis in [Lyu22] and the projected SGD analysis for nonconvex optimization in [DD19] using Moreau envelope. This approach lends itself to a simple and general analysis framework resulting in optimal complexity in terms of $\varepsilon$ dependence for constrained nonconvex optimization with dependent data.

### 1.1. 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 $\varepsilon$-ball centered at the origin. We also use the distance function defined as $\text{dist}(\theta, \Theta) = \min_{\theta' \in \Theta} \| \theta - \theta' \|^2$. We denote $f : \Theta \to \mathbb{R} \cup \{ +\infty \}$ to be a generic objective function, where $\Theta \subseteq \mathbb{R}^p$ is
closed and convex. Let \( \iota \) denote the indicator function of the set \( \Theta \), where \( \iota(\theta) = 0 \) if \( \theta \in \Theta \) and \( \iota(\theta) = +\infty \) if \( \theta \not\in \Theta \). Note that

\[
\min_{\theta \in \Theta} f(\theta) = \arg\min_{\theta \in \Theta} \left( \varphi(\theta) := f(\theta) + \iota_{\Theta}(\theta) \right).
\]

For each \( \theta^* \in \Theta \), the subdifferential of \( f \) at \( \theta^* \in \mathbb{R}^p \) is defined as

\[
\partial f(\theta^*) := \{ v \in \mathbb{R}^p \mid f(\theta) - f(\theta^*) \geq \langle v, \theta - \theta^* \rangle + o(\|\theta - \theta^*\|) \text{ as } \theta \rightarrow \theta^* \}.
\]

An element of \( \partial f(\theta^*) \) is called a subgradient of \( f \) at \( \theta^* \). When \( f \) is differentiable at \( \theta^* \), then \( \partial f(\theta^*) \) coincides with the singleton set containing the gradient \( \nabla f(\theta^*) \). The normal cone \( N_{\Theta}(\theta^*) \) of \( \Theta \) at \( \theta^* \) is defined as

\[
N_{\Theta}(\theta^*) := \{ u \in \mathbb{R}^p \mid \langle u, \theta - \theta^* \rangle \leq 0 \forall \theta \in \Theta \}.
\]

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

2. Background

2.1. Direct stationarity measures. In this subsection, we introduce some notions on stationarity conditions and related quantities. A first-order necessary condition for \( \theta^* \in \Theta \) to be a first order stationary point of \( f \) over \( \Theta \) is that there exists a subgradient \( v \in \partial f(\theta^*) \) such that \( -v \) belongs to the normal cone \( N_{\Theta}(\theta^*) = \partial \iota_{\Theta}(\theta^*) \), which is also equivalent to the variational inequality

\[
\inf_{\theta \in \Theta} \langle v, \theta - \theta^* \rangle \geq 0 \forall \theta \in \Theta.
\]

Hence we introduce the following notion of first-order stationarity for constrained minimization problem:

\[
\theta^* \text{ is a stationary point for } f \text{ over } \Theta \iff 0 \in v + N_{\Theta}(\theta^*) \text{ for some } v \in \partial f(\theta^*) \iff \inf_{\theta \in \Theta} \langle v, \theta - \theta^* \rangle \geq 0 \text{ for some } v \in \partial f(\theta^*).
\]

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

In view of the preceding discussion and (1), we can also say that \( \theta \) is a stationary point of \( f \) over \( \Theta \) if and only if \( 0 \in \partial \varphi(\theta) \). Accordingly, we may use \( \text{dist}(0, \partial \varphi(\theta)) = 0 \) as an equivalent notion of stationarity.

For iterative algorithms, such a first-order optimality condition may hardly be satisfied exactly in a finite number of iterations, so it is more important to know how the worst-case number of iterations

| Method                  | \( \min_{\theta \in \Theta} f(\theta) \) | \( \min_{\theta \in \Theta} f(\theta) \) | \( \min_{\theta \in \Theta} f(\theta) \) | Markovian data | Constrained |
|-------------------------|----------------------------------------|----------------------------------------|----------------------------------------|----------------|-------------|
| SMM [Lyu22]             | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-8}) \)             | -                                      | ✓              | ✓           |
| SGD [SSXY20]            | \( O(\varepsilon^{-1}) \)             | -                                      | -                                      | ✓              | x           |
| Proj. SGD [DD19]        | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-4}) \)             | x              | ✓           |
| Proj. SGD-Sec. 4        | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-4}) \)             | ✓              | ✓           |
| AdaGrad-Sec. 5          | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-1}) \)             | \( O(\varepsilon^{-4}) \)             | ✓              | ✓           |

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: \( \mathbb{E}\|\nabla f(\theta)\| \leq \varepsilon \), \( \mathbb{E}[\text{dist}(0, \partial \varphi(\theta))] \leq \varepsilon \), and \( \mathbb{E}[\|\nabla f(\lambda)\|] \leq \varepsilon \), respectively. This work showed the improved complexity \( O(\varepsilon^{-4}) \) under the additional assumption that the iterates of the algorithm are in the interior of \( \Theta \), which does not necessarily hold in the constrained case. We do not make such an assumption in this paper.
required to achieve an ‘ε-approximate’ solution scales with the desired precision ε. To this end, we can relax the above first-order optimality conditions as follow: For each ε > 0,

θ∗ is an ε-stationary point for f over Θ def dist(0, θ) ⇐⇒ dist(0, ∇ϕ(θ∗)) ≤ ε.

(2)

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

\[ \text{Gap}_θ(f, θ^*) := \inf_{v \in ∇f(θ^*)} \left\{ -\inf_{θ \in Θ} \left( \nu, \frac{θ - θ^*}{\|θ - θ^*\|} \right) \right\} = \text{dist}(0, ∇ϕ(θ^*)), \]

which is justified in Proposition A.1 in Appendix A. We call the quantity Gapθ(f, θ∗) above the stationarity gap at θ∗ for f over Θ. This measure of approximate stationarity was used in [Lyu20, Lyu22], and it is also equivalent to a similar notion in [Nes13, Def. 1]. When θ∗ is in the interior of Θ and if f is differentiable at θ∗, then (2) is equivalent to \( \|∇f(θ^*)\| \leq ε \). In Proposition A.1, we provide an equivalent definition of ε-stationarity using the normal cone.

2.2. Moreau envelope and near-stationarity measures. In this subsection, we introduce the notion of Moreau envelope and another related ‘near-stationarity’ measure. Roughly speaking, we say a point θ∗ is ‘approximately near stationary’ for f over Θ if there exists some point ˆθ near θ that is approximately stationary for f over Θ. We will make this notion precise through the following discussion.

One of the central notions in the recent breakthrough in analyzing convergence rates of first-order methods for constrained nonconvex problems [DD19] is called the ‘Moreau envelope’, which is a smooth approximation of an objective function that is closely related with proximal mapping. For a constant λ > 0, we define the Moreau envelope \( ϕ_λ \) of \( ϕ \) defined in (1) as

\[ \phi_λ(θ) := \min_{θ' \in ℝ^p} \left( \phi(θ') + \frac{1}{2λ} \|θ' - θ\|^2 \right). \]

If f is ρ-weakly convex and if λ < ρ−1, then the minimum in the right hand side is uniquely achieved at a point ˆθ, which we call the proximal point of θ. Accordingly, we define the proximal map θ → proxλϕ(θ) as

\[ ˆθ := \text{prox}_ϕ(θ) := \arg\min_{θ' \in ℝ^p} \left( \phi(θ') + \frac{1}{2λ} \|θ' - θ\|^2 \right). \]

(5)

Also in this case, the Moreau envelope \( ϕ_λ \) is C1 with gradient given by (See [DD19])

\[ \nabla ϕ_λ(θ) = λ^{-1}(θ - \text{prox}_λϕ(θ)). \]

(6)

Note that when θ is a stationary point of \( ϕ \), then its proximal point ˆθ should agree with θ. Hence the gradient norm of the Moreau envelope \( ϕ_λ \) may provide an alternative measure of stationarity. Indeed, it provides a measure of ‘near stationarity’ in the sense that if \( \|∇ϕ_λ(θ)\| \) is small, then the proximal point ˆθ in (5) is within \( λ\|∇f_λ(θ)\| \) from θ and it is approximately stationary in terms of dist(0, ∇ϕ(θ∗)) as well as the stationarity gap Gapθ(f, ˆθ):

\[ \begin{align*}
\|θ - ˆθ\| &\leq λ\|∇ϕ_λ(θ)\| \\
ϕ(ˆθ) &\leq ϕ(θ) \\
\text{Gap}_θ(f, ˆθ) &= \text{dist}(0, ∇ϕ(θ∗)) ≤ λ\|∇ϕ_λ(θ)\|.
\end{align*} \]

(7)

Note that the first and the last inequality above follows from the first-order optimality condition for ˆθ together with (6) (See Propositions A.2 and A.1 in Appendix A).

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 [DD19, DP19] 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 \cite{Nes13}

$$\|\mathcal{G}_{1/\rho}(\theta_t)\| = \hat{\rho} \left\| \theta_t - \text{proj}_\Theta \left( \theta_t - \frac{1}{\rho} \nabla f(\theta_t) \right) \right\| =: \hat{\rho} \| \theta_t - \tilde{\theta}_t \|,$$

for any $\lambda > 0$, where we have also defined the variable $\tilde{\theta}_t$. The results \cite[Sec. 2.2]{DD19} \cite[Thm. 5.3]{DL18} showed how to translate the guarantees on the gradient of the Moreau envelope to gradient mapping by proving that

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

It is easy to show that a small gradient mapping implies that $\theta_t$ is close to the point $\text{proj}_\Theta \left( \theta_t - \frac{1}{\rho} \nabla f(\theta_t) \right)$ which itself is approximately stationary in view of Sec. 2.1. Even though such an approximately stationary point can be computed in the deterministic case, computation of $\nabla f(\theta_t)$ might not be tractable in the stochastic case. We show in Sec. 8 how to output a point with our algorithms which is approximately stationary with the claimed complexity results in the dependent data setting.

### 3. STOCHASTIC GRADIENT ESTIMATION

Denote $\Delta_{n-k,n}$ to be the worst-case total variation distance between conditional distribution of $x_n$ given $x_1, \ldots, x_{n-k}$ and the stationary distribution $\pi$. Namely,

$$\Delta_{n-k,n} := \sup_{x_1, \ldots, x_{n-k} \in \Omega} \| \pi_n(\cdot | x_1, \ldots, x_{n-k}) - \pi \|_{TV},$$

where $\pi_n(x | x_1, \ldots, x_{n-k})$ denotes the probability distribution of $x_n$ conditional on the past points $x_1, \ldots, x_{n-k}$.

(A1). We can sample a sequence of points $(x_n)_{n \geq 1}$ in $\Omega$ in a way that there exists a diverging sequence $k_t \in \{1, t\}$, $t \geq 1$ such that

$$\Delta_{[t-k_t,t]} \to 0, \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_t \Delta_{[t-k_t,t+1]} < \infty.$$

In a special case, suppose $x_t$ is given by a function $g$ of some underlying Markov chain $X_t$. If $X_t$ is irreducible and aperiodic on a finite state space, then $X_t$ has a unique stationary distribution, say $\pi$, and the total variation distance between the empirical distribution of $X_t$ and $\pi$ vanishes exponentially fast. In this case, $\Delta_{n-k,n} = O(\exp(-ck))$ for some constant $c > 0$ independent of $n$, so (A1) is verified for any $\alpha_t = O(1)$ and $k_t \geq C \log t$ for $C > 0$ large enough. In 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 (A1) (see, e.g., \cite{LP17, MT12}). See \cite[Sec. 5, Sec. 6]{LNB20} for concrete applications and sampling methods that satisfy this assumption.

(A2). The function $f$ is $\rho$-weakly convex and the set $\Theta$ is closed and convex. There exists an open set $U \subset \mathbb{R}^p$ containing $\Theta$ and a mapping $G : U \times \Omega \to \mathbb{R}^p$ such that for all $\theta \in \Theta$,

$$\mathbb{E}_{x \sim \pi}[G(\theta, x)] \in \partial f(\theta).$$

Furthermore, there exists a constant $L_1 > 0$ such that the mapping $\theta \to G(\theta, x)$ is $L_1$-Lipschitz for each $x \in \Omega$.

The last assumption in (A2) is not required in \cite{DD19} for i.i.d. case, but it is required for previous works in SGD for non-i.i.d. data in the unconstrained case \cite{SSY18}.

