A MINI-BATCH PROXIMAL ADAM ALGORITHM FOR NONSMOOTH NONCONVEX STOCHASTIC COMPOSITE OPTIMIZATION VIA MCMC SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose MCMC-ProxAdam, a novel stochastic proximal gradient algorithm designed to solve constrained nonconvex stochastic composite optimization problems with a nonsmooth convex regularizer, in settings where the gradient of smooth part is intractable and must be estimated via MCMC sampling. We depart from the traditional assumption of bounded stochastic gradient by introducing bounded sub-exponential norm for the stochastic gradient, then we derive a Bernstein-type inequality for the mini-batch stochastic estimator under non-stationary Markov chains, allowing our framework to accommodate biased and unbounded stochastic gradient common in practical scenarios. Leveraging this Bernstein-type inequality, we derive error bounds for the mini-batch gradient estimator and establish the convergence result for MCMC-ProxAdam in the nonsmooth nonconvex setting. Numerical experiments on energy-based models (EBMs) for image generation tasks across the MNIST, CIFAR-10, and CelebA datasets demonstrate that the proposed algorithm is capable of solving nonconvex composite optimization problems with a convex nonsmooth term, with numerical performance empirically validating our theoretical convergence results.

1 INTRODUCTION

In this paper, we study the following problem

$$\Phi^{\rm opt} := \min_{\theta\in\Theta}\{\Phi(\theta) := f(\theta) + h(\theta)\}, \tag{1}$$

where Θ is a closed convex set in Euclidean space $\mathbb{R}^l$, $f : \Theta \to \mathbb{R}$ is continuously differentiable, but possibly nonconvex, and h is a simple nonsmooth convex function with efficiently computable proximal operator. We also assume that the gradient of $f \in \mathcal{C}_L^{1,1}(\Theta)$ is L-Lipschitz continuous for some $L > 0$, i.e.,

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \le L\|\theta_1 - \theta_2\|, \quad \text{for any } \theta_1, \theta_2 \in \Theta, \tag{2}$$

and Φ is bounded below over Θ, i.e., $\Phi^{\rm opt}$ is finite. This formulation is widespread to machine learning and deep learning, where $f(\theta)$ represents the loss function, typically expressed as the online formulation $f(\theta) = \mathbb{E}_{z\sim\mathbb{P}}[F(\theta; z)]$ or the finite-sum formulation $f(\theta) = \frac{1}{N}\sum_{i=1}^N f_i(\theta)$. The function $h(\theta)$ acts as a regularization term on the parameter θ to mitigate overfitting and encourages θ to exhibit low-dimensional structural characteristics such as sparsity and low-rank structures.

In large-scale optimization scenarios, evaluating the exact gradient $\nabla f(\theta)$ is computationally prohibitive. To tackle this issue, stochastic first-order algorithms are widely adopted, which utilize gradient estimator instead of $\nabla f(\theta)$ to solve problems. The online setting has been explored in (Davis & Drusvyatskiy, 2019; Ding et al., 2025; Ghadimi et al., 2016; Pham et al., 2020), while the finite-sum setting has been extensively investigated in (Allen-Zhu, 2017; Chen et al., 2019b; Defazio et al., 2014; Li & Li, 2018; Paquette et al., 2018; Pham et al., 2020; Reddi et al., 2016; Wang et al., 2019). In the aforementioned literature, $\nabla f(\theta)$ is typically estimated via the stochastic gradient $G(\theta; z)$, which is assumed to be unbiased and possess a bounded variance. Specifically, at each iteration t, given the current iterate $\theta_t \in \Theta$ and a random variable $z_t \sim \mathbb{P}$, the stochastic gradient $G(\theta_t; z_t)$ is provided by a stochastic first-order oracle ($\mathcal{SFO}$).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In many applications, such as graphical models (Ekeberg et al., 2013) and energy-based models (Du & Mordatch, 2019), $\nabla f(\theta)$ can be expressed as a very high-dimensional integral over the sample space $\mathcal{X}$ with respect to the associated Gibbs measure μ_θ:

$$\nabla f(\theta) = \int_{z\in\mathcal{X}} G(\theta; z)\mu_\theta(\mathrm{d}z). \tag{3}$$

In Ekeberg et al. (2013) and Du & Mordatch (2019), the Gibbs measure μ_θ is parameterized as

$$\mu_\theta(z) = \frac{e^{\phi_\theta(z)}}{\int_{z\in\mathcal{X}} e^{\phi_\theta(z)}\mathrm{d}z} \propto e^{\phi_\theta(z)}, \tag{4}$$

where $\phi_\theta(z)$ is an energy function that may be highly complex, and could even take the form of a neural network. This results in $\mu_\theta(z)$ being unknown as the normalization constant $\int_{z\in\mathcal{X}} e^{\phi_\theta(z)}\mathrm{d}z$ is typically intractable to compute in practice, which makes it inaccessible to estimate $\nabla f(\theta)$. Another challenge arises in Monte Carlo EM for latent-variable models in statistics. In the expectation step of Monte Carlo EM, one needs to sample from the posterior distribution of the latent variables given the observed data. Similar to the Gibbs measure, this posterior distribution depends on the parameters being optimized and features an intractable normalization constant. Consequently, MCMC methods are required to generate these posterior samples, which are subsequently used to approximate the expectation and estimate the gradient for parameter optimization (McCulloch, 1997). As a result, the MCMC method is employed to solve these difficulties. A defining strength of the MCMC method lies in its ability to construct a Markov chain with suitable transition kernel, such that the Markov chain's stationary distribution coincides exactly with the target distribution $\mu_\theta(z)$. Critically, the sampling process relies on the ratio of unnormalized densities that circumvents the need to compute the normalization constant. Common MCMC algorithms include Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953), Hamiltonian Monte Carlo (Duane et al., 1987), Stochastic Gradient Langevin Dynamics (Welling & Teh, 2011), etc.

In cases where the stochastic gradients/subgradients are computed using samples generated via MCMC method, the samples $z_1, z_2, \ldots$ are generated from a Markov chain whose target distribution is $\mu_\theta(z)$, but which in practice is run for only finitely many steps and thus does not reach its stationary distribution. Consequently, the gradient estimator is biased, and the correlation between samples may cause the variance boundedness condition to fail in the mini-batch setting, rendering classical convergence frameworks that rely on the unbiased stochastic gradients/subgradients with bounded variance no longer applicable here. Existing literature on optimization with Markovian data is limited to convex problems that can be nonsmooth, or smooth nonconvex problems (Alacaoglu & Lyu, 2023; Doan, 2022; Doan et al., 2020; Dorfman & Levy, 2022; Duchi et al., 2012; Karimi et al., 2019; Li et al., 2023; Nagaraj et al., 2020; Sun et al., 2018). To the best of our knowledge, our MCMC-ProxAdam is the first stochastic approximation algorithm with solid convergence to solve the nonsmooth nonconvex problem (1) by employing MCMC sampling to obtain the gradient estimator of $\nabla f(\theta)$ in the form of (3). In particular, Dorfman & Levy (2022) and Li et al. (2023) analyze the convergence of their algorithms by establishing upper bounds on the bias and variance of the mini-batch gradient estimator. Specifically, Dorfman & Levy (2022) derives upper bounds via a Hoeffding-type inequality but under the condition of bounded stochastic gradient—an assumption that may not hold in practical scenarios. A pioneering work Li et al. (2023) relaxes the conventional bounded gradient requirement by considering a class of sub-exponential functions. They derive upper bounds via a Bernstein-type inequality for the random variable to establish the subsequent convergence results. However, in this literature the stochastic gradient $G(\theta; z)$ must satisfy $\mathbb{E}_{z\sim\mu_\theta}[G(\theta; z)] = 0$ for any $\theta \in \Theta$ and $\mu_\theta(z)$ must be the form of (4), alongside a uniform boundedness requirement on $\nabla_\theta\phi_\theta(z)$ for any $z \in \mathcal{X}$ and $\theta \in \Theta$. Furthermore, the second-order moment of $\nabla^2_\theta\phi_\theta(x)$ must remain bounded in expectation with respect to the stationary distribution μ_θ for any $\theta \in \Theta$. In Atchadé et al. (2017), perturbed proximal gradient (PPG) algorithm is proposed which combines a stochastic proximal gradient algorithm with the MCMC method. To analyze Markov noise, they employ a uniform-in-θ ergodic assumption, but this analysis is restricted to nonsmooth convex problems. We compare our work with previous literature in Table 1.