(A3). There exists $L \in (0, \infty)$ such that $\|G(\theta, x)\| \leq L$ for all $\theta \in \Theta$ and $x \in \Omega$. 

The next result 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. This lemma uses a similar idea as [Lyu22]. Importantly, the result of the lemma makes it explicit the dependence on the step size and gradient norms to be applicable with AdaGrad.

**Lemma 3.1** (Key lemma). Let (A2), (A3) hold and $\theta_t$ be generated according to Algorithm 1, 2 or 3. Fix $\hat{\rho} > 0$ and denote $\hat{\theta} = \text{prox}_{\rho/(\hat{\rho})}(\theta)$ and fix $1 \leq k \leq t$. Then

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

**Proof.** Let $\pi_{t+1|t-k} = \mathbb{P} \mathcal{F}_{t-k}$ denote the distribution of $\mathbf{x}_{t+1}$ conditional on the information $\mathcal{F}_{t-k} = \sigma(\mathbf{x}_1, \ldots, \mathbf{x}_{t-k})$. Denote $V(\mathbf{x}, \theta) := \langle \theta - \theta_t, G(\theta, \mathbf{x}) \rangle$. Note that $\mathbb{E}_{\pi} \mathbb{E}[V(\mathbf{x}, \theta)] = \langle \hat{\theta}_t - \theta_t, \mathbb{E}_{x \sim \pi}[G(\theta_t, \mathbf{x})] \rangle$.
Observe that by triangle inequality

$$
\left| \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_t) | \mathcal{F}_{t-k} \right] - \mathbb{E}_{\pi} \left[ V(\mathbf{x}, \theta_t) \right] \right| 
\leq \left| \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_t) | \mathcal{F}_{t-k} \right] - \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_{t-k}) | \mathcal{F}_{t-k} \right] \right| 
+ \left| \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_{t-k}) | \mathcal{F}_{t-k} \right] - \mathbb{E}_{\pi} \left[ V(\mathbf{x}, \theta_{t-k}) \right] \right| 
+ \left| \mathbb{E}_{\pi} \left[ V(\mathbf{x}, \theta_{t-k}) \right] - \mathbb{E}_{\pi} \left[ V(\mathbf{x}, \theta_t) \right] \right|.
$$

(8)

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

First, by $\|\theta_s - \theta_{s-1}\| \leq \alpha_s \|G(\theta_{s-1}, \mathbf{x}_s)\|$ for $s \geq 1$, using triangle inequality, we have

$$
\left| \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_t) | \mathcal{F}_{t-k} \right] - \mathbb{E} \left[ V(\mathbf{x}_{t+1}, \theta_{t-k}) | \mathcal{F}_{t-k} \right] \right| 
\leq \frac{2LL_1 + 2\hat{\rho}L}{\hat{\rho} - \rho} \mathbb{E} \left[ \|\theta_{t-k} - \theta_t\| \right| \mathcal{F}_{t-k} \right] 
\leq \frac{2LL_1 + 2\hat{\rho}L}{\hat{\rho} - \rho} \left\| \sum_{s=t-k}^{t-1} \alpha_s \|G(\theta_s, \mathbf{x}_{s+1})\| \right\| \mathcal{F}_{t-k} \right],
$$

(9)

where the first inequality used the estimation

$$
V(\mathbf{x}_{t+1}, \theta_t) - V(\mathbf{x}_{t+1}, \theta_{t-k}) = \langle \hat{\theta}_t - \theta_t, G(\theta_t, \mathbf{x}_{t+1}) - \langle \hat{\theta}_{t-k} - \theta_{t-k}, G(\theta_{t-k}, \mathbf{x}_{t+1}) \rangle \rangle 
= \langle \hat{\theta}_t - \theta_t, G(\theta_t, \mathbf{x}_{t+1}) - G(\theta_t, \mathbf{x}_{t-k}) + \langle \hat{\theta}_t - \theta_{t-k}, G(\theta_{t-k}, \mathbf{x}_{t+1}) \rangle \rangle + \langle \theta_{t-k} - \theta_t, G(\theta_{t-k}, \mathbf{x}_{t+1}) \rangle.
$$

along with Cauchy-Schwarz inequality, $L_1$-Lipschitz continuity of $\theta \to G(\mathbf{x}, \theta)$ (see (A2)), Lemma A.4, Lemma A.3 and (A3). A similar argument shows

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

(10)

We continue by estimating the first term on the RHS of (8). Recall that by Scheffé’s lemma, if two probability measures $\mu$ and $\nu$ on the same probability space have densities $\alpha$ and $\beta$ with respect to a reference measure $dm$, then $\|\mu - \nu\|_{TV} = \frac{1}{2} \int |\alpha - \beta| d m$ (see, e.g., [Tsy04]). Let $\pi_{t+1|t-k}$ and $\pi_t$ denote the density functions of $\pi_{t+1|t-k}$ and $\pi_t$ with respect to the Lebesgue measure, which we denote by $d \xi$. Using this identity and since $\theta_t$ is deterministic with respect to $\mathcal{F}_{t-k}$, we have that

$$
\mathbb{E} \left[ G(\theta_{t-k}, \mathbf{x}_{t+1}) | \mathcal{F}_{t-k} \right] - \mathbb{E}_{\pi} \left[ G(\theta_{t-k}, \mathbf{x}) \right] 
= \mathbb{E} \left[ \int_{\Omega} G(\theta_{t-k}, \mathbf{x})(\pi_{t+1|t-k}(\mathbf{x}) - \pi_t(\mathbf{x})) d \xi \right] 
\leq \int_{\Omega} \|G(\theta_{t-k}, \mathbf{x})\| |\pi_{t+1|t-k}(\mathbf{x}) - \pi_t(\mathbf{x})| d \xi 
\leq L \|\pi_{t+1|t-k} - \pi_t\|_{TV}
$$
Hence by Cauchy-Schwarz inequality, and using that \( \theta_{t-k}, \tilde{\theta}_{t-k} \) are deterministic with respect to \( \mathcal{F}_{t-k} \), we get

\[
\leq L \Delta_{[t-k,t+1]}.
\]

4. Projected Stochastic Subgradient Method with Dependent Data

Below we give a formal statement of the standard projected stochastic subgradient descent algorithm:

**Algorithm 1** Projected Stochastic Subgradient Algorithm

1. **Input:** Initialize \( \theta_1 \in \Theta \subseteq \mathbb{R}^p; T > 0; (\alpha_t)_{t \geq 1} \)
2. Sample \( \tau \) from \( \{1, \ldots, T\} \) independently of everything else where \( p(\tau = k) = \frac{\alpha_k}{\sum_{i=1}^T \alpha_i} \).
3. For \( t = 1, 2, \ldots, T \) do:
   4. Sample \( x_{t+1} \) from the conditional distribution \( \pi_{t+1} = \pi_{t+1}(\cdot|x_1, \ldots, x_t) \)
   5. \( \theta_{t+1} \leftarrow \text{proj}_\Theta(\theta_t - \alpha_t G(\theta_t, x_{t+1})) \)
6. End for
7. **Return:** \( \theta_T \) (Optionally, \( \theta_T^{\text{out}} \) as either \( \theta_T \) or \( \text{argmin}_{\theta \in \Theta} \| \nabla \phi_1/\hat{\beta}(\theta) \|^2 \)).

Now we state our first main result in this work, which extends the convergence result of projected SGD with i.i.d. samples in [DD19] to the general dependent sample setting. This result improves the existing complexity of stochastic algorithms from [Lyu22] for solving constrained nonconvex stochastic optimization under dependent data. In Theorem 8.1 (ii), we use the notion of 'global convergence' with respect to arbitrary initialization.

**Theorem 4.1** (Projected stochastic subgradient method). Assume (A1)-(A3) hold. Let \( (\theta_t)_{t \geq 1} \) be a sequence generated by Algorithm 1. Fix \( \hat{\beta} > \beta \). Then the following hold:

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

\[
\mathbb{E} \left[ \| \nabla \phi_{1/\hat{\beta}}(\theta_T^{\text{out}}) \|^2 \right] \leq \frac{\hat{\beta}^2 L^2}{\hat{\beta} - \beta} \sum_{k=1}^T a_k \left[ \frac{\phi_{1/\hat{\beta}}(\theta_1) - \inf_{\theta} \phi_{1/\hat{\beta}}}{\hat{\beta} L^2} + \frac{1}{2} \sum_{t=1}^T a_t^2 + \frac{2(L_1 + \hat{\beta})}{\hat{\beta} - \beta} \sum_{t=1}^T k_t \alpha_t \alpha_{t-k_t} + \frac{2}{\hat{\beta} - \beta} \sum_{t=1}^T a_t \mathbb{E}[\Delta_{[t-k_t,t+1]}] \right] - \epsilon
\]

In particular, with \( \alpha_t = \frac{c}{\sqrt{t}} \) for some \( c > 0 \) and under exponential mixing (for example, with aperiodic and irreducible Markov Chain with finite state space), we have that

\[
\mathbb{E} \left[ \| \nabla \phi_{1/\hat{\beta}}(\theta_T^{\text{out}}) \|^2 \right] \leq \epsilon \quad \text{with} \quad \tilde{O}(\epsilon^{-4}) \quad \text{samples}.
\]

(ii) (Global convergence) Further assume that \( \sum_{t=0}^\infty k_t \alpha_t \alpha_{t-k_t} < \infty \). Then \( \| \nabla \phi_{1/\hat{\beta}}(\tilde{\theta}_t) \| \to 0 \) as \( t \to \infty \) almost surely. Furthermore, if \( f \) is continuously differentiable, then \( \tilde{\theta}_t \) converges to the set of all stationary points of \( \phi \) over \( \Theta \).

**Remark 4.2.** If \( (x_t)_{t \geq 1} \) is exponentially mixing, then Theorem 6.2(ii) holds with \( \alpha_t = t^{-1/2} \log t \) for any fixed \( \epsilon > 0 \) and \( k_t = O(\log t) \).
**Proof.** Recall the definition of \( \varphi_{1/\hat{\rho}} \) from (4). We start as in [DD19] with the difference of conditioning on \( \mathcal{F}_{t-k} \) instead of the latest iterate, and use Lemma 3.1 to handle the additional bias due to dependent sampling. Denote \( \hat{\theta}_t = \text{prox}_{\varphi_{1/\hat{\rho}}}(\theta_t) \) for \( t \geq 1 \) and fix \( k \in \{0, \ldots, t\} \). Observe that

\[
\mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\theta_{t+1}) \big| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ f(\hat{\theta}_t) + \frac{\hat{\rho}}{2} \| \theta_{t+1} - \hat{\theta}_t \|_2^2 \big| \mathcal{F}_{t-k} \right]
\]

\[
= \mathbb{E} \left[ f(\hat{\theta}_t) \big| \mathcal{F}_{t-k} \right] + \frac{\hat{\rho}}{2} \mathbb{E} \left[ \| \text{proj}_{\Theta} (\theta_t - \alpha_t \psi(\theta_t, x_{t+1}) \big| \mathcal{F}_{t-k} \right] \tag{12}
\]

\[
\leq \mathbb{E} \left[ f(\hat{\theta}_t) \big| \mathcal{F}_{t-k} \right] + \frac{\hat{\rho}}{2} \mathbb{E} \left[ \| \theta_t - \hat{\theta}_t - \alpha_t \psi(\theta_t, x_{t+1}) \|_2^2 \big| \mathcal{F}_{t-k} \right]
\]

\[
\leq \mathbb{E} \left[ f(\hat{\theta}_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2 \big| \mathcal{F}_{t-k} \right] + \hat{\rho} \alpha_t \mathbb{E} \left[ (\hat{\theta}_t - \theta_t, \psi(\theta_t, x_{t+1})) \big| \mathcal{F}_{t-k} \right] + \frac{\alpha_t^2 \hat{\rho} L^2}{2} \tag{13}
\]

Namely, the first and the last inequalities use the definition of Moreau envelope \( \varphi_{1/\hat{\rho}} \) and \( \hat{\theta}_t \in \Theta \), the second inequality uses 1-Lipschitzness of the projection operator. The last inequality uses Lemma 3.1 as well as the uniform bound \( \alpha_s \| \psi(\theta_s, x_{t+1}) \| \leq \alpha_s L \) (see A2) and that \( \alpha_s \) is non-increasing in \( s \).

By using (A2) and the weak convexity of \( g \), we have

\[
\langle \hat{\theta}_t - \theta_t, \mathbb{E}_{x \sim \pi} [\psi(\theta_t, x)] \rangle \leq f(\hat{\theta}_t) - f(\theta_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2. \tag{14}
\]

By using this estimate in (13) and then integrating out \( \mathcal{F}_{t-k} \), we get

\[
\mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\theta_{t+1}) \right] - \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\theta_t) \right] \leq \hat{\rho} \alpha_t \mathbb{E} \left[ f(\hat{\theta}_t) - f(\theta_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2 \right] + \frac{\alpha_t^2 \hat{\rho} L^2}{2} \tag{15}
\]

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

\[
\hat{\rho} \sum_{i=1}^{T} \alpha_t \mathbb{E} \left[ f(\theta_i) - f(\hat{\theta}_i) - \frac{\hat{\rho}}{2} \| \theta_i - \hat{\theta}_i \|_2^2 \right] \leq \left( \varphi_{1/\hat{\rho}}(\theta_1) - \inf_{\theta \in \Theta} \varphi_{1/\hat{\rho}}(\theta) \right) + \frac{\hat{\rho} L^2}{\hat{\rho} - \rho} \sum_{i=1}^{T} \alpha_t^2 + \frac{\hat{\rho} L^2}{\hat{\rho} - \rho} \sum_{i=1}^{T} \alpha_t \mathbb{E}_{[\Delta_{t-k, t+1}]} \tag{15}
\]

\[
+ \frac{2L^2(L + \hat{\rho})}{\hat{\rho} - \rho} \sum_{i=1}^{T} k_i \alpha_t \alpha_{t-k}.
\]

Next, we use the fact that the function \( \theta \mapsto f(\theta) + \frac{\hat{\rho}}{2} \| \theta - \hat{\theta}_t \|_2^2 \) is strongly convex with parameter \((\hat{\rho} - \rho)/2\) that is minimized at \( \hat{\theta}_t \) to get

\[
f(\theta_t) - f(\hat{\theta}_t) - \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2
\]

\[
= \left( f(\theta_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2 \right) - \left( f(\hat{\theta}_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|_2^2 \right) + \frac{\hat{\rho} - \rho}{2} \| \theta_t - \hat{\theta}_t \|_2^2 \geq (\hat{\rho} - \rho) \| \theta_t - \hat{\theta}_t \|_2^2
\]

\[
= \frac{\hat{\rho} - \rho}{\hat{\rho}^2} \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \|_2^2. \tag{16}
\]
where the second to the last equality uses (6). Combining with (15), this implies
\[
\frac{\hat{\rho} - \rho}{\hat{\rho}} \sum_{t=1}^{T} \alpha_t \mathbb{E} \left[ \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \|^2 \right] \leq \left( \varphi_{1/\hat{\rho}}(\theta_1) - \inf_{\theta} \varphi_{1/\hat{\rho}} \right) + \frac{\hat{\rho} L^2}{2} \sum_{t=1}^{T} \alpha_t^2 + \frac{2 \hat{\rho} L^2}{\hat{\rho} - \rho} \sum_{t=1}^{T} \alpha_t \mathbb{E} \left[ \Delta_{t-k_t, t+1} \right] 
\]
\[
+ \frac{2 L^2 \hat{\rho} (L_1 + \hat{\rho})}{\hat{\rho} - \rho} \sum_{t=1}^{T} k_t \alpha_t \alpha_{t-k_t}.
\]
This shows the assertion when \( \theta^\text{out} = \theta \). If \( \theta^\text{out} T \in \arg \min_{\theta \in [\theta_1, \ldots, \theta_T]} \| \nabla \varphi_{1/\hat{\rho}}(\theta) \|^2 \), the assertion follows from (17) and Lemma A.5 in Appendix A.

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

Next, we show (ii). We will first show that \( \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| \to 0 \) almost surely as \( t \to \infty \). Under the hypothesis, by (17), we have
\[
\sum_{t=1}^{\infty} \alpha_t \mathbb{E} \left[ \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \|^2 \right] < \infty.
\]
By Fubini’s theorem, this implies
\[
\sum_{t=1}^{\infty} \alpha_t \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \|^2 < \infty \quad \text{almost surely.}
\]
We will then use Lemma A.5 (ii) to deduce that \( \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| \to 0 \) almost surely as \( t \to \infty \). To this end, it suffices to verify
\[
\left\| \nabla \varphi_{1/\hat{\rho}}(\theta_{t+1}) \right\|^2 - \left\| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \right\|^2 = O(\alpha_t).
\]
Indeed, by using (6) and Lemma A.3 in Appendix A,
\[
\frac{1}{\hat{\rho}} \| \nabla \varphi_{1/\hat{\rho}}(\theta_{t+1}) - \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| \leq \| \theta_{t+1} - \theta_t \| + \| \text{proj}_{\hat{\rho}/\hat{\rho}}(\theta_{t+1}) - \text{proj}_{\hat{\rho}/\hat{\rho}}(\theta_t) \|
\]
\[
\leq \frac{2 \hat{\rho} - \rho}{\hat{\rho} - \rho} \| \theta_{t+1} - \theta_t \|
\]
\[
= \frac{2 \hat{\rho} - \rho}{\hat{\rho} - \rho} \| \text{proj}_\theta(\theta_t - \alpha_t g(\theta_t, x_t)) - \text{proj}_\theta(\theta_t) \|
\]
\[
\leq \alpha_t \frac{2 \hat{\rho} - \rho}{\hat{\rho} - \rho} L,
\]
where the last inequality uses (A3). This estimate and Lemma A.4 imply
\[
\left\| \nabla \varphi_{1/\hat{\rho}}(\theta_{t+1}) \right\|^2 - \left\| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \right\|^2 \leq \| \nabla \varphi_{1/\hat{\rho}}(\theta_{t+1}) - \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| \left( \| \nabla \varphi_{1/\hat{\rho}}(\theta_{t+1}) \| + \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| \right)
\]
\[
\leq \alpha_t \frac{4 \hat{\rho}^2 - \rho^2}{\hat{\rho} - \rho} L^2.
\]
Hence (18) follows, as desired.

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

Next, fix arbitrary \( \theta \in \Theta \setminus \{ \hat{\theta}_\infty \} \). Since \( \hat{\theta}_t \to \hat{\theta}_\infty \neq \theta \), it holds that \( \theta \neq \hat{\theta}_t \) for all sufficiently large \( t \). Note that
\[
\left| \left\langle \nabla f(\hat{\theta}_\infty), \frac{\theta - \hat{\theta}_\infty}{\| \theta - \hat{\theta}_\infty \|} \right\rangle - \left\langle \nabla f(\hat{\theta}_t), \frac{\theta - \hat{\theta}_t}{\| \theta - \hat{\theta}_t \|} \right\rangle \right| \leq \| \nabla f(\hat{\theta}_\infty) - \nabla f(\hat{\theta}_t) \|.
\]
The last term tends to zero since $\tilde{\theta}_t \to \tilde{\theta}_\infty$ and the function $\theta' \to \frac{\theta - \theta'}{\|\theta - \theta'\|}$ is continuous whenever $\theta' \neq \theta$. Also, since $\nabla f$ is continuous and $\tilde{\theta}_t \to \tilde{\theta}_\infty$, the first term also tends to zero as $t \to \infty$. Then by using the relation (3), we get
\[
\left\langle \nabla f(\tilde{\theta}_\infty), \frac{\theta - \tilde{\theta}_\infty}{\|\theta - \tilde{\theta}_\infty\|} \right\rangle \geq \left\langle \nabla f(\tilde{\theta}_t), \frac{\theta - \tilde{\theta}_t}{\|\theta - \tilde{\theta}_t\|} \right\rangle - o(1) \geq -\text{dist}(0, \partial \phi(\tilde{\theta}_t)) - o(1)
\]
for all sufficiently large $t \geq 1$. By the first part of (ii) and (7), we have $\text{dist}(0, \partial \phi(\tilde{\theta}_t)) \to 0$ 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 $\theta \in \Theta \setminus \{\tilde{\theta}_\infty\}$ is arbitrary, we conclude that $\tilde{\theta}_\infty$ is a stationary point of $f$ over $\Theta$. 

\section{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, which is also referred to as AdaGrad-norm [WWB19, Lev17]. For this section, we introduce an additional assumptions on the boundedness of the objective values.

(A4). There exists $C_q \in (0, \infty)$ such that $|f(\theta)| \leq C_q$ for all $\theta \in \Theta$.

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 [Lyu22]) since they critically rely on knowing the decay rate of the step sizes. To be able to apply such an analysis for adaptive algorithms, we use a generalized result in Lem. 3.1 and use the specific form of AdaGrad step size in Thm. 5.1 to achieve the optimal $O(\epsilon^{-4})$ complexity.

\begin{algorithm}
\caption{AdaGrad}
\begin{algorithmic}
\State \textbf{Input:} Initialize $\theta_1 \in \Theta \subseteq \mathbb{R}^p$; $T > 0$; $\{\alpha_t\}_{t \geq 1}$; $\nu_0 > 0$; $\alpha > 0$
\State Optionally, sample $\tau$ from $\{1, \ldots, T\}$ independently of everything else where $\mathbb{P}(\tau = k) = \frac{1}{T}$.
\For {$t = 1, 2, \ldots, T$}
\State Sample $x_{t+1}$ from the conditional distribution $\pi_{t+1} = \pi_{t+1}(\cdot | x_1, \ldots, x_t)$
\State $v_t = v_{t-1} + \|G(\theta_t, x_{t+1})\|^2$\hspace{1cm}$\alpha_t = \frac{\nu}{\sqrt{v_t}}$
\State $\theta_{t+1} \leftarrow \text{proj}_\Theta(\theta_t - \alpha_t G(\theta_t, x_{t+1}))$
\EndFor
\State \textbf{Return:} $\theta_T$ (Optionally, $\theta_T^{\text{out}}$ as either $\theta_{\tau}$ or $\arg\min_{\theta \in \Theta \setminus \{\tilde{\theta}_\infty\}} \|\nabla f_{1/\rho}(\theta)\|^2$.)
\end{algorithmic}
\end{algorithm}

\textbf{Theorem 5.1} (Adaptive gradient method-AdaGrad). Let (A1)-(A4) hold. Let $\{\theta_t\}_{t \geq 1}$ be a sequence generated by Algorithm 2. Fix $\hat{\rho} > \rho$ and $\alpha$ and a nondecreasing, diverging sequence $\{k_t\}_{t \geq 1}$. Then, for each $T \geq 1$,
\[
\mathbb{E}\left[\|\nabla f_{1/\hat{\rho}}(\theta_T^{\text{out}})\|^2\right] \leq \frac{\hat{\rho}^2 L}{T(\hat{\rho} - \rho)} \left( \frac{C_q \sqrt{\nu_0 + TL^2}}{\alpha \hat{\rho} L} + \sqrt{T} + \frac{2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \left( \sqrt{T}k_T + \frac{\sqrt{T}k_T \alpha^2}{2} \log(1 + \nu_0^{-1} TL^2) \right) \right)
\]
\[
\hspace{5cm} + \frac{2L}{\hat{\rho} - \rho} \sum_{t=1}^{T} \mathbb{E}[\Delta_{[1-k_t,t+1]}]
\]
\[
= O\left( \frac{k_T \log(TL^2)}{\sqrt{T}} + \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\Delta_{[1-k_t,t+1]}] \right).
\]

\textbf{Remark 5.2.} We note that unlike Thm. 4.1, for AdaGrad we only prove nonasymptotic complexity results and not asymptotic convergence statements for the output sequence of the algorithms. Even tough asymptotic convergence of AdaGrad with i.i.d. data is proven in [LO19], the technique in that
We continue to upper bound the terms on the RHS of this inequality. We use Lem. A.6 to bound the stochastic subgradient that depends on. However, the specific form of (19) is important in our development to use Lem. 3.1 to handle the dependent data, since \( \alpha_t \) brings additional stochastic dependencies. Even though we believe an appropriate modification of Lem. 3.1 can be possible, we do not pursue such generalization in the present work.

**Remark 5.3.** Since the step size in this case is nonincreasing, (A1) reduces to \( \sum_{t=1}^{\infty} \Delta_{[t-k,t+1]} < \infty \). This, for example, is satisfied for the exponential mixing case that is mentioned in Theorem 4.1 and considered in the previous work [Lyu22].