Stochastic gradient descent (SGD) (Robbins & Monro, 1951) is fundamental to large-scale optimization, yet controlling its gradient variance through hand-tuned learning rate schedules remains challenging. This has motivated the development of algorithms with adaptive learning rate. AdaGrad (Duchi et al., 2011) pioneers adaptive scaling using accumulated squared gradients, yet its

Table 1: Comparison with existing literature on stochastic algorithms with Markovian data

Refs.	Algorithms	Setting	Batch Type	Convergence rate	Assumptions
Problem type I: $\min_\theta f(\theta) = \mathbb{E}_{z\sim\mathbb{P}}[F(\theta; z)]$					
Duchi et al. (2012)	EMD	convex	single	$\mathcal{O}(\tau_{\text{mix}}/\sqrt{T})$	1, 2, 3
Sun et al. (2018)	MCGD	convex/smooth nonconvex	single	$\mathcal{O}(1/\sqrt{T})$	2, 5, 6, 7, 8
Doan (2022)	MSGD	strong convex/convex/smooth nonconvex	single	$\mathcal{O}(\log T/T), \mathcal{O}(\log T/\sqrt{T})$	4, 8, 9, 10
Doan et al. (2020)	AMGD	strong convex/convex/smooth nonconvex	single	$\mathcal{O}(\log T/T), \mathcal{O}(\log T/\sqrt{T})$	4, 5, 7, 9
Alacaoglu & Lyu (2023)	MC-SFOM	smooth nonconvex	single	$\mathcal{O}(\log T/\sqrt{T})$	7, 8, 11, 12, 19
Karimi et al. (2019)	MCSA	smooth nonconvex	single	$\mathcal{O}(\log T/\sqrt{T}) + c_0$	22, 23, 24
Dorfman & Levy (2022)	MAG	convex/smooth nonconvex	mini-batch	$\mathcal{O}(\sqrt{\tau_{\text{mix}}/T})$	4, 5, 7, 11
Li et al. (2023)	MCMC-SGD	smooth nonconvex	mini-batch	$\mathcal{O}(\log T/\sqrt{nT})$	13, 14, 15,16, 17
Problem type II: $\min_\theta \Phi(\theta) = f(\theta) + h(\theta)$					
Atchadé et al. (2017)	PPG	f: smooth convex, h: nonsmooth convex	mini-batch	$\mathcal{O}(1/\sqrt{nT})$	4, 18, 20
ours	MCMC-ProxAdam	f: smooth nonconvex, h: nonsmooth convex	mini-batch	$\mathcal{O}(1/\sqrt{nT})$ (Corollary 2)	4, 16, 21

1. Finite single-step variance of $G(\theta; z) \in \partial_\theta F(\theta; z)$;
2. Lipschitz continuity of $F(\theta; z)$;
3. Uniformity or stochastic uniformity of the mixing time τ_{mix};
4. Lipschitz continuity of $\nabla f(\theta)$;
5. Compactness of the constraint set;
6. Boundedness of $F(\theta; z)$;
7. Boundedness of $\nabla_\theta F(\theta; z)$ or $G(\theta; z) \in \partial_\theta F(\theta; z)$;
8. Lipschitz continuity of $\nabla_\theta F(\theta; z)$;
9. Uniform geometric ergodicity of Markov process;
10. PL condition;
11. Boundedness of $f(\theta)$;
12. Boundedness of first moment of stochastic gradient $G(\theta; z)$;
13. $\mathbb{E}_{z\sim\mathbb{P}}[\nabla_\theta F(\theta; z)] = 0$;
14. Target distribution satisfies (4);
15. Sub-exponential norm boundedness of $F(\theta; z)$ for any $\theta \in \Theta$;
16. A uniform lower bound of spectral gap;
17. Uniform boundedness of $\nabla_\theta \phi_\theta(z)$ for any $\theta \in \Theta$ and $z \in \mathcal{X}$ and uniformly bounded second-order moment of $\nabla^2_\theta \phi_\theta(x)$ under the stationary distribution for any $\theta \in \Theta$;
18. Uniform-in-θ ergodic of Markov kernel;
19. Mixing condition on Markov chain;
20. Regularity conditions on Markov kernels and step sizes;
21. Sub-exponential norm boundedness of stochastic gradient $G(\theta; z)$ for any $\theta \in \Theta$;
22. Stability and descent properties of the Lyapunov function;
23. Existence and Lipschitz continuity of the Poisson solution;
24. Uniform boundedness of drift term $\|\nabla_\theta F(\theta; z) - \nabla f(\theta)\|$ for any $\theta \in \Theta$ and $z \in \mathcal{X}$.

performance often stagnates in dense settings due to the rapid, monotonic decay of the learning rate. To address this, RMSProp (Tieleman & Hinton, 2012) replaces the simple accumulation with an exponential moving average, effectively circumventing the monotonic step-size decay inherent in AdaGrad. By further incorporating first-order momentum, Adam (Kingma & Ba, 2015) is proposed with convergence guarantees for online convex optimization. However, Reddi et al. (2018) later identifies a flaw in the original proof and constructs a counterexample where Adam fails to converge. This theoretical gap has catalyzed the development of various convergent Adam variants (Chen et al., 2019a; De et al., 2018; Guo et al., 2021; Zaheer et al., 2018; Zhang et al., 2022; Zou et al., 2019). All the aforementioned literature fail to address nonsmooth composite optimization. From another perspective, AdaGrad, RMSProp, and Adam use historical gradients to build adaptive diagonal preconditioners that mitigate gradient anisotropy and optimize dimension-wise step sizes, yet they differ significantly in their handling of gradient information. In Yun et al. (2021), a unified framework for stochastic proximal gradient methods is introduced that supports arbitrary positive preconditioners and nonsmooth, possibly nonconvex, regularizers. This framework is also the first proximal variant of Adam. To the best of our knowledge, no convergence analysis of Adam-type algorithms exists for optimization with Markovian data as considered in this paper.

In this work, by assuming a bounded sub-exponential norm, our theoretical framework accommodates unbounded stochastic gradients—a more practical scenario than the bounded assumptions in Dorfman & Levy (2022). Moreover, we newly develop a Bernstein-type inequality for random vectors that is different from the Bernstein-type inequality for random variable in Li et al. (2023), which can make us remove the assumptions 13, 14, 17 of Table 1 used in Li et al. (2023) that are not easy to satisfy in real problems. Compared with Yun et al. (2021), our algorithm applies not only to the setting where the gradient estimator is unbiased and bounded, but also to the more general case where gradient estimate is obtained via MCMC sampling, potentially biased and unbounded, with Markovian data. Going beyond the scope of (Alacaoglu & Lyu, 2023; Doan, 2022; Doan et al., 2020; Dorfman & Levy, 2022; Duchi et al., 2012; Karimi et al., 2019; Li et al., 2023; Nagaraj et al., 2020; Sun et al., 2018), whose analyses are restricted to smooth nonconvex settings, our algorithm accommodates a broader class of nonsmooth nonconvex problems. In contrast to PPG algorithm proposed by Atchadé et al. (2017), which integrates MCMC sampling with stochastic proximal gradient framework to solve nonsmooth convex problems, our MCMC-ProxAdam extends this capability to the more challenging nonsmooth nonconvex regime. Furthermore, by incorporating the Adam mechanism, our algorithm achieves faster and more stable convergence when solving large-scale optimization problems in deep learning.

In this paper, we propose an algorithm, MCMC-ProxAdam, which combines the stochastic proximal gradient method with mini-batching, Adam and MCMC sampling to solve the constrained nonsmooth

nonconvex composite optimization problem (1) in scenarios where $\nabla f(\theta)$ is in the form of (3) and must be approximated by the MCMC method. Our main contributions can be summarized as follows:

- **Weaker assumptions for analyzing the error of MCMC estimators:** Building on the Bernstein-type inequality for sums of bounded random matrices on stationary Markov chains in Neeman et al. (2024), we establish a Bernstein-type inequality for random vectors in non-stationary cases, where the boundedness assumption is relaxed to the sub-exponential norm boundedness condition. By leveraging this newly developed Bernstein-type inequality, we derive error bounds for the mini-batch gradient estimator which is fundamental to the convergence analysis.
- **Convergence guarantee of MCMC-ProxAdam:** MCMC-ProxAdam is the first Adam-type algorithm with solid convergence analysis that can solve nonsmooth nonconvex problems (1) with Markovian data. We give the solid convergence result in Theorem 2 that is for general batch size n_t, as well as the convergence rate $\mathcal{O}\left(1/\sqrt{nT}\right)$ with T iterations and a constant batch size n in Corollary 2 under a $4/3$-moment gradient-mapping criterion. The weaker $4/3$-moment stationarity measure is the price paid for simultaneously accommodating adaptive coordinate-wise preconditioning, Markovian dependence, and sub-exponential rather than bounded stochastic gradients.
- **Numerical experiments on EBMs:** We conduct numerical experiments on energy-based model (EBMs) generation tasks to evaluate the nonsmooth nonconvex model (1) and the proposed MCMC-ProxAdam algorithm. The experimental results confirm that our algorithm effectively handles the convex nonsmooth L_1 regularizer within a nonconvex framework, empirically supporting the theoretical convergence.

2 MCMC-ProxAdam Algorithm

The mathematical notations employed throughout this paper are listed in Table 3.

Algorithm 1 MCMC-ProxAdam: A Stochastic Proximal Adam Method via MCMC Sampling

1: **Input:** Stepsizes $\{\alpha_t\}_{t=0}^{T-1}$, momentum parameters $\{\beta_t, \rho_t\}_{t=0}^{T-1}$, and small constant $\epsilon > 0$.
2: **Initialize:** $\theta_0 \in \mathbb{R}^l$, $m_{-1} = (0, \ldots, 0)^\top \in \mathbb{R}^l$, and $v_{-1} = (0, \ldots, 0)^\top \in \mathbb{R}^l$.
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: Draw samples $S_t = \{z_{n_b+1}^{(t)}, \ldots, z_{n_b+n_t}^{(t)}\}$ by MCMC sampling after a burn-in period of n_b;
5: Compute the estimated gradient $g_t = \frac{1}{n_t} \sum_{j=n_b+1}^{n_b+n_t} G(\theta_t; z_j^{(t)})$;
6: $v_t = (1-\rho_t)v_{t-1} + \rho_t g_t^2$;
7: $m_t = (1-\beta_t)m_{t-1} + \beta_t g_t$;
8: $D_t = \text{Diag}(\sqrt{v_t} + \boldsymbol{\epsilon})$ with $\boldsymbol{\epsilon} = (\epsilon, \ldots, \epsilon)^\top \in \mathbb{R}^l$;
9: $\theta_{t+1} \in \underset{\theta \in \Theta}{\text{argmin}} \left\{ \langle m_t, \theta \rangle + h(\theta) + \frac{1}{2\alpha_t}(\theta - \theta_t)^\top D_t (\theta - \theta_t) \right\}$;
10: **end for**
11: **Output:** θ_R chosen from $\{\theta_t\}_{t=0}^{T-1}$ with $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1} \alpha_t}$.

Step 4 in Algorithm 1 represents a general MCMC sampling step, and the specific choice of the sampler depends on the target domain. For instance, Gibbs sampling can be used for discrete models such as Potts models (Atchadé et al., 2017). For energy-based models (EBMs) in Du & Mordatch (2019), we can adapt Langevin MCMC to perform sampling via:

$$\tilde{z}^k = \tilde{z}^{k-1} - \eta \nabla_z E_{\theta_t}(\tilde{z}^{k-1}) + \omega^k, \quad \omega^k \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}) \tag{5}$$

where $E_\theta(z)$ is a deep neural energy function and we let the above iterative procedure define a distribution $q_\theta(z)$ such that $\tilde{z}^k \sim q_\theta$. The noise variance σ_t^2 is the standard form of Langevin MCMC and is a parameter of the sampler rather than of the optimizer. Our analysis imposes no constraint on it: It enters the theoretical analysis solely through the transition kernel P_{θ_t} and hence only through spectral gap under (H1). Thus, σ_t can theoretically vary across iterations t without violating our

theoretical bounds Theorem 1; in practice, we can fix it following the setup in Du & Mordatch (2019). The update rule in Line 9 of Algorithm 1 can be written as the proximal projection:

$$\begin{aligned}\theta_{t+1} &\in \operatorname*{argmin}_{\theta\in\Theta}\left\{\langle m_t,\theta\rangle + h(\theta) + \frac{1}{2\alpha_t}(\theta-\theta_t)^\top D_t(\theta-\theta_t)\right\} \\ &= \operatorname{prox}^{D_t}_{\alpha_t h(\cdot)}\left(\theta_t - \alpha_t D_t^{-1} m_t\right)\end{aligned} \tag{6}$$

where the proximal projection $\operatorname{prox}^A_h(z) := \arg\min_{\theta\in\Theta}\left\{h(\theta) + \frac{1}{2}\|\theta - z\|^2_A\right\}$ is defined with respect to the weighted norm induced by the matrix A. Here, the positive definite matrix acts as a preconditioner, which geometrically realizes element-wise scaling of the parameter space. The generalized projection is defined as follows:

$$\mathcal{G}^D(\theta, g, \alpha) = \frac{1}{\alpha} D\left(\theta - \operatorname{prox}^D_{\alpha h(\cdot)}\left(\theta - \alpha D^{-1} g\right)\right). \tag{7}$$

3 Analysis of the MCMC Error

The core of our theoretical analysis is to establish error bounds for the mini-batch gradient estimator obtained via MCMC sampling under non-stationary Markov chains and potentially unbounded settings. The technical roadmap is outlined as follows: in Section 3.1, we start from a Bernstein-type inequality for sums of random matrices (8) of Lemma 1 on stationary Markov chains in Neeman et al. (2024) (see there for its proof), which is restricted to the maximum eigenvalue of square random matrices and fails to hold for other matrix norms or rectangular matrices. Then we broaden this inequality to the Frobenius norm of rectangular matrices on stationary Markov chains in Lemma 2 via Hermitian dilation (Definition A.2) and probabilistic techniques to accommodate random vector cases in our later analysis; in Section 3.2, by introducing the assumption of bounded sub-exponential norms for random vectors and leveraging the Bernstein-type inequality in Lemma 2, we derive a novel Bernstein-type inequality for the mini-batch gradient estimator on non-stationary Markov chains in Lemma 5. As our core technical pivot, Lemma 5 enables our framework to accommodate stochastic gradients that are biased and unbounded. Building upon Lemma 5, we further establish error bounds for the mini-batch gradient estimator in Theorem 1, which is fundamental to the convergence analysis of our MCMC-ProxAdam algorithm.