**Proof of Theorem 5.1.** We proceed as the proof of Thm. 4.1, but with the difference that \( \alpha_t \) is random and depends on the history of observed stochastic subgradients, with \( G(\theta_t, x_{t+1}) \) being the last stochastic subgradient that \( \alpha_t \) depends on.

We estimate as in the first chain of inequalities in the proof of Thm. 4.1 with \( \alpha_t \) dividing both sides and by omitting the expectation because of the randomness of \( \alpha_t \). In particular, we have

\[
\frac{1}{\alpha_t} \varphi_{1/\hat{p}}(\theta_{t+1}) \leq \frac{1}{\alpha_t} \left[ \frac{f(\hat{\theta}_t) + \hat{\beta} \| \theta_{t+1} - \hat{\theta}_t \|^2}{2} \right] = \frac{1}{\alpha_t} \left[ f(\hat{\theta}_t) + \hat{\beta} \| \text{proj}_\theta(\theta_t - \alpha_t G(\theta_t, x_{t+1})) - \text{proj}_\theta(\hat{\theta}_t) \|^2 \right]
\]

\[
\leq \frac{1}{\alpha_t} \left[ f(\hat{\theta}_t) + \hat{\beta} \| (\theta_t - \hat{\theta}_t) - \alpha_t G(\theta_t, x_{t+1}) \|^2 \right] \leq \frac{1}{\alpha_t} \left[ f(\hat{\theta}_t) + \hat{\beta} \| \theta_t - \hat{\theta}_t \|^2 \right] + \hat{\beta} \langle \hat{\theta}_t - \theta_t, G(\theta_t, x_{t+1}) \rangle + \frac{\alpha_t \hat{\beta} \| G(\theta_t, x_{t+1}) \|^2}{2}
\]

\[
= \frac{1}{\alpha_t} \varphi_{1/\hat{p}}(\theta_t) + \hat{\beta} \langle \hat{\theta}_t - \theta_t, G(\theta_t, x_{t+1}) \rangle + \frac{\alpha_t \hat{\beta} \| G(\theta_t, x_{t+1}) \|^2}{2}.
\]

Proceeding as in the proof of Theorem 4.1, namely, by taking expectation conditional on \( \mathcal{F}_{t-k} \), using Lemma 3.1, using (14), and then integrating \( \mathcal{F}_{t-k} \) out, we obtain

\[
\mathbb{E} \left[ \frac{1}{\alpha_t} \varphi_{1/\hat{p}}(\theta_{t+1}) \right] - \mathbb{E} \left[ \frac{1}{\alpha_t} \varphi_{1/\hat{p}}(\theta_t) \right] \leq \hat{\beta} \mathbb{E} \left[ f(\hat{\theta}_t) - f(\theta_t) + \frac{\hat{\beta}}{2} \| \theta_t - \hat{\theta}_t \|^2 \right] + \mathbb{E} \left[ \frac{\alpha_t \hat{\beta} \| G(\theta_t, x_{t+1}) \|^2}{2} \right]
\]

\[
+ \hat{\beta} \mathbb{E} \left[ \frac{2L^2}{\hat{\beta} - \rho} \Delta_{[t-k,t+1]} + \frac{2L(L_1 + \hat{\beta})}{\hat{\beta} - \rho} \sum_{s=t-k}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right].
\]

The only difference from before is that while bounding \( \| \theta_s - \theta_{s-1} \| \) in Lem. 3.1 we did not use the worst case bound for \( \| G(\theta_s, x_{s+1}) \| \) as in (A2).

We use (16) on this inequality with \( k = k_t \) where \( k_t \) is nondecreasing, sum for \( t \in \{1, 2, \ldots, T\} \) and rearrange to get

\[
\frac{\hat{\beta} - \rho}{\hat{\beta}} \sum_{t=1}^{T} \mathbb{E} \left[ \| \nabla \varphi_{1/\hat{p}}(\theta_t) \|^2 \right] \leq \sum_{t=1}^{T} \mathbb{E} \left[ \frac{\varphi_{1/\hat{p}}(\theta_{t+1}) - \varphi_{1/\hat{p}}(\theta_t)}{\alpha_t} \right] + \sum_{t=1}^{T} \mathbb{E} \left[ \frac{\alpha_t \hat{\beta} \| G(\theta_t, x_{t+1}) \|^2}{2} \right]
\]

\[
+ \sum_{t=1}^{T} \hat{\beta} \mathbb{E} \left[ \frac{2L^2}{\hat{\beta} - \rho} \Delta_{[t-k_t,t+1]} + \frac{2L(L_1 + \hat{\beta})}{\hat{\beta} - \rho} \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right].
\]

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

\[
\sum_{t=1}^{T} \alpha_t \| G(\theta_t, x_{t+1}) \|^2 = \sum_{t=1}^{T} \frac{\alpha}{\sqrt{v_0 + \sum_{j=1}^{t} \| G(\theta_j, x_{j+1}) \|^2}} \| G(\theta_t, x_{t+1}) \|^2 \leq 2 \sqrt{\sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2}
\]

\[
\sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2 \leq T \sqrt{\sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2} \quad \text{(21)}
\]
By taking expectation, and using Jensen’s inequality, we get
\[
\mathbb{E} \left[ \sum_{t=1}^{T} \alpha_t \| G(\theta_t, x_{t+1}) \| ^2 \right] \leq \sqrt{2} \mathbb{E} \left[ \sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \| ^2 \right] \leq 2 \sqrt{\mathbb{E} \left[ \sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \| ^2 \right]} \leq 2 \sqrt{TVL}.
\]

We next use (A4) to obtain
\[
\sum_{t=1}^{T} \left( \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \right) | \varphi_{1/\beta}(\theta_t) | \leq \sum_{t=1}^{T} \left| \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \right| \leq C_p \sum_{t=1}^{T} \left( \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} \right) \leq C_p \frac{\sqrt{v_0 + TL^2}}{\alpha}.
\]

since \( \frac{1}{\alpha_t} \) is monotonically nondecreasing in \( t \).

It remains to estimate the last term on (20) which is the main additional error term that is due to dependent data. For convenience, let us define \( \alpha_s \| G(\theta_s, x_{s+1}) \| = 0 \) for \( s \leq 0 \). Then we have
\[
\sum_{t=1}^{T} \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \leq \sum_{t=1}^{T} \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \|,
\]
where \( \alpha_s = \frac{\alpha}{\sqrt{v_0 + \sum_{j=1}^{s} \| G(\theta_j, x_{j+1}) \|}^2} \) and the inequality used that \( k_t \) is nondecreasing.

By Young’s inequality, we can upper bound this term as
\[
\sum_{t=1}^{T} \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \leq \sum_{t=1}^{T} \sum_{s=t-k_t}^{t-1} \left( \frac{k_t}{2} \right)^{1/4} \left( \frac{2}{k_t} \right)^{1/2} \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \leq \sum_{t=1}^{T} \frac{k_t + 1}{2\sqrt{t}} + \sum_{t=1}^{T} \sqrt{t} \left( \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right)^2 \leq \sqrt{t} k_t + \sum_{t=1}^{T} \sqrt{t} \left( \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right)^2.
\]

We now have that the rightmost summation in (24) is of the form in the first inequality in Lem. A.6. We continue from (24) by using the definition of \( \alpha_t \)
\[
\sum_{t=1}^{T} \frac{\sqrt{t}}{2k_t} \left( \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right)^2 \leq \frac{\sqrt{t}}{2} \sum_{t=1}^{k_t} \sum_{s=t-k_t}^{t-s} \alpha_s^2 \| G(\theta_s, x_{s+1}) \| ^2 \leq \sqrt{t} \sum_{s=1}^{k_t} \sum_{t=1}^{T-s} c_t.
\]

since for any \( (c_t) \), we have \( \sum_{t=1}^{T} \sum_{s=t-k_t+1}^{T-s} c_s = (c_2 - c_1 + c_3 - c_2 + \cdots + c_1) + (c_3 - c_2 + c_4 - c_3 + \cdots + c_2) + \cdots + (c_T - c_{T-1} + c_T - c_{T-2} + \cdots + c_{T-1}) = (c_2 - c_1 + c_3 - c_2 + \cdots + c_T - c_{T-1}) + (c_3 - c_2 + c_4 - c_3 + \cdots + c_T - c_{T-2} + \cdots + c_1 + c_2 + \cdots + c_T) = \sum_{s=1}^{k_t} \sum_{t=1}^{T-s} c_t. \)

Since in our case \( c_t = 0 \) for \( t < 1 \), we have also that \( \sum_{t=1}^{k_t} \sum_{t=1}^{T-s} c_t = \sum_{t=1}^{k_t} \sum_{t=1}^{T-s} c_t. \)

We now have that the rightmost summation in (24) is of the form in the first inequality in Lem. A.6. We continue from (24) by using the definition of \( \alpha_t \)
\[
\sum_{t=1}^{T} \frac{\sqrt{t}}{2k_t} \left( \sum_{s=t-k_t}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \right)^2 \leq \frac{\sqrt{t}}{2} \sum_{t=1}^{k_t} \sum_{s=t-k_t}^{t-s} \alpha_s^2 \| G(\theta_s, x_{s+1}) \| ^2 \]
\[
= \frac{\sqrt{t}}{2} \sum_{s=1}^{k_t} \sum_{t=1}^{T-s} \alpha_s^2 \| G(\theta_s, x_{s+1}) \| ^2 \]
Collecting (21), (22) and (25) on (20) results in the bound
\[ \frac{\sqrt{T} \alpha^2}{2} \sum_{s=1}^{k_T} \log \left( 1 + v_0^{-1} \sum_{t=1}^{T-s} \| G(\theta_t, x_{t+1}) \|^2 \right) \]
\[ \leq \frac{\sqrt{T} k_T \alpha^2}{2} \log \left( 1 + v_0^{-1} \sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2 \right), \]
where the third line applies the second inequality in Lem. A.6. Using this estimation on (23) gives us
\[ \sum_{t=1}^{T} \sum_{s=t-k_T}^{t-1} \alpha_s \| G(\theta_s, x_{s+1}) \| \leq \sqrt{T} k_T + \frac{\sqrt{T} k_T \alpha^2}{2} \log \left( 1 + v_0^{-1} \sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2 \right). \] (25)
Collecting (21), (22) and (25) on (20) results in the bound
\[ \frac{\hat{\rho} - \rho}{\hat{\rho}} \sum_{t=1}^{T} \mathbb{E} \left[ \| \nabla \psi_{1/\hat{\rho}}(\theta_t) \|^2 \right] \leq \frac{\sqrt{v_0 + TL^2 C^2}}{\alpha} + \rho L \sqrt{T} + \sum_{t=1}^{T} \rho E \left[ \frac{2 L^2}{\hat{\rho} - \rho} \Delta_{t-k,t+1} \right] + 2 L \rho \left( \sqrt{T} k_T + \frac{\sqrt{T} k_T \alpha^2}{2} \log \left( 1 + v_0^{-1} \sum_{t=1}^{T} \| G(\theta_t, x_{t+1}) \|^2 \right) \right). \]
We divide both sides by $T$ to conclude.

\[ \square \]

6. 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 [Pol64]. This method is analyzed for convex optimization in [GFJ15] and for constrained and stochastic nonconvex optimization with i.i.d. data in [MJ20]. Some features of our analysis simplify and relax some conditions from the analysis in [MJ20] even with i.i.d. data, see Lem. 6.1 and Remark 6.3 for the details.

**Algorithm 3** Stochastic heavy ball (momentum SGD)

1: **Input:** Initialize $\theta_0 \in \Theta \subseteq \mathbb{R}^p$; $T > 0$; $(\alpha_t)_{t \geq 1}$; $\beta > 0$; $z_1 > 0$
2: Optionally, sample $\tau$ from $\{1, \ldots, T\}$ independently of everything else where $P(\tau = k) = \frac{a_k}{\sum_{i=1}^k a_i}$.
3: **For** $t = 1, 2, \ldots, T$ **do**:
4: Sample $x_{t+1}$ from the conditional distribution $\pi_{t+1} = \pi_{t+1}(\cdot | x_1, \ldots, x_t)$
5: $\theta_{t+1} \leftarrow \text{proj}_{\Theta}(\theta_t - \alpha_t z_t)$
6: $z_{t+1} = \beta G(\theta_{t+1}, x_{t+1}) + \frac{1-\beta}{\alpha_{t+1}} (\theta_t - \theta_{t+1})$
7: **End** **for**
8: **Return:** $\theta_T$ (Optionally, return $\theta_t$)

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 [MJ20] that analyzed the algorithm in the i.i.d. case.