To maintain the continuity of discussion, we have deferred the fundamental definitions and background material on Markov chains and high-dimensional probability to Appendix A. We define $\ell_2(\mu\otimes\mathbf{1}) = \left\{\mathbf{f}: \mathcal{X}\to\mathbb{C}^{l^2} \,\middle|\, \mathbb{E}_\mu\left[\|\mathbf{f}(z)\|^2\right] < \infty\right\}$ for a measurable vector-valued function $\mathbf{f}: \mathcal{X}\to\mathbb{C}^{l^2}$. We assume that all random matrices and random vectors are $\ell_2(\mu\otimes\mathbf{1})$ measurable.

3.1 Matrix Bernstein-type Inequalities and Properties under Sub-exponential Norm

The following assumptions are required for Lemma 1:

- **(A1)** P is a discrete-time Markov chain on a continuous state space $\mathcal{X}$ with stationary distribution μ and absolute spectral gap $1-\lambda$ (Definition A.1). $z_1,\ldots,z_{\hat{n}}$ is a sequence of states driven by P with initial distribution $\nu = \mu$.
- **(A2)** For $j\in[\hat{n}]$, the random matrix $\mathcal{F}_j(z)\in\mathbb{R}^{l\times l}$ satisfies $\mathbb{E}_\mu[\mathcal{F}_j(z)] = 0$ and there exists a constant $V_j > 0$ such that $\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)^2\right]\right\|_{\text{op}} \le V_j$ for any $z\in\mathcal{X}$, where $\mathcal{F}_j(z)^2$ refers to matrix multiplication. Moreover, there exists a constant $M>0$ such that $\|\mathcal{F}_j(z)\|_{\text{op}} \le M$ for any $j\in[\hat{n}]$ and $z\in\mathcal{X}$.

Lemma 1. *Assume (A1)-(A2) hold. Let* $\sigma^2 = \sum_{j=1}^{\hat{n}} V_j$, *then for any* $0<\rho<\log(1-\lambda)/M$,

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{\hat{n}}\exp\left(\frac{\rho}{2}\mathcal{F}_j(z_j)\right)\right\|^2_{\text{F}}\right] \le l\exp\left(\frac{\sigma^2}{M^2}\left(e^{\rho M}-\rho M-1\right) + \frac{\sigma^2}{M^2}\cdot\frac{\lambda\left(e^{\rho M}-1\right)^2}{1-\lambda e^{\rho M}}\right).$$

Furthermore, for any $t > 0$,

$$\mathbb{P}_\mu\left(\lambda_{\max}\left(\sum_{j=1}^{\hat{n}}\mathcal{F}_j(z_j)\right)\geq t\right)\leq l^{2-\pi/4}\exp\left(\frac{-\pi^2t^2}{32\alpha(\lambda)\sigma^2+32\beta(\lambda)Mt}\right), \tag{8}$$

where $\beta(\lambda)$ is $\frac{4}{3\pi}$ if $\lambda = 0$ and $\frac{8/\pi}{1-\lambda}$ if $0 < \lambda < 1$, $\alpha(\lambda) = \frac{1+\lambda}{1-\lambda}$, and λ is defined in Definition A.1.

To facilitate the generalization from Lemma 1 to Lemma 2, we impose the following assumptions:

(A3) For $j \in [\hat{n}]$, the random matrix $\mathcal{F}_j(z) \in \mathbb{R}^{l_1\times l_2}$ satisfies $\mathbb{E}_\mu[\mathcal{F}_j(z)] = 0$ and there exist constants $W_j > 0$ such that for any $z \in \mathcal{X}$,

$$\max\left\{\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\mathrm{F}},\ \left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\mathrm{F}}\right\}\leq W_j,$$

where $\mathcal{F}_j^\top(z) = (\mathcal{F}_j(z))^\top$ denotes the matrix transpose of $\mathcal{F}_j(z)$. Moreover, there exists a constant $M' > 0$ such that $\|\mathcal{F}_j(z)\|_{\mathrm{F}} \leq M'$ for $j \in [\hat{n}]$ and any $z \in \mathcal{X}$.

Lemma 2. *Assume (A1) and (A3) hold. Let $\tau^2 = \sum_{j=1}^{\hat{n}} W_j$. Then the following tail bound holds:*

$$\mathbb{P}_\mu\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}}\mathcal{F}_j(z_j)\right\|_{\mathrm{F}}\geq s\right)\leq (l')^{2-\pi/4}\exp\left(\frac{-\gamma\pi^3\hat{n}^2s^2}{32\pi l'\tau^2+128\sqrt{2l'}M'\hat{n}s}\right),\quad \forall s>0, \tag{9}$$

where $l' = l_1 + l_2$, M' is given in (A3) and $\gamma = 1 - \lambda$.

To accommodate biased and unbounded stochastic gradients, we make the following assumptions. (E2) imposes a sub-exponential norm (Definition A.4 and Definition A.5) bound on the random vector. This provides a quantitative control on the tail behavior of the random vector, and it is strictly weaker than the requirement of boundedness. The following Lemma 3 shows that the bounded sub-exponential norm of a random vector implies a variance bound, and Lemma 4 establishes a tail probability inequality for $\mathcal{F}_j$. These provide the necessary foundation for deriving Lemma 5.

(E1) P is a discrete-time Markov chain on a continuous state space $\mathcal{X}$ with stationary distribution μ and absolute spectral gap $1 - \lambda$. $z_1, \ldots, z_{\hat{n}}$ is a sequence of states driven by P with initial distribution ν.

(E2) For $j \in [\hat{n}]$, let $\mathcal{F}_j : \mathcal{X} \to \mathbb{R}^l$ be a random vector satisfying $\|\mathcal{F}_j\|_{\psi_1(\mu;\mathcal{X})} \leq B$ for a constant $B > 0$.

Lemma 3. *Let $\mathcal{F}_j : \mathcal{X} \to \mathbb{R}^l$, $j = 1, \ldots, \hat{n}$, satisfy assumption (E2) . Then*

$$\max\left\{\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\mathrm{F}},\left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\mathrm{F}}\right\}\leq 4lC_0^2B^2,$$

where C_0 is a constant.

Lemma 4. *Let $\mathcal{F}_j : \mathcal{X} \to \mathbb{R}^l$, $j = 1, \ldots, \hat{n}$, satisfy assumption (E2). Then for any $t > 0$ and given ε-net, $\varepsilon \in (0, 1)$,*

$$\mathbb{P}_\mu\left(\|\mathcal{F}_j(z)\|_2 > t\right) = \mathbb{P}_\mu\left(\sup_{v\in\mathbb{S}^{l-1}} v^\top\mathcal{F}_j(z) > t\right)\leq 2\left(\frac{2}{\varepsilon}+1\right)^l\exp\left(-\frac{(1-\varepsilon)t}{B}\right).$$

3.2 Bernstein-type Inequality and Error Analysis for MCMC Gradient Estimator

The derivation of the Bernstein-type inequality for the mini-batch gradient estimator in Lemma 5 requires only (H1) and (H2). These assumptions characterize the stochastic gradient $G(\theta; z)$ and originate from (E1) and (E2). According to Definition A.5, $\|G_\theta\|_{\psi_1(\mu_\theta;\mathcal{X})} = \sup_{v\in\mathbb{S}^{l-1}}\|\langle G(\theta; z), v\rangle\|_{\psi_1(\mu_\theta;\mathcal{X})}$.

(H1) Let P_θ be the Markov operator induced by the MCMC algorithm with absolute spectral gap $\gamma(P_\theta) = 1 - \lambda(P_\theta)$. There is a uniform lower bound of absolute spectral gaps $\gamma^\star := \inf_{\theta\in\Theta}\gamma(P_\theta) > 0$. $S = \{z_j\}_{n_b+1}^{n_b+\hat{n}}$ is the set of samples driven by P_θ with initial distribution ν.

(H2) There exists a constant $B > 0$ such that $\|G_\theta\|_{\psi_1(\mu_\theta;\mathcal{X})} \le B$ for any $\theta \in \Theta$.

(H1) is difficult to verify in general, particularly for a complex deep energy network. That said, an absolute spectral gap is often assumed in the convergence analysis of Markov chains (Jiang et al., 2018; Neeman et al., 2024). In practice it can be enforced by the Metropolis step and clamping the iterates to a compact box.

Lemma 5. *Assume assumptions (H1)-(H2) are satisfied. Suppose the initial distribution ν is absolutely continuous with respect to the stationary distribution μ_θ, and its RN derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu_\theta} \in \ell_2(\mu_\theta)$. Furthermore, let $\chi = \left\| \frac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1 \right\|_\mu < +\infty$. Then for given ε-net, $\varepsilon \in (0,1)$ and $s \ge s_0$, the gradient estimator $g(\theta; S) = \frac{1}{\hat{n}} \sum_{j=n_b+1}^{n_b+\hat{n}} G(\theta; z_j)$ has the following tail bound:*

$$\begin{aligned} \mathbb{P}_\nu \left(\|g(\theta;S) - \nabla f(\theta)\|_2 \ge s \right) \le C(l+1)^{2-\pi/4} \exp\left(\frac{-\gamma^\star \hat{n} \pi^2 s^2}{2048(l^2+l)C_0^2 B^2} \right) \\ + 2C(l+1)^{2-\pi/4} \exp\left(\frac{-\pi^3 \sqrt{\gamma^\star \hat{n} s}}{1024 C_1 \sqrt{2B(l+1)}} \right), \end{aligned} \tag{10}$$

where $s_0 = \frac{(\log N_3)^2}{(N_1 - N_2)^2 \gamma^\star \hat{n}}$, $N_1 = \frac{(1-\varepsilon)C_1}{\sqrt{B}}$, $N_2 = \frac{\pi^3}{512 C_1 \sqrt{2B(l+1)}}$, $N_3 = \frac{2\hat{n}(\frac{2}{\varepsilon}+1)^l}{(l+1)^{2-\frac{\pi}{4}}}$, $C_1 = \frac{\pi^{\frac{3}{2}}}{(2l+2)^{\frac{1}{4}}(256(1-\varepsilon))^{\frac{1}{2}}}$, $C = (1+\chi^2)^{\frac{1}{2}}$, *$C_0$ is a constant, and $\gamma^\star$ is defined in assumption (H1).*

Theorem 1. *Suppose that all the conditions in Lemma 5 are satisfied. Then the gradient estimator $g(\theta; S) = \frac{1}{\hat{n}} \sum_{j=n_b+1}^{n_b+\hat{n}} G(\theta; z_j)$ has the following error bounds:*

$$\mathbb{E}_\nu \left[\|g(\theta;S) - \nabla f(\theta)\|_2 \right] \le \mathcal{O}\left(\frac{\sigma}{\sqrt{\gamma^\star \hat{n}}} \right), \quad \mathbb{E}_\nu \left[\|g(\theta;S) - \nabla f(\theta)\|_2^2 \right] \le \mathcal{O}\left(\frac{\sigma^2}{\gamma^\star \hat{n}} \right), \tag{11}$$

where $C = (1+\chi^2)^{\frac{1}{2}}$, $\sigma = \sqrt{\log(C(l+1)^{2-\frac{\pi}{4}})(l^2+l)B^2}$.

4 Convergence Analysis

Theorem 2. *Suppose that all the conditions in Lemma 5 are satisfied. Let $\alpha_t \le \frac{\epsilon}{2L}$, $\beta_{t+1} = \frac{\alpha_t}{c\epsilon}$ for all $t = 0, \ldots, T-1$, and an arbitrarily $c \in (0, \frac{1}{4L}]$. Then it holds that*

$$\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}} \right] \right)^{\frac{3}{2}} \le \left(\sqrt{B'} + \epsilon \right) \frac{D + 4c\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_0}\right) + 8\sum_{t=0}^{T-1} \frac{\alpha_t}{\epsilon} \mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_{t+1}}\right)}{\sum_{t=0}^{T-1} \alpha_t}, \tag{12}$$

where $D = 8\left(\Phi(\theta_0) - \Phi^{\mathrm{opt}}\right) + 8cB'$ with $\Phi, \Phi^{\mathrm{opt}}$ being defined in (1), *$B' = \frac{8\sqrt{2}CC'^2 B^2 (\frac{2}{\varepsilon}+1)^{\frac{l}{2}}}{(1-\varepsilon)^2}$, C' is a constant, and R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1} \alpha_t}$, for $t = 0, \ldots, T-1$.*

Theorem 2 presents the main convergence result for our MCMC-ProxAdam algorithm, specifically designed for nonsmooth nonconvex optimization involving biased gradient estimator derived from Markovian samples. Theorem 4 in Zou et al. (2019) provides a similar result for the convergence of Adam for smooth nonconvex problems under the assumption of unbiased gradient estimator. It is worth mentioning that our algorithmic framework and the underlying proof techniques are totally different from the approach in Zou et al. (2019). Moreover, the convergence in expectation characterized by Theorem 2 is on the term $\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}} \right] \right)^{\frac{3}{2}}$, which is slightly weaker than the standard metric $\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^2 \right]$. The weakening is a cost of using a coordinate-wise adaptive learning rate, relaxing the bounded stochastic gradient assumption, and considering Markovian data in this paper. As explained in detail in Remark 6 (ii) of Zou et al. (2019), a coordinate-wise adaptive learning rate and unbounded stochastic gradient under unbiased gradient and bounded variance assumptions make it technically difficult to obtain the results for the convergence rate using the standard metric. Here we in addition consider Markovian data with biased gradient. Recovering

the classical $\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^2\right]$ metric would require the stronger bounded-gradient condition that our analysis is designed to avoid. Under this weaker metric, we also establish a high-probability convergence bound for $\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^2$ with an explicit polynomial dependence on δ; see Corollary 4 for details.

Corollary 1. *Suppose that all the conditions in Lemma 5 are satisfied. Let $\alpha_t = \frac{\epsilon}{2L}$, $\beta_{t+1} = \frac{\alpha_t}{c\epsilon}$ for all $t = 0, \ldots, T-1$, and an arbitrarily $c \in (0, \frac{1}{4L}]$. Let the batch size $n_t \equiv n$. Then we have*

$$\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \leq \left(\sqrt{B'} + \epsilon\right)\left(\frac{2LD}{\epsilon T} + \frac{8cL}{\epsilon T}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n}\right) + \frac{8}{\epsilon}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n}\right)\right),$$

where D and B' are defined as those in Theorem 2, and R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1}\alpha_t}$, for $t = 0, \ldots, T-1$.

We note that the bound in Corollary 1 depends on the batch size n. The third term in this bound will dominate the first and second ones for sufficiently large T. Thus, the batch size n must be carefully balanced against the total iteration number T, which eventually depends on the total computational budget given by the user. Corollary 2 presents a reasonable choice of n tailored to the total number of calls to the $\mathcal{SFO}$.

Corollary 2. *Suppose that all the conditions in Lemma 5 are satisfied. Given a fixed total number of calls $\bar{N}$ to the $\mathcal{SFO}$, if $\alpha_t = \frac{\epsilon}{2L}$, $\beta_{t+1} = \frac{\alpha_t}{c\epsilon}$ for all $t = 0, \ldots, T-1$, and an arbitrarily $c \in (0, \frac{1}{4L}]$. Let the batch size $n = \mathcal{O}\left(\bar{N}^{\frac{1}{2}}\right)$. Then we have*

$$\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \leq \mathcal{O}\left(\frac{D_1}{\bar{N}}\right) + \mathcal{O}\left(\frac{D_2}{\sqrt{\bar{N}}}\right),$$

where $D_1 = \frac{16cL\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\gamma^\star}$, $D_2 = \frac{4L\left(\sqrt{B'}+\epsilon\right)D}{\epsilon} + \frac{8\sigma^2\left(\sqrt{B'}+\epsilon\right)}{\epsilon\gamma^\star}$ with D and B' being defined as those in Theorem 2, and R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1}\alpha_t}$, for $t = 0, \ldots, T-1$.