**Lemma 6.1.** Let $(z_t)$ be defined as Alg. 3 and let (A3) hold. Then, we have
\[ \| z_{t+1} \|^2 \leq \beta L + (1 - \beta)(\alpha_t / \alpha_{t+1})^2 \| z_t \|^2 \quad \text{for all } t \geq 1 \]
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
\[ \| z_{t+1} \|^2 \leq \beta \| G(\theta_{t+1}, x_{t+1}) \|^2 + \frac{1-\beta}{\alpha_{t+1}^2} \| \theta_t - \theta_{t+1} \|^2 \]
\[ \leq \beta \| G(\theta_{t+1}, x_{t+1}) \|^2 + \frac{(1 - \beta) \alpha_t^2}{\alpha_{t+1}^2} \| z_t \|^2, \]

where the second inequality used that \( \theta_t \in \Theta \) and that \( \text{proj}_\Theta \) is nonexpansive. Using (A3) and dividing both sides by \( \alpha_{t+1}^2 \) gives the first inequality in the assertion. Also, by multiplying both sides of the inequality by \( \alpha_{t+1}^2 \), we have

\[ \alpha_{t+1}^2 \| z_{t+1} \|^2 \leq \beta \alpha_{t+1}^2 \| G(\theta_t, x_{t+1}) \|^2 + (1 - \beta) \alpha_t^2 \| z_t \|^2. \]  

(26)

By using (A3) in (26), then rearranging, multiplying both sides by \( t^\delta \), and summing (26) give

\[ \sum_{t=1}^T \beta t^\delta \alpha_t^2 \| z_t \|^2 \leq -T^\delta \alpha_{T+1}^2 \| z_{T+1} \|^2 + \alpha_2^2 \| z_1 \|^2 + \beta L^2 \sum_{t=1}^T t^\delta \alpha_t^2. \]

Removing the nonpositive term on the RHS gives the result.

\[ \square \]

**Theorem 6.2** (Stochastic heavy ball). Let (A1)-(A3) hold. Let \( \{\theta_t\}_{t \geq 1} \) be a sequence generated by Algorithm 3. Fix \( \hat{\rho} \geq 2\rho \). Then, for any \( \beta \in (0, 1) \),

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

\[ \mathbb{E} \left[ \| \nabla \varphi_{1/\hat{\rho}}(\bar{\theta}_T) \|^2 \right] \leq \frac{\hat{\rho}}{\sum_{t=1}^T \alpha_t} \left( \varphi_{1/\hat{\rho}}(\bar{\theta}_0) - \inf_{\varphi_{1/\hat{\rho}}} + \frac{(1 + \beta(1 - \beta))L}{2\beta^2} \sum_{t=1}^T \alpha_t^2 + \frac{1 - \beta}{2\beta^2} \alpha_t \| z_t \|^2 \right) \]

\[ + \frac{2L^2}{\beta - \rho} \sum_{t=1}^T \alpha_t \| \Delta_{t-k, t} \|_1 + \frac{2L^2(1 + \rho)}{\beta - \rho} \sum_{t=1}^T k_t \alpha_t \alpha_{t-k}. \]

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

**Remark 6.3.** Our analysis is more flexible compared to [MJ20] even when restricted to the i.i.d. case. In this case, we allow variable step sizes \( \alpha_t = \frac{T}{T+1} \) whereas [MJ20] requires constant step size \( \alpha_t = \alpha = \frac{T}{\sqrt{T}} \). We can also use any \( \beta \in (0, 1) \) whereas [MJ20] restricts to \( \beta = \alpha \). This point is important since in practice \( \beta \) is used as a tuning parameter.

**Proof.** We proceed as the proof of Thm. 4.1. However, following the existing analyses for SHB [GFJ15, MJ20] we use the following iterate \( \hat{\theta}_t = \theta_t + \frac{1}{\hat{\rho}} (\theta_t - \theta_{t-1}) \) and also \( \hat{\theta}_t = \text{prox}_{\varphi_{1/\hat{\rho}}}(\bar{\theta}_t) \). The useful property of \( \hat{\theta}_t \) exploited in [MJ20] with constant step sizes (and also in [GFJ15] in the unconstrained setting), is that

\[ \| \hat{\theta}_{t+1} - \hat{\theta}_t \|^2 = \| \theta_{t+1} + \frac{1 - \beta}{\beta} (\theta_{t+1} - \theta_t) - \hat{\theta}_t \|^2 = \frac{1}{\beta^2} \| \theta_{t+1} - [(1 - \beta)\theta_t + \beta \hat{\theta}_t] \|^2 \]

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

\[ = \| \hat{\theta}_t - \bar{\theta}_t - \alpha_t G(\theta_t, x_t) \|^2. \]

(27)

where the inequality used that \( \theta_t, \theta_{t+1}, \hat{\theta}_t \) 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. 4.1, we evaluate \( \varphi_{1/\hat{\rho}} \) at \( \hat{\theta}_{t+1} \) instead of \( \theta_{t+1} \) and then use the inequality in (27) to deduce

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

\[ \leq \mathbb{E} \left[ f(\hat{\theta}_t) + \frac{\hat{\rho}}{2} \| \hat{\theta}_t - \alpha_t G(\theta_t, x_t) - \hat{\theta}_t \|^2 | \mathcal{F}_{t-k} \right]. \]

(28)
We expand the square to obtain
\[ \| \hat{\theta}_t - \alpha_t G(\theta_t, x_t) - \hat{\theta}_t \|^2 = \| \hat{\theta}_t - \theta_t \|^2 - 2 \alpha_t \langle \hat{\theta}_t - \theta_t, G(\theta_t, x_t) \rangle + \alpha_t^2 \| G(\theta_t, x_t) \|^2. \]

By using the last estimate on (28) and using the definition of \( \varphi_{1/\hat{\rho}}(\hat{\theta}_t) \) along with (A3) gives
\[ \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\hat{\theta}_{t+1}) \big| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\hat{\theta}_t) - \hat{\rho} \alpha_t (\theta_t - \hat{\theta}_t, G(\theta_t, x_t)) - \frac{\hat{\rho} \alpha_t (1-\beta)}{\beta} (\theta_t - \theta_{t-1}, G(\theta_t, x_t)) \right. \\
\left. + \frac{\hat{\rho} \alpha_t^2 L^2}{2} \bigg| \mathcal{F}_{t-k} \right]. \tag{29} \]

We estimate the third term on RHS by Young’s inequality, the nonexpansiveness of the projection and (A3)
\[ - \frac{\hat{\rho} \alpha_t (1-\beta)}{\beta} \langle \theta_t - \theta_{t-1}, G(\theta_t, x_t) \rangle \leq \frac{\hat{\rho} (1-\beta)}{2 \beta} \left( \| \theta_t - \theta_{t-1} \|^2 + \alpha_t^2 \| G(\theta_t, x_t) \|^2 \right) \]
\[ \leq \frac{\hat{\rho} (1-\beta)}{2 \beta} \left( \alpha_{t-1}^2 \| z_{t-1} \|^2 + \alpha_t^2 L^2 \right). \]

We insert this estimate back to (29) and use Lem. 3.1 as in the proof of Thm. 4.1 to obtain
\[ \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\hat{\theta}_{t+1}) \big| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\hat{\theta}_t) - \hat{\rho} \alpha_t \langle \theta_t - \hat{\theta}_t, \theta_{t-1}, G(\theta_t, x_t) \rangle + \frac{\hat{\rho} (1-\beta)}{2 \beta} \alpha_{t-1}^2 \| z_{t-1} \|^2 + \frac{\hat{\rho} (2-\beta) \alpha_t^2 L^2}{2 \beta} \right| \mathcal{F}_{t-k} \]
\[ \leq \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\hat{\theta}_t) \big| \mathcal{F}_{t-k} \right] - \hat{\rho} \alpha_t \langle \theta_t - \hat{\theta}_t, \theta_{t-1}, G(\theta_t, x_t) \rangle \]
\[ + \hat{\rho} \alpha_t \left\{ \frac{2L^2}{\beta-\rho} \Delta_{[t-k,t+1]} + k \frac{2L^2 L_1 + \hat{\rho} L^2}{\beta-\rho} \right\} + \frac{\hat{\rho} (1-\beta)}{2 \beta} \alpha_{t-1}^2 \mathbb{E} \left[ \| z_{t-1} \|^2 \bigg| \mathcal{F}_{t-k} \right]. \tag{30} \]

We now estimate the second term on the RHS
\[ \langle \theta_t - \hat{\theta}_t, \theta_{t-1}, G(\theta_t, x_t) \rangle = f(\theta_t) - f(\hat{\theta}_t) - \frac{\rho}{2} \| \theta_t - \hat{\theta}_t \|^2 \]
\[ = \left( f(\theta_t) + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|^2 \right) - \left( f(\hat{\theta}_t) + \frac{\hat{\rho}}{2} \| \hat{\theta}_t - \hat{\theta}_t \|^2 \right) - \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|^2 + \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|^2 - \frac{\rho}{2} \| \theta_t - \hat{\theta}_t \|^2 \]
\[ \geq \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|^2 - \frac{\rho}{2} \| \theta_t - \hat{\theta}_t \|^2 \geq \frac{\hat{\rho}}{2} \| \theta_t - \hat{\theta}_t \|^2 - \frac{\hat{\rho} (1-\beta)^2 \alpha_{t-1}^2}{2 \beta^2} \| z_{t-1} \|^2. \tag{31} \]

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

We use (31) on (30), insert \( k = k_t \), integrate out \( \mathcal{F}_{t-k} \) and sum to get
\[ \sum_{t=1}^T \hat{\rho}^2 \alpha_t \mathbb{E} \left[ \| \theta_t - \hat{\theta}_t \|^2 \right] \leq - \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\theta_{T+1}) \right] + \mathbb{E} \left[ \varphi_{1/\hat{\rho}}(\theta_{T+1}) \right] + \sum_{t=1}^T \frac{\hat{\rho} (2-\beta) \alpha_t^2 L^2}{2 \beta} + \sum_{t=1}^T \frac{\hat{\rho} (1-\beta)^2 \alpha_{t-1}^2}{2 \beta^2} \mathbb{E} \left[ \| z_{t-1} \|^2 \right] \]
\[ + \sum_{t=1}^T \hat{\rho} \alpha_t \left( \frac{2L^2}{\beta-\rho} \mathbb{E} \left[ \| \theta_t - \hat{\theta}_t \|^2 \right] \right) + k_t \frac{2L^2 L_1 + \hat{\rho} L^2}{\beta-\rho} \mathbb{E} \left[ \| \theta_t - \hat{\theta}_t \|^2 \right] + \sum_{t=1}^T \frac{\hat{\rho} (1-\beta)^2 \alpha_{t-1}^2}{2 \beta^2} \mathbb{E} \left[ \| z_{t-1} \|^2 \right]. \tag{32} \]

Using Lem. 6.1 for the terms involving \( \| z_{t} \|^2 \) and using \( \| \nabla \varphi_{1/\hat{\rho}}(\theta_t) \| = \hat{\rho} \| \theta_t - \hat{\theta}_t \| \) finishes the proof of (i) after simple arrangements.

Next, we show (ii). The argument for the second part is identical to that of Theorem 6.2 (ii). The argument for the first part is also similar to that of Theorem 4.1 (ii) with a minor modification. Namely,
from (32) and the hypothesis,
\[ \sum_{t=1}^{T} \beta^2 \alpha_t \mathbb{E} \left[ \| \theta_t - \hat{\theta}_t \|^2 \right] < \infty. \]

Using Fubini’s theorem and (6), this implies
\[ \sum_{t=1}^{T} \alpha_t \| \nabla \phi_{1/ \beta}(\hat{\theta}_t) \|^2 < \infty \quad \text{almost surely.} \]

Hence by Lemma A.5, it suffices to show that
\[ \left| \| \nabla \phi_{1/ \beta}(\hat{\theta}_{t+1}) \|^2 - \| \nabla \phi_{1/ \beta}(\hat{\theta}_t) \|^2 \right| = O(\alpha_t). \]

Proceeding as in the proof of Theorem 4.1 (iii), 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 6.1. 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 6.1, we have
\[ \| z_{t+1} \|^2 \leq \beta L + 2(1 - \beta)(\alpha_t / \alpha_{t+1})^2 L. \]

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 \( t > t_0 \),
\[ \| z_{t+1} \|^2 \leq \beta L + (1 - \beta/2)(2L) = 2L. \]

This shows the assertion. \( \square \)

7. EXTENSION TO PROXIMAL CASE

In this section, we describe how our developments for stochastic subgradient method extends to the proximal case, using the ideas from [DD19]. In particular, the problem we solve in this section is
\[ \theta^* \in \text{arg min}_{\theta \in \mathbb{R}^p} \{ \phi(\theta) := f(\theta) + r(\theta) \}, \quad f(\theta) = \mathbb{E}_{x \sim \pi} [\ell(\theta, x)], \]

where \( r: \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) is a convex, proper, closed function. In this case, in step 5 of Algorithm 1, we use \( \text{prox}_{\alpha_t, r} \) instead of \( \text{proj}_\Theta \) to define \( \theta_{t+1} \).