Corollary 3. *Suppose that all the conditions in Lemma 5 are satisfied. Let $\alpha_t = \frac{\hat{\alpha}\sqrt{n}}{\sqrt{t+1}}$, $\beta_{t+1} = \frac{\alpha_t}{c\epsilon}$ for all $t = 0, \ldots, T-1$, and an arbitrarily $c \in (0, \frac{1}{4L}]$. Given a fixed total number of calls $\bar{N}$ to the $\mathcal{SFO}$, if the batch size $n = \mathcal{O}\left(\bar{N}^{\frac{1}{2}}\right)$, then we have*

$$\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \leq \mathcal{O}\left(\frac{D_3}{\bar{N}}\right) + \mathcal{O}\left(\frac{D_4}{\sqrt{\bar{N}}}\right),$$

where $D_3 = \frac{\left(\sqrt{B'}+\epsilon\right)c\sigma^2}{\gamma^\star}$, $D_4 = \frac{\left(\sqrt{B'}+\epsilon\right)\left(D\epsilon\gamma^\star+\sigma^2\right)}{\epsilon\gamma^\star}$ with D and B' being defined as those in Theorem 2, and R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1}\alpha_t}$, for $t = 0, \ldots, T-1$.

Corollary 4. *Suppose that all the conditions in Corollary 2 are satisfied. Given a fixed total number of calls $\bar{N}$ to the $\mathcal{SFO}$, for any $\delta > 0$, the following bound holds with probability at least $1 - \delta^{\frac{2}{3}}$:*

$$\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^2 \leq \mathcal{O}\left(\frac{D_1}{\bar{N}\delta}\right) + \mathcal{O}\left(\frac{D_2}{\sqrt{\bar{N}}\delta}\right),$$

where D_1, D_2 are defined as those in Corollary 2, and R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1}\alpha_t}$, for $t = 0, \ldots, T-1$.

5 Numerical Experiments

We evaluate MCMC-ProxAdam on our proposed nonconvex Energy-Based Models (EBMs) with non-smooth L_1 regularization in the context of generation tasks. For a data point z, EBMs characterize the data distribution through a deep neural energy function $E_\theta(z)$, with the density given by $p_\theta(z) = \frac{\exp(-E_\theta(z))}{Z_\theta}$, where $Z_\theta = \int \exp(-E_\theta(z))\mathrm{d}z$ denotes the partition function. Then we can fit $p_\theta(z)$ to

Table 2: FID results of different optimizers across three datasets.

Method	MNIST	CIFAR-10	CelebA
MCMC-ProxAdam	25.99 ± 1.85	50.07 ± 4.36	22.86 ± 2.59
MCMC-Adam	27.08 ± 2.92	53.21 ± 6.43	25.61 ± 3.73
MCMC-SubAdam	28.16 ± 2.10	55.51 ± 8.60	23.97 ± 5.33

$p_{\text{data}}(z)$ by maximizing the expected log-likelihood $\mathbb{E}_{z\sim p_{\text{data}}(z)}\left[\log p_\theta(z)\right]$, which is equivalent to as follows.

$$\min_\theta \ell(\theta) := \mathbb{E}_{z\sim p_{\text{data}}(z)}\left[-\log p_\theta(z)\right] = \mathbb{E}_{z\sim p_{\text{data}}(z)}\left[E_\theta(z) + \log Z(\theta)\right]. \tag{13}$$

We also proposed EBMs with L_1 regularization for θ, leading to the nonsmooth nonconvex model

$$\min_\theta L(\theta) := \ell(\theta) + \lambda_{\text{reg}}\|\theta\|_1. \tag{14}$$

The gradient of $\ell(\theta)$ is given by $\nabla\ell(\theta) = \mathbb{E}_{z^+\sim p_{\text{data}}}\left[\nabla_\theta E_\theta(z^+)\right] - \mathbb{E}_{z^-\sim p_\theta}\left[\nabla_\theta E_\theta(z^-)\right]$ (Du & Mordatch, 2019). While the first term is estimated over mini-batches of real data, the intractable partition function Z_θ precludes direct sampling from $p_\theta(z)$; to address this, we adapt Langevin MCMC to perform sampling via (5).

We train EBMs on MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015). For the non-regularized model (13), we use MCMC-Adam (MCMC-ProxAdam without L_1 regularization). For the L_1-regularized model (14), we employ MCMC-ProxAdam and MCMC-SubAdam (a subgradient variant of MCMC-Adam). We need to point that all three algorithms require MCMC sampling for gradient estimation. Theoretically, when the L_1 regularization is not considered, Theorem 2 includes the convergence guarantee of MCMC-Adam as a special case. However, no existing work has established the convergence of MCMC-SubAdam under Markovian data. We compute the Fréchet Inception Distance (FID) (Heusel et al., 2017) between 60,000 generated images and 60,000 real images (lower is better). Each optimizer is run five times independently, with the mean and standard deviation of FID scores reported in Table 2; across all three datasets, MCMC-ProxAdam achieves the lowest mean FID and the smallest standard deviation. Generated samples are shown in Fig. 1.

We also include PPG (Atchadé et al., 2017) and MCMC-SGD (Li et al., 2023) as baselines; both yield poor performance even on MNIST (detailed in Table 5 of Appendix D.3). Furthermore, we conduct ablation studies on MCMC-ProxAdam, MCMC-Adam and MCMC-SubAdam on the MNIST dataset, specifically targeting the momentum parameter β_t and the proximal operator, while comparing fixed-value and dynamic parameter settings (detailed in Appendix D.4). Results in Table 6 and Table 7 consistently demonstrate that dynamic parameter schedules achieve faster convergence compared to their fixed-value counterparts. Finally, we conduct a sensitivity analysis of MCMC-ProxAdam with respect to the burn-in n_b and the noise parameter σ_t (detailed in Appendix D.5). As summarized in Table 8 and Table 9, while larger σ_t values yield a faster initial descent, a moderate value (e.g., $\sigma_t = 0.005$) optimally balances convergence speed and the final FID. Overall, our algorithm exhibits robust convergence and stability across different parameter configurations as the number of epochs increases. These results confirm that our algorithm could solve nonconvex optimization problems with a convex nonsmooth term, supporting the theoretical convergence.

6 Conclusion

We provide a rigorous convergence analysis of the proposed MCMC-ProxAdam, a framework that integrates the proximal operator, mini-batching, and the Adam mechanism to solve nonsmooth nonconvex problems where the gradient of the smooth part can only be estimated via MCMC sampling. This results in a biased gradient estimator that may further exhibit unbounded variance. We establish error bounds for the gradient estimator under relaxed assumptions compared with existing literature. We prove convergence of the proposed method and derive a convergence rate of $\mathcal{O}(1/\sqrt{nT})$ under a $4/3$-moment gradient-mapping criterion. Numerical experiments on energy-based models (EBMs) for image generation across the MNIST, CIFAR-10, and CelebA datasets demonstrate its capability in solving nonconvex composite optimization problems with a convex nonsmooth term, which is consistent with the theoretical convergence results developed in this paper.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for language editing and proofreading. We have not used generative AI tools for experimental data generation and core mathematical proofs, and other disclosure tasks are not applicable to this work. Additionally, we have reviewed all AI-assisted work by manually checking the revised text to ensure accuracy, clarity, and faithfulness to our original manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work presents fundamental research in mathematical optimization and machine learning algorithms. All datasets used in our experiments (e.g., MNIST, CIFAR-10, and CelebA) are standard, publicly available benchmark datasets widely used in the scientific community. To the best of our knowledge, this research does not involve human subjects directly, nor does it present direct privacy, security, or immediate negative societal impacts.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

To ensure the reproducibility of our theoretical and empirical results, we have provided details across the paper and supplementary materials. Specifically, all detailed mathematical assumptions and complete proofs for our main theoretical results are provided in Appendix A, Appendix B and Appendix C. The complete experimental setup, including hyperparameter configurations, and implementation details, is described in Appendix D. Furthermore, the complete source code along with instructions for reproducing experimental results has been included in the supplementary materials.