Recall also that
\[ \hat{\theta}_t = \text{prox}_{\alpha_t/ \beta}(\theta_t). \]

In the projected case, when \( r(\theta) \) is the indicator function of the set \( \Theta \), we had that \( \hat{\theta}_t \in \Theta \). This was used, for example, in (12) to use nonexpansiveness for bounding \( \| \theta_{t+1} - \theta_t \|^2 \) since \( \theta_{t+1} = \text{proj}_\Theta(\theta_t - \alpha_t g_t) \). In this case, for the same step, one needs an intermediate result derived by [DD19].

**Lemma 7.1.** ([DD19, Lem. 3.2]) Given the definition of \( \hat{\theta}_t \), we have for \( t \geq 0 \)
\[ \hat{\theta}_t = \text{prox}_{\alpha_t/ \beta}(\alpha_t \hat{\theta}_t - \alpha_t \hat{\theta}_t + (1 - \alpha_t / \beta) \theta_t), \]

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

We include the following result combining the ideas from Lem. 3.1, Thm. 4.1 and [DD19, Sec. 3.2] for proving convergence of proximal stochastic subgradient algorithm with dependent data.

**Theorem 7.2** (Stochastic prox-subgradient method). Assume (A1)-(A3) hold. Let \( (\theta_t)_{t \geq 0} \) be a sequence generated by Algorithm 1 where we use \( \text{prox}_{\alpha_t, r} \) instead of \( \text{proj}_\Theta \) in step 5. Fix \( \hat{\beta} > \rho \). For each \( T \geq 1 \),
\[
\mathbb{E} \left[ \| \nabla \phi_{1/ \hat{\beta}}(\theta_T^{\text{out}}) \|^2 \right] \leq \frac{\hat{\beta}^2 L^2}{\hat{\beta} - \rho} \left[ \frac{1}{\sum_{k=1}^{T} \alpha_k} \left( \phi_{1/ \beta}(\theta_0) - \inf \phi_{1/ \beta} + 2 \sum_{t=1}^{T} \alpha_t^2 + 2(\hat{\beta} + L) \right) + \frac{2}{\hat{\rho} - \rho} \sum_{k=1}^{T} k_1 \alpha_t \alpha_t - k_1 + \frac{2}{\hat{\rho} - \rho} \sum_{t=1}^{T} \alpha_t \mathbb{E} [\Delta_{t-k_1, t+1}] \right].
\]
Proof. We start the same as Thm. 4.1 and note by the definition of $\theta_{t+1}$

$$\varphi_{1/\rho}(\theta_{t+1}) \leq \varphi(\hat{\theta}) + \frac{\hat{\rho}}{2} \|\theta_{t+1} - \hat{\theta}\|^2. \quad (33)$$

We next estimate $\frac{\delta}{2} \|\theta_{t+1} - \hat{\theta}\|^2$ similar to [DD19, Lem. 3.3] by using 1-Lipschitzness of prox$_{\alpha_t g}$. Let $\delta = 1 - \alpha_t \hat{\rho}$ and estimate

$$\|\theta_{t+1} - \hat{\theta}\|^2 = \|\text{prox}_{\alpha_t g}(\theta_t - \alpha_t g_t) - \text{prox}_{\alpha_t \hat{\rho} \theta_t - \alpha_t \hat{\nu}_t + \delta \hat{\theta})\|^2$$

$$\leq \delta^2 \|\theta_t - \hat{\theta}\|^2 - 2\delta \alpha_t \langle \theta_t - \hat{\theta}, G(\theta_t, x_t) - \hat{\nu}_t \rangle + \alpha_t^2 \|G(\theta_t, x_t) - \hat{\nu}_t\|^2. \quad (34)$$

where we skipped some intermediate steps, which are already in [DD19, Lem. 3.3]. We note that by Lem. 3.1, we have

$$-2\delta \alpha_t \mathbb{E} \left[ \langle \theta_t - \hat{\theta}, G(\theta_t, x_t) \rangle \bigg| \mathcal{F}_{t-k} \right] = -2\delta \alpha_t \langle \theta_t - \hat{\theta}, \mathbb{E}_{\pi}[G(\theta_t, x_t)] \rangle + 2\delta \alpha_t \left( \frac{2\rho^2}{\hat{\rho} - \rho} \Delta_{[t-k, t+1]} + \frac{2\rho^2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \alpha_{t-k} \right). \quad (35)$$

We take the conditional expectation of (33) and use (34) with (35) to derive

$$\mathbb{E} \left[ \varphi_{1/\rho}(\theta_{t+1}) \bigg| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ \varphi(\hat{\theta}) \bigg| \mathcal{F}_{t-k} \right] + \delta^2 \|\theta_t - \hat{\theta}\|^2$$

$$-2\delta \alpha_t \langle \theta_t - \hat{\theta}, \mathbb{E}_{\pi}[G(\theta_t, x_t)] \rangle - 2\delta \alpha_t \mathbb{E} \left[ \langle \theta_t - \hat{\theta}, -\hat{\nu}_t \rangle \bigg| \mathcal{F}_{t-k} \right]$$

$$+ \alpha_t^2 \mathbb{E} \left[ \|G(\theta_t, x_{t+1}) - \hat{\nu}_t\|^2 \bigg| \mathcal{F}_{t-k} \right] + 2\delta \alpha_t \left( \frac{2\rho^2}{\hat{\rho} - \rho} \Delta_{[t-k, t+1]} + \frac{2\rho^2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \alpha_{t-k} \right).$$

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

$$\mathbb{E} \left[ \varphi_{1/\rho}(\theta_{t+1}) \bigg| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ \varphi(\hat{\theta}) \bigg| \mathcal{F}_{t-k} \right] + \delta^2 \|\theta_t - \hat{\theta}\|^2 - 2\delta \alpha_t \mathbb{E} \left[ \langle \theta_t - \hat{\theta}, \mathbb{E}_{\pi}[G(\theta_t, x_t)] \rangle - \hat{\nu}_t \right]$$

$$+ \alpha_t^2 \mathbb{E} \left[ \|G(\theta_t, x_{t+1}) - \hat{\nu}_t\|^2 \bigg| \mathcal{F}_{t-k} \right] + 2\delta \alpha_t \left( \frac{2\rho^2}{\hat{\rho} - \rho} \mathbb{E}[\Delta_{[t-k, t+1]}] + \frac{2\rho^2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \alpha_{t-k} \right). \quad (36)$$

Next, we use that the subdifferential of $\rho$-weakly convex $g$ is $\rho$-hypo-monotone (see [DD19, Lem. 2.1]) and $\mathbb{E}_{\pi}[G(\theta_t, x_t)] \in \partial f(\theta_t)$ and $\hat{\nu}_t \in \partial f(\hat{\theta})$ to derive

$$\langle \theta_t - \hat{\theta}, \mathbb{E}_{\pi}[G(\theta_t, x_t)] - \hat{\nu}_t \rangle \geq -\rho \|\theta_t - \hat{\theta}\|^2. \quad (37)$$

We combine (37) with $\|\hat{\nu}_t\|^2 \leq L^2$ (see [DD19, Sec. 3.2]) on (36) to derive

$$\mathbb{E} \left[ \varphi_{1/\rho}(\theta_{t+1}) \bigg| \mathcal{F}_{t-k} \right] \leq \mathbb{E} \left[ \varphi(\hat{\theta}) \bigg| \mathcal{F}_{t-k} \right] - \rho(\hat{\rho} - \rho) \alpha_t \mathbb{E}\|\theta_t - \hat{\theta}\|^2 + 4\alpha_t^2 L^2$$

$$+ 2\delta \alpha_t \left( \frac{2\rho^2}{\hat{\rho} - \rho} \mathbb{E}[\Delta_{[t-k, t+1]}] + \frac{2L^2(L_1 + \hat{\rho})}{\hat{\rho} - \rho} \alpha_{t-k} \right).$$

We sum the inequality and argue similarly as in the proof of Theorem 4.1 to finish the proof. \qed

8. Complexity for constrained smooth optimization

We next specialize our results to the case when $g$ is differentiable with Lipschitz continuous gradient. In this case, we are going to compare our complexity with the one derived in [Lyu22] for constrained smooth nonconvex optimization. First, we introduce the next assumption to replace (A2).

(A2'). The function $g$ is $C^1$ smooth and has $\rho$ Lipschitz gradient.

We will now show how to translate the result to a direct stationarity measure in view of Sec. 2.1 to compare with the result in [Lyu22].
Theorem 8.1 (Sample complexity). Assume (A1), (A2'), (A3) hold. Let \((\theta_t)_{t \geq 1}\) be a sequence generated by any of the Algorithms 1, 2, and 3. Fix \(\hat{\rho} > \rho\) and a diverging sequence \((k_t)_{t \geq 0}\). Further assume that \(\Theta\) is compact. Pick \(\hat{t}\) randomly from \([1, \ldots, T]\) as in the respective theorems for the algorithms and let \(\hat{\theta}_{t+1} = \text{proj}_\Theta \left( \theta_{\hat{t}} - \frac{1}{N} \sum_{i=1}^{N} \nabla \ell (\theta_{\hat{t}}, x^{(i)}) \right) \) with \(\hat{N} = O(\varepsilon^{-2})\) samples. Then we have that
\[
E \left[ \text{dist}(0, \partial (f + t\theta) (\hat{\theta}_{t+1})) \right] \leq \varepsilon \quad \text{with} \quad \bar{O}(\varepsilon^{-4}) \text{ samples.}
\]

Proof. When \(g\) is smooth, we can use [DD19, Sec. 2.2] to show that for any \(\theta\),
\[
\|g_{1/2\hat{\rho}}(\theta)\| \leq \frac{3}{2} \|\nabla \psi_{1/\hat{\rho}}(\theta)\|.
\]
This establishes that the upper bound of Thm. 4.1 also upper bounds the norm of the gradient mapping \(\|g_{1/2\hat{\rho}}(\theta)\|\). By invoking Thm. 4.1 with a randomly selected iterate, this establishes the bound required for Lem. 8.2 and then applying Lem. 8.2 gives the result.

Even though gradient of the Moreau envelope is a near approximate stationarity measure, in the specific case of minimizing smooth nonconvex functions over a constraint set, we can output a point that is approximately stationary with respect to the direct stationarity measure in Prop. A.1(i), which permits a direct comparison with the previous result on constrained nonconvex optimization with dependent data [Lyu22].

Lemma 8.2. Let (A1), (A2'), (A3) hold, \(\Theta\) be compact, and \(\Delta_{t-k, t+1} = O(\lambda^{k})\) for \(\lambda < 1\). Let an algorithm output \(\theta_t\) (for example, a randomly selected iterate) such that \(E \|\theta_t - \text{proj}_\Theta (\theta_t - \nabla f(\theta_t))\| \leq \varepsilon\) with \(\bar{O}(\varepsilon^{-4})\) queries to \(\nabla \ell(\theta, x)\). Then, for \(\hat{\theta}_{t+1} = \text{proj}_\Theta (\theta_t - \tilde{\nabla} f(\theta_t))\) with \(\tilde{\nabla} f(\theta_t) = \frac{1}{\hat{N}} \sum_{i=1}^{\hat{N}} \nabla \ell(\theta_t, x^{(i)})\) with \(\hat{N} = O(\varepsilon^{-2})\) samples, we have that
\[
E \left[ \text{dist}(0, \partial (f + t\theta) (\hat{\theta}_{t+1})) \right] \leq \varepsilon \quad \text{with} \quad \bar{O}(\varepsilon^{-4}) \text{ samples.}
\]

Proof. By the definition of \(\hat{\theta}_{t+1}\), we have that
\[
\theta_t - \tilde{\nabla} f(\theta_t) - \hat{\theta}_{t+1} \in \partial_{\hat{\theta}} (\hat{\theta}_{t+1}).
\]
As a result, we have
\[
E \left[ \text{dist}(0, \partial (f + t\theta) (\hat{\theta}_{t+1})) \right] = E \left[ \min_{v \in \partial_{\hat{\theta}} (\hat{\theta}_{t+1})} \|\nabla f(\hat{\theta}_{t+1}) + v\| \right] 
\leq E \|\nabla f(\hat{\theta}_{t+1}) - \hat{\theta}_{t+1} + \theta_t - \tilde{\nabla} f(\theta_t)\|.
\]
For convenience, let \(\hat{\theta}_{t+1} = \text{proj}_\Theta (\theta_t - \nabla f(\theta_t))\). We continue estimating the last inequality by using this definition, triangle inequality, nonexpansiveness of \(\text{proj}_\Theta\), and \(\rho\)-smoothness of \(f\)
\[
E \left[ \text{dist}(0, \partial (f + t\theta) (\hat{\theta}_{t+1})) \right] \leq E \left[ \|\theta_t - \hat{\theta}_{t+1}\| + \|\nabla f(\hat{\theta}_{t+1}) - \nabla f(\theta_t)\| + \|\tilde{\nabla} f(\theta_t) - \nabla f(\theta_t)\| \right] 
\leq E \left[ (1 + \rho) \|\theta_t - \hat{\theta}_{t+1}\| + \|\tilde{\nabla} f(\theta_t) - \nabla f(\theta_t)\| \right] 
\leq E \left[ (1 + \rho) \left(\|\theta_t - \theta_{t+1}\| + \|\hat{\theta}_{t+1} - \hat{\theta}_{t+1}\| + \|\tilde{\nabla} f(\theta_t) - \nabla f(\theta_t)\| \right) \right] 
\leq E \left[ (1 + \rho) \|\theta_t - \theta_{t+1}\| + (2 + \rho) \|\tilde{\nabla} f(\theta_t) - \nabla f(\theta_t)\| \right].
\]
By the assumption in the lemma, recall that we have \(\|\theta_t - \hat{\theta}_{t+1}\| \leq \varepsilon\), therefore we have to estimate the last term in the last inequality. We use [Lyu22, Lem. 7.1] (see also Lemma A.7) with \(\psi = \nabla \ell\) to get \(E \|\tilde{\nabla} f(\theta_t) - \nabla f(\theta_t)\| = O(\hat{N}^{-1/2})\) with \(\hat{N}\) samples and finish the proof. \(\square\)
Consider the online dictionary learning (ODL) problem, which is stated in the following stochastic program

\[
\min_{\theta \in \Theta} \left( f(\theta) := \mathbb{E}_{X \sim \pi} \left[ \ell(X, \theta) := \inf_{H \in \Theta} d(X, \theta H) + R(H) \right] \right),
\]

where \( d(\cdot, \cdot) : \mathbb{R}^{p \times n} \times \mathbb{R}^{p \times n} \to [0, \infty) \) is 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 \( X \). In words, we seek to learn a single dictionary matrix \( \theta \in \mathbb{R}^{p \times r} \) within the constraint set \( \Theta \) (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 \( \pi \). Here, we may put \( L_1 \)-regularization on \( H \) in order to promote dictionary \( \theta \) that enable sparse representation of observed data.

The most widely investigated instance of the above ODL problem is when \( d \) equals the squared Frobenius distance. In this case, Mairal et al. [MBPS10] provided an online algorithm based on the framework of stochastic majorization-minimization [Mai13]. The algorithm proceeds by minimizing a recursively defined upper-bounding quadratic surrogate of the empirical loss function:

\[
\text{Upon arrival of } X_n:
\begin{align*}
H_n &= \arg\min_{H \in \Theta} \|X_n - \theta_{n-1} H\|_F^2 + \lambda \|H\|_1 \\
A_n &= (1 - w_n) A_{n-1} + w_n H_n H_n^T \\
B_n &= (1 - w_n) B_{n-1} + w_n H_n X_n^T \\
\theta_n &= \arg\min_{\theta \in \Theta} \left( \text{tr}(\theta^T A_n \theta) - 2\text{tr}(\theta B_n) \right),
\end{align*}
\]

where \((w_n)_{n \geq 1}\) is a prescribed sequence of ‘adaptivity parameters’. A well-known result in [MBPS10] states that the above algorithm converges almost surely to the set of stationary points of the expected loss function \( f \) in (38), provided the data matrices \((X_n)_{n \geq 1}\) are i.i.d. according to the stationary distribution \( \pi \). Later, Lyu, Needell, and Balzano [LNB20] generalized the analysis to the case where \( X_n \)'s are given by a function of some underlying Markov chain. Recently, Lyu [Lyu22] provided the first convergence rate bound of the ODL algorithm (39) of order \( O((\log n)^{1+\epsilon} / n^{1/4}) \) for the empirical loss function and \( O((\log n)^{1+\epsilon} / n^{1/8}) \) for the expected loss function for arbitrary \( \epsilon > 0 \).

Suppose we are given a sequence of data matrices \((X_n)_{n \geq 1}\) that follows \( \pi \) in some asymptotic sense. Under some mild assumptions, one can compute the subgradient of the loss function \( \theta \rightarrow \ell(X_n, \theta) \) in two steps and can perform a standard stochastic projected gradient descent:

\[
\begin{align*}
H_n &= \arg\min_{H \in \Theta} \|X_n - \theta_{n-1} H\|_F^2 + \lambda \|H\|_1, \\
G(\theta_{n-1}, X_n) &\in \partial (\theta \rightarrow d(X_n, \theta H_n)), \\
\theta_n &= \text{Proj}_{\Theta}(\theta_{n-1} - \alpha_n G(\theta_{n-1}, X_n)).
\end{align*}
\]

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

\((A5)\). For each \( X \) and \( \theta \), \( \inf_{H \in \Theta} d(X, \theta H) + R(H) \) admits a unique solution in \( \Theta' \).

Note that \((A5)\) is trivially satisfied if \( R(H) \) contains a Tikhonov regularization term \( \kappa_2 \|H\|_F^2 \) for some \( \kappa_2 > 2 \). Under \((A5)\), Danskin's theorem [Ber97] implies that the function \( \theta \rightarrow \ell(X, \theta) \) is differentiable and satisfies \( \nabla_\theta \ell(X, \theta) = \nabla_\theta d(X, \theta H^*) \), where \( H^* \) is the unique solution of \( \inf_{H \in \Theta} d(X, \theta H) + \lambda \|H\|_1 \). Hence we may choose \( G(\theta_{n-1}, X_n) = \nabla_\theta d(X_n, \theta H_n) \) in (40) in this case.

Notice that (40) is a PSGD algorithm for the ODL problem (38), which is a constrained and non-convex minimization problem. Zhao et al. [ZTX17] provided asymptotic analysis of algorithm of the form (40) (especially for online nonnegative matrix factorization) for general dissimilarity metric \( d \) using projected SGD. For a wide class of dissimilarity metrics such as Csiszár \( f \)-divergence, Bregman divergence, \( \ell_1 \) and \( \ell_2 \) metrics, and Huber loss, they showed that when the data matrices are i.i.d. and
the stepsizes $\alpha_n$ are non-summable ($\sum_{n=1}^{\infty} \alpha_n = \infty$) and square-summable ($\sum_{n=1}^{\infty} \alpha_n^2 < \infty$), then the sequence of dictionary matrices $(\theta_n)_{n \geq 1}$ obtained by (40), regardless of initialization, converges almost surely to the set of stationary points of (38). Their asymptotic analysis uses a rather involved technique inspired from dynamical systems literature and does not provide any rate of convergence. Moreover, being a constrained nonconvex optimization problem, such asymptotic guarantees has not been available to the more general Markovian data setting, despite a Markovian SGD has recently been analyzed in the unconstrained nonconvex setting by Sun et al. [SSY18].

When the function $\theta \rightarrow \ell(X, \theta)$ for each $X$ is $\rho$-weakly convex for some $\rho > 0$, then the expected loss function in (38) is also $\rho$-weakly convex, so in this case a direct application of the main result in [DDKL20] would yield a rate of convergence of (40) with i.i.d. data matrices $X_n$ of rate $O((\log n) / n^{1/4})$. Such hypothesis of weak convexity of the loss function is implied by the following smoothness assumption:

$(A6)$. For each $X$ and $\theta$, the function $\theta \rightarrow \ell(X, \theta) = \inf_{H \in \Theta} (d(X, \theta H) + R(H))$ is $L$-smooth for some $L > 0$.

In [MBPS10, Prop. 2], it was shown that both $(A5)$ and $(A6)$ are verified when $d$ equals the following squared $\ell_2$-loss with Tikhonov regularization on $H$:

$$d(X, \theta H) = \|X - \theta H\|^2_2 + \kappa_2 \|H\|^2 + \lambda \|H\|_1,$$

where $\kappa_2 > 0$ and $\lambda \geq 0$. In such cases, our main results extends the theoretical guarantees of 40 to more general setting when $X_n$’s 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). To our best knowledge, this is the first time that PSGD with adaptive step sizes has been applied to ODL problems with provable global asymptotic convergence and complexity bounds for the general Markovian data case. We state these results in the following corollary.

**Corollary 9.1.** Consider the online dictionary learning problem (38) and assume $(A5)$. Suppose we have a sequence of data matrices $(X_t)_{t \geq 0}$ and let $(\theta_1)_{t \geq 1}$ be the sequence of dictionary matrices in $\Theta \subseteq \mathbb{R}^{p \times r}$ obtained by either of the three algorithms: Projected SGD (Algorithm 1), AdaGrad (Algorithm 2), and stochastic heavy ball (Algorithm 3). Suppose the following holds:

- $(a1)$ $\Theta$ is compact and the sequence of data matrices $(X_t)_{t \geq 0}$ that satisfy the assumption (A1) and has a compact support;
- $(a2)$ For each $X$, the function $\theta \rightarrow \ell(X, \theta)$ is $\rho$-weakly convex for some $\rho > 0$ over $\Theta$.

Then in all cases, the following rate of convergence bound holds:

$$\min_{1 \leq k \leq n} \mathbb{E} \left[ - \inf_{\theta \in \Theta} \left\langle \nabla f(\hat{\theta}_k), \frac{\theta - \hat{\theta}_k}{\|\theta - \hat{\theta}_k\|} \right\rangle \right] = O\left( \frac{\log n}{n^{1/4}} \right).$$

Moreover, sample $\hat{t} \in \{1, \ldots, T\}$ and compute $\hat{\theta}_{\hat{t}+1}$ as in Theorem 8.1. Then

$$\mathbb{E} \left[ \text{dist}(0, \delta(f + \epsilon_{\hat{t}})(\hat{\theta}_{\hat{t}+1})) \right] \leq \epsilon \text{ with } T = \tilde{O}(\epsilon^{-4}) \text{ samples}.$$

Furthermore, Projected SGD and SHB converges almost surely to the set of stationary point of the objective function for (38). In particular, the above results hold under $(a1)$ and when $d$ is as in (41).

**Proof.** Follows immediately from Theorems 4.1, 5.1, 6.2 and 8.1. For the last statement for squared Frobenius loss, see [MBPS10, Prop. 2] for verifying (A5) and (A6) and recall that (A6) implies $(a2)$. □

**Acknowledgments**

The authors are grateful to Stephen Wright for valuable discussions. Research of A. Alacaoglu was supported in part by NSF award 2023239 and DOE ASCR Subcontract 8F30039 from Argonne National Laboratory. Research of H. Lyu was partially supported by NSF DMS-2010035.
REFERENCES

[AMC21] Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher, Convergence of adaptive algorithms for constrained weakly convex optimization, Advances in Neural Information Processing Systems 34 (2021).

[BC11] Heinz H Bauschke and Patrick L Combettes, Convex analysis and monotone operator theory in hilbert spaces, vol. 408, Springer, 2011.

[Bert97] Dimitri P Bertsekas, Nonlinear programming, Journal of the Operational Research Society 48 (1997), no. 3, 334–334.

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

[DD19] Damek Davis and Dmitriy Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, SIAM Journal on Optimization 29 (2019), no. 1, 207–239.