REFERENCES

Ahmet Alacaoglu and Hanbaek Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *International Conference on Machine Learning*, pp. 458–489. PMLR, 2023.

Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pp. 89–97. PMLR, 2017.

Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019a.

Zaiyi Chen, Yi Xu, Haoyuan Hu, and Tianbao Yang. Katalyst: Boosting convex Katyusha for non-convex problems with a large condition number. In *International Conference on Machine Learning*, pp. 1102–1111. PMLR, 2019b.

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

Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Kuangyu Ding, Jingyang Li, and Kim-Chuan Toh. Nonconvex stochastic Bregman proximal gradient method with application to deep learning. *Journal of Machine Learning Research*, 26(39):1–44, 2025.

Thinh T Doan. Finite-time analysis of Markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2022.

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

Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In *International Conference on Machine Learning*, pp. 5429–5446. PMLR, 2022.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

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

Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E-Statistical, Nonlinear, and Soft Matter Physics*, 87(1):012707, 2013.

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

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. In *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2021.

W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequalities for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tianyou Li, Fan Chen, Huajie Chen, and Zaiwen Wen. Convergence analysis of stochastic gradient descent with MCMC estimators. *arXiv preprint arXiv:2303.10599*, 2023.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170, 1997.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of Markov-dependent random matrices. *Information and Inference: A Journal of the IMA*, 13(4):iaae032, 2024.

Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst for gradient-based nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 613–622. PMLR, 2018.

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

Sashank J. Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J. Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Daniel Rudolf. Explicit error bounds for Markov chain Monte Carlo. *arXiv preprint arXiv:1108.3201*, 2011.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 681–688, 2011.

Jihun Yun, Aurélie C Lozano, and Eunho Yang. Adaptive proximal gradient methods for structured neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24365–24378, 2021.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135, 2019.

Part I

Appendix

Table of Contents

A Definitions **15**
- A.1 Markov chains . . . 15
- A.2 Sub-exponential norm . . . 16

B Proof of Results in Section 3 **17**
- B.1 Proof of Lemma 2 . . . 17
- B.2 Proof of Lemma 3 . . . 18
- B.3 Proof of Lemma 4 . . . 19
- B.4 Proof of Lemma 5 . . . 20
- B.5 Proof of Theorem 1 . . . 22

C Proof of Results in Section 4 **24**
- C.1 Key Lemmas in Section 4 . . . 24
- C.2 Proof of Theorem 2 . . . 28
- C.3 Proof of Corollary 2 . . . 31
- C.4 Proof of Corollary 3 . . . 31
- C.5 Proof of Corollary 4 . . . 32

D Additional Experimental Details and Results **32**
- D.1 Experimental Setup and Hyperparameters of Table 2 . . . 32
- D.2 Explanation for Selecting the Regularization Parameter λ_{reg} . . . 32
- D.3 Failures of PPG and MCMC-SGD as baselines . . . 33
- D.4 Ablation Experiments . . . 33
- D.5 Sensitivity Study . . . 34
- D.6 Generated Samples . . . 36

In the appendix, we present supplementary materials and detailed technical proofs for the main results. Specifically, Appendix A provides relevant definitions for Markov chains and the sub-exponential norm in high-dimensional probability to facilitate reading. Appendix B establishes the error bounds for mini-batch gradient estimator using MCMC sampling. In Appendix C, we present the comprehensive convergence analysis and stability results for the proposed algorithms. In Appendix D, we provide some additional experimental details and results.

Table 3: Notation used throughout the paper.

Symbol	Description
$\hat{n}$	An arbitrarily given positive integer, $\hat{n} \in \mathbb{Z}^+$.
n_t	The batch size at the t-th iteration.
n	The constant batch size when $n_t \equiv n$.
S_t	The set of samples generated from a Markov chain to approximate expectations with respect to μ_{θ_t}, following a burn-in period of n_b steps, $S_t = \{z_i\}_{i=n_b+1}^{n_b+n_t}$.
n_b	The number of burn-in steps of the Markov chain.
g_t	The mini-batch gradient estimator obtained via MCMC, $g_t = \frac{1}{n_t}\sum_{j=n_b+1}^{n_b+n_t} G(\theta_t; z_j)$.
α_t	The base learning rate.
m_t	The biased first moment estimate with parameter $\beta_t \geq 0$.
v_t	The biased second raw moment estimate with parameter $\rho_t \geq 0$.
ϵ	A small positive constant, $\epsilon > 0$.
ε	A fixed constant associated with the ε-net (Definition A.6), with $\varepsilon \in (0, 1)$.
$\|v_t\|_1$	The standard L_1 norm, $\|v_t\|_1 = \sum_{k=1}^{l} \|v_{t,k}\|$.
$\|v_t\|_2^2$	The squared standard L_2 norm, $\|v_t\|_2^2 = \langle v_t, v_t\rangle = \sum_{k=1}^{l} \|v_{t,k}\|^2$.
$\|v_t\|_{D_t}^2$	The weighted norm induced by D_t, $\|v_t\|_{D_t}^2 = \langle v_t, D_t v_t\rangle = \langle v_t, v_t\rangle_{D_t}$.
$\|A\|_{\mathrm{op}}$	The spectral norm of a matrix $A \in \mathbb{R}^{l_1 \times l_2}$, $\|A\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Ax\|_2$.
$\|A\|_F$	The Frobenius norm of a matrix $A \in \mathbb{R}^{l_1 \times l_2}$, $\|A\|_F = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$.
$\lambda_{\max}(B)$	The largest eigenvalue of a symmetric matrix $B \in \mathbb{R}^{l_3 \times l_3}$.

A Definitions

A.1 Markov chains

Let $(\mathcal{X}, \mathcal{B}, \mathbb{P})$ be a probability space. $\{Z_i\}_{i=1}^{\hat{n}}$ be a time-homogeneous Markov chain on a state space $\mathcal{X}$, governed by an initial distribution ν and a transition kernel P. That is, for any Borel set $A \in \mathcal{B}(\mathcal{X})$,

$$\nu(A) = \mathbb{P}(Z_1 \in A) \quad \text{and} \quad P(z, A) = \mathbb{P}(Z_k \in A \mid Z_{k-1} = z).$$

A distribution μ is stationary with respect to P if it satisfies the invariance condition:

$$\mu(A) = \int P(z, A)\, \mu(\mathrm{d}z) \quad \text{for all } A \in \mathcal{B}(\mathcal{X}).$$

The chain is said to be stationary if its initial distribution $\nu = \mu$. We denote the chi-squared divergence by $\chi^2 = \chi^2(\nu, \mu) := \|\frac{\mathrm{d}\nu}{\mathrm{d}\mu} - 1\|_\mu^2$ and let $C = \sqrt{1 + \chi^2}$.

Define $\ell_2(\mu) = \left\{f : \mathcal{X} \to \mathbb{C} \mid \mathbb{E}_\mu\left[f(z)^2\right] < \infty\right\}$ for $f : \mathcal{X} \to \mathbb{C}$ measurable. This is a Hilbert space equipped with inner product $\langle f, g\rangle_\mu = \int f(z)g(z)\mu(\mathrm{d}z), \forall f, g \in \ell_2(\mu)$ and the corresponding norm is $\|f\|_\mu = \sqrt{\langle f, f\rangle_\mu}$. The norm of a linear operator T is defined as $\|T\|_\mu = \sup_{\|f\|_\mu = 1} \|Tf\|_\mu$.