[DDKL20] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee, Stochastic subgradient method converges on tame functions, Foundations of computational mathematics 20 (2020), no. 1, 119–154.

[DHS10] John Duchi, Elad Hazan, and Yoram Singer, Adaptive subgradient methods for online learning and stochastic optimization, Tech. Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010.

[DL18] Dmitriy Drusvyatskiy and Adrian S Lewis, Error bounds, quadratic growth, and linear convergence of proximal methods, Mathematics of Operations Research 43 (2018), no. 3, 919–948.

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

[DP19] Dmitriy Drusvyatskiy and Courtney Paquette, Efficiency of minimizing compositions of convex functions and smooth maps, Mathematical Programming 178 (2019), no. 1, 503–558.

[GFJ15] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson, Global convergence of the heavy-ball method for convex optimization, 2015 European control conference (ECC), IEEE, 2015, pp. 310–315.

[JRJ07] Björn Johansson, Maben Rabi, and Mikael Johansson, A simple peer-to-peer algorithm for distributed optimization in sensor networks, 2007 46th IEEE Conference on Decision and Control, IEEE, 2007, pp. 4705–4710.

[JRJ10] ______., A randomized incremental subgradient method for distributed optimization in networked systems, SIAM Journal on Optimization 20 (2010), no. 3, 1157–1170.

[Lev17] Kfir Levy, Online to offline conversions, universality and adaptive minibatch sizes, Advances in Neural Information Processing Systems 30 (2017).

[LNB20] Hanbaek Lyu, Deanna Needell, and Laura Balzano, Online matrix factorization for markovian data and applications to network dictionary learning, Journal of Machine Learning Research 21 (2020), no. 251, 1–49.

[LO19] Xiaoyu Li and Francesco Orabona, On the convergence of stochastic gradient descent with adaptive stepsizes, The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 983–992.

[LP17] David A Levin and Yuval Peres, Markov chains and mixing times, vol. 107, American Mathematical Soc., 2017.

[Lyu20] Hanbaek Lyu, Convergence of block coordinate descent with diminishing radius for nonconvex optimization, arXiv preprint arXiv:2012.03503 (2020).

[Lyu22] ______., Convergence and complexity of stochastic block majorization-minimization, arXiv preprint arXiv:2201.01652 (2022).

[Mai13] Julien Mairal, Stochastic majorization-minimization algorithms for large-scale optimization, Advances in Neural Information Processing Systems 26 (2013).

[MBPS10] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, Online learning for matrix factorization and sparse coding, Journal of Machine Learning Research 11 (2010), no. Jan, 19–60.

[MJ20] Vien Mai and Mikael Johansson, Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization, International Conference on Machine Learning, PMLR, 2020, pp. 6630–6639.

[MT12] Sean P Meyn and Richard L Tweedie, Markov chains and stochastic stability, Springer Science & Business Media, 2012.
Yu Nesterov, *Gradient methods for minimizing composite functions*, Mathematical programming **140** (2013), no. 1, 125–161.

Boris T Polyak, *Some methods of speeding up the convergence of iteration methods*, Ussr computational mathematics and mathematical physics **4** (1964), no. 5, 1–17.

R Tyrrell Rockafellar and Roger J-B Wets, *Variational analysis*, vol. 317, Springer Science & Business Media, 2009.

Maurice Sion, *On general minimax theorems.*, Pacific Journal of mathematics **8** (1958), no. 1, 171–176.

Tao Sun, Yuejiao Sun, Yangyang Xu, and Wotao Yin, *Markov chain block coordinate descent*, Computational Optimization and Applications **75** (2020), no. 1, 35–61.

Tao Sun, Yuejiao Sun, and Wotao Yin, *On markov chain gradient descent*, Advances in Neural Information Processing Systems, 2018, pp. 9896–9905.

Alexandre B Tsybakov, *Introduction to nonparametric estimation*, 2009, URL https://doi.org/10.1007/b13794. Revised and extended from the **9** (2004), no. 10.

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

Rachel Ward, Xiaoxia Wu, and Leon Bottou, *Adagrad stepsizes: Sharp convergence over nonconvex landscapes*, International Conference on Machine Learning, PMLR, 2019, pp. 6677–6686.

Renbo Zhao, Vincent Tan, and Huan Xu, *Online nonnegative matrix factorization with general divergences*, Artificial Intelligence and Statistics, 2017, pp. 37–45.
APPENDIX A. APPENDIX

The next result illustrates the connection between the two stationarity measures given in (3) to compare with the existing result in [Lyu22].

**Proposition A.1.** For each \( \theta^* \in \Theta \), \( v \in \partial f(\theta^*) \), and \( \epsilon > 0 \), following conditions are equivalent:

(i) \( \text{dist}(0, \partial f(\theta^*)) \leq \epsilon \);

(ii) \( \inf_{\theta \in \partial f(\theta^*)} \left\langle v, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle \leq \epsilon. \)

In particular, it holds that

\[
\text{dist}(0, \partial f(\theta^*) + N_{\Theta}(\theta^*)) = \inf_{v \in \partial f(\theta^*)} \left[ - \inf_{\theta \in \Theta} \left\langle v, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle \right].
\]

**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 \( u \in N_{\Theta}(\theta^*) \) and \( w \in B_\epsilon \) such that \( v + u + w = 0 \). So \( -v - w \in N_{\Theta}(\theta^*) \), which is equivalent to

\[
\inf_{\theta \in \Theta} \left\langle v + w, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle \geq 0.
\]

By Cauchy-Schwarz inequality, this implies

\[
- \inf_{\theta \in \Theta} \left\langle v, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle \leq \|w\| \leq \epsilon.
\]

Conversely, suppose (ii) holds. We let \( \mathcal{D}_{\leq 1}(\theta^*) \) denote the set of all feasible directions at \( \theta^* \) of norm bounded by 1, which consists of vectors of form \( a(\theta - \theta^*) \) for \( \theta \in \Theta \) and \( a \in (0, \|\theta - \theta^*\|^{-1}] \). Being the intersection of two convex sets, \( \mathcal{D}_{\leq 1}(\theta^*) \) is convex. Then applying the minimax theorem [Sio58, 3.3 Cor.] for the bilinear map \( \langle x, u \rangle \rightarrow \langle v + \epsilon u, x \rangle \) defined on the product of convex sets \( \mathcal{D}_{\leq 1}(\theta^*) \times B_1 \), observe that

\[
\sup_{u \in B_1} \inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \langle v + \epsilon u, x \rangle = \inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \sup_{u \in B_1} \langle v + \epsilon u, x \rangle
\]

\[
= \inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \langle v, x \rangle + \sup_{u \in B_1} \langle \epsilon u, x \rangle
\]

\[
= \inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \langle v, x \rangle + \epsilon \|x\|
\]

\[
= \inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \|x\| \left( \left\langle v, \frac{x}{\|x\|} \right\rangle + \epsilon \right) \geq 0.
\]

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

\[
\left\langle v, \frac{x}{\|x\|} \right\rangle + \epsilon = \left\langle v, \frac{\theta_x - \theta^*}{\|\theta_x - \theta^*\|} \right\rangle + \epsilon \geq \inf_{\theta \in \partial f(\theta^*)} \left\langle v, \frac{\theta - \theta^*}{\|\theta - \theta^*\|} \right\rangle + \epsilon \geq -\epsilon + \epsilon \geq 0.
\]

Attainment of the supremum at a \( u^* \) in (42) is guaranteed by strong duality, see [BC11, Prop. 19.19(ii, v)].

The above implies

\[
\inf_{x \in \mathcal{D}_{\leq 1}(\theta^*)} \langle v + \epsilon u^*, x \rangle \geq 0.
\]

Thus we conclude that \( -v - \epsilon u^* \in N_{\Theta}(\theta^*) \). Then (i) holds since \( \|u^*\| \leq 1 \). \( \square \)

**Proposition A.2.** Suppose \( f \) is \( p \)-weakly convex and \( \lambda < \rho^{-1} \). Then for each \( \theta \in \Theta \),

\[
\sup_{v \in \partial f(\hat{\theta})} \left\langle v(\hat{\theta}), \frac{\theta' - \hat{\theta}}{\|\theta' - \hat{\theta}\|} \right\rangle \leq \lambda^{-1} \|\hat{\theta} - \theta\| \leq \lambda^{-2} \|
\]

\[
\|f(\theta) - f(\hat{\theta})\|.
\]

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

\[
\langle v(\hat{\theta}) + \lambda^{-1}(\hat{\theta} - \theta), \theta' - \hat{\theta} \rangle \geq 0, \quad \forall \theta' \in \Theta.
\]
By rearranging and using Cauchy-Schwarz, this yields for all $\theta' \in \Theta$,
$$
\langle v(\hat{\theta}), \hat{\theta} - \theta' \rangle \leq \lambda^{-1} \langle \hat{\theta} - \theta, \theta' - \hat{\theta} \rangle \leq \lambda^{-1}\|\hat{\theta} - \theta\| \cdot \|\theta' - \hat{\theta}\|.
$$

Now assume $\theta' \neq \hat{\theta}$. Dividing both sides by $\|\theta' - \hat{\theta}\|$, we get
$$
-\left(\frac{v(\hat{\theta}) \cdot \theta' - \hat{\theta}}{\|\theta' - \hat{\theta}\|}\right) \leq \lambda^{-1}\|\hat{\theta} - \theta\| = \lambda^{-2}\|\nabla \varphi(\theta)\|.
$$

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

The next two results will be used in Lem. 3.1 to control the bias due to dependent data.

**Lemma A.3.** [RW09, Prop. 12.19] For any $\tilde{\rho} \geq \rho$ and $\rho$-weakly convex function $\varphi$, it follows that $\theta \mapsto \text{prox}_{\varphi/\tilde{\rho}}(\theta)$ is $\frac{\tilde{\rho}}{\rho-\tilde{\rho}}$-Lipschitz.

**Lemma A.4.** [AMC21, Lem. 1] Let $\hat{\theta} \geq \rho$ and $\|G(x, \theta)\| \leq L$. Then,
$$
\|\hat{\theta} - \theta\| \leq \frac{2L}{\hat{\rho} - \rho}.
$$

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

**Lemma A.5.** [Mai13, Lem. A.5] 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) $\min_{1 \leq k \leq n} b_k \leq \sum_{k=0}^{\infty} a_k b_k = O\left(\left(\sum_{k=1}^{n} a_k\right)^{-1}\right)$.

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

The proof of (ii) is omitted and can be found in [Mai13, Lem. A.5].

The next lemma is commonly used for adaptive gradient algorithms. For example, see [Lev17, Lem. D.2, Lem. A.1] or [DHS10, Lem. 12].

**Lemma A.6.** [Lev17, Lem. D.2, Lem. A.1] For nonnegative real numbers $a_i$ for $i \geq 1$, we have for any $v_0 > 0$
$$
\frac{\sum_{i=1}^{n} a_i}{v_0 + \sum_{j=1}^{i} a_j} \leq \log \left(1 + \frac{\sum_{i=1}^{n} a_i}{v_0}\right) \quad \text{and} \quad \frac{\sum_{i=1}^{n} a_i}{v_0 + \sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i}.
$$

The following uniform concentration lemma for vector-valued parameterized observables is due to [Lyu22, Lem. 7.1].

**Lemma A.7.** [Lyu22, Lem. 7.1] Fix compact subsets $\mathcal{X} \subseteq \mathbb{R}^q$, $\Theta \subseteq \mathbb{R}^p$ and a bounded Borel measurable function $\psi: \mathcal{X} \times \Theta \to \mathbb{R}^r$. 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 $\Omega$ and $\varphi: \Omega \to \mathcal{X}$ is a measurable function. Assume the following:

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

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 $\tilde{\psi}(\cdot) := E_{x \sim \pi} [\psi(x, \cdot)]$ and $\psi_n: \Theta \to \mathbb{R}^r$ recursively as $\psi_0 \equiv 0$ and
$$
\psi_n(\cdot) = (1 - w_n) \psi_{n-1}(\cdot) + w_n \psi(X_n, \cdot).
$$

Then there exists a constant $C > 0$ such that for all $n \geq 1$,
$$
\sup_{\theta \in \Theta} \|\psi(\theta) - E[\psi_n(\theta)]\| \leq C w_n, \quad E\left[\sup_{\theta \in \Theta} \|\psi(\theta) - \psi_n(\theta)\|\right] \leq C w_n \sqrt{n}.
$$

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