The transition kernel P acts as an operator on $\ell_2(\mu)$:

$$(Pf)(x) = \int f(z)P(x, \mathrm{d}z), \quad \forall x \in \mathcal{X}, f \in \ell_2(\mu).$$

The projection operator Π corresponding to the distribution μ is defined as $(\Pi f)(x) = \mathbb{E}_\mu[f]\mathbf{1}$, where $\mathbf{1}$ is the function satisfying $\mathbf{1}(x) = 1$ for all x and $\mathbb{E}_\mu[f] = \langle \mathbf{1}, f\rangle_\mu$. An important subset of elements in $\ell_2(\mu)$ is the class of 'mean-zero functions', denoted as

$$\ell_2^0(\mu) = \{f \in \ell_2(\mu) \mid \Pi f = 0\}.$$

Recall that $\ell_2(\mu \otimes \mathbf{1}) = \left\{\mathbf{f} : \mathcal{X} \to \mathbb{C}^{l^2} \,\middle|\, \mathbb{E}_\mu\left[\|\mathbf{f}(z)\|^2\right] < \infty\right\}$ for $\mathbf{f} : \mathcal{X} \to \mathbb{C}^{l^2}$ measurable as a vector-valued function. This space serves as the natural extension of the scalar-valued $\ell_2(\mu)$ to

matrix-valued functions. By identifying the space of $l \times l$ complex matrices $\mathbb{M}^{l\times l}$ with the tensor product space $\mathbb{C}^l \otimes \mathbb{C}^l \cong \mathbb{C}^{l^2}$, a matrix-valued function $\mathbf{F} : \mathcal{X} \to \mathbb{C}^{l\times l}$ can be represented as a vector-valued function $\mathbf{f} : \mathcal{X} \to \mathbb{C}^{l^2}$ via the vectorization map vec($\mathbf{F}$). $\mathbb{H}^{l_1+l_2}$ denotes the set of all $(l_1 + l_2) \times (l_1 + l_2)$ Hermitian matrices. For a matrix $B \in \mathbb{M}^{l_1\times l_2}$, B^* refers to the conjugate transpose of B.

Definition A.1. *(Absolute spectral gap) A Markov kernel P with stationary measure μ admits an absolute spectral gap $1 - \lambda(P)$ if*

$$\lambda(P) := \sup_{\substack{f\in\ell_2^0(\mu),\\ f\neq 0}} \frac{\|Pf\|_\mu}{\|f\|_\mu} = \|P - \Pi\|_\mu < 1.$$

The absolute spectral gap characterizes the convergence rate of the Markov chain toward its stationary distribution and ensures the Markov chain's ergodicity, and it is noteworthy that $\lambda(P) \leq 1$ always holds (Rudolf, 2011). We will use $\lambda = \lambda(P)$ for simplicity.

Definition A.2. *(Hermitian dilation) Hermitian dilation $\mathcal{H} : \mathbb{M}^{l_1\times l_2} \to \mathbb{H}^{l_1+l_2}$ is a mapping from a general matrix to a Hermitian matrix, defined as*

$$\mathcal{H}(B) = \begin{pmatrix} 0_{l_1\times l_1} & B \\ B^* & 0_{l_2\times l_2} \end{pmatrix},\ B \in \mathbb{M}^{l_1\times l_2}.$$

Property A.1. For a matrix $B \in \mathbb{R}^{l_1\times l_2}$, $\mathcal{H}(B) = \begin{pmatrix} 0_{l_1\times l_1} & B \\ B^\top & 0_{l_2\times l_2} \end{pmatrix} \in \mathbb{R}^{(l_1+l_2)\times(l_1+l_2)}$. Hermitian matrix $\mathcal{H}(B)$ has the following properties (Tropp et al., 2015):

(a) $\mathcal{H}(B)$ is a symmetric matrix;

(b) $\|\mathcal{H}(B)\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathcal{H}(B)^\top \mathcal{H}(B)\right) = 2\|B\|_{\mathrm{F}}^2$;

(c) $\|\mathcal{H}(B)\|_{\mathrm{op}} = \|B\|_{\mathrm{op}} \leq \|B\|_{\mathrm{F}}$.

Definition A.3. *(Random matrix variance) Let $\mathcal{F} : \mathcal{X} \to \mathbb{C}^{l_1\times l_2}$ be a random matrix and $\mathbb{E}[\mathcal{F}] = 0$, the matrix variance of $\mathcal{F}$ is* $\max\left\{\|\mathbb{E}[\mathcal{F}\mathcal{F}^*]\|_{\mathrm{op}},\ \|\mathbb{E}[\mathcal{F}^*\mathcal{F}]\|_{\mathrm{op}}\right\}$.

A.2 Sub-exponential norm

We explain some definitions in high-dimensional probability (Vershynin, 2018). The unit sphere in R^l is defined as $\mathbb{S}^{l-1} = \left\{v \in \mathbb{R}^l : \|v\| = 1\right\}$.

Definition A.4. *(Sub-exponential random variable) A random variable $Z : \mathcal{X} \to \mathbb{R}$ is called sub-exponential if it satisfies any of the following equivalent properties, where the parameters $K_i > 0$ differ by at most an absolute constant factor:*
(i) (Tail) There exists $K_1 > 0$ such that

$$\mathbb{P}(|Z| \geq t) \leq 2\exp\left(-\frac{t}{K_1}\right) \quad \textit{for all } t \geq 0.$$

(ii) (Moments) There exists $K_2 > 0$ such that

$$\|Z\|_{L^p} = \left(\mathbb{E}[|Z|^p]\right)^{1/p} \leq K_2 p \quad \textit{for all } p \geq 1.$$

(iii) (Moment Generating Function of $|Z|$) There exists $K_3 > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{|Z|}{K_3}\right)\right] \leq 2.$$

The sub-exponential norm of Z, denoted $\|Z\|_{\psi_1}$, is defined as the smallest parameter K_3 in property (iii), i.e.,

$$\|Z\|_{\psi_1} = \inf\left\{K > 0 : \mathbb{E}\left[\exp\left(\frac{|Z|}{K}\right)\right] \leq 2\right\}.$$

This norm provides a canonical measure of the sub-exponential decay rate of Z. Moreover, let $\tilde{f}: \mathbb{R} \to \mathbb{R}$ be a mapping, the sub-exponential norm of a measurable function $\tilde{f}(Z): \mathcal{X} \to \mathbb{R}$ is defined by

$$\|\tilde{f}\|_{\psi_1(\pi;\mathcal{X})} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{|\tilde{f}(Z)|}{t}\right)\right] \leq 2\right\}.$$

Definition A.5. *(Sub-exponential random vector) Let $q > 1$ be a positive integer. A random vector $Y: \mathcal{X} \to \mathbb{R}^q$ is called sub-exponential if the one-dimensional marginals $\langle Y, v\rangle$ are sub-exponential random variables for all $v \in \mathbb{R}^q$. The sub-exponential norm of Y is defined by taking the maximal sub-exponential norm of $\langle Y, v\rangle$ over all unit vectors v:*

$$\|Y\|_{\psi_1(\mu;\mathcal{X})} = \sup_{v \in \mathbb{S}^{q-1}} \|\langle Y, v\rangle\|_{\psi_1(\mu;\mathcal{X})}.$$

Moreover, let $\mathcal{F}: \mathbb{R}^q \to \mathbb{R}^l$ be a mapping, the sub-exponential norm of a measurable vector function $\mathcal{F}(Y): \mathcal{X} \to \mathbb{R}^l$ is defined by

$$\|\mathcal{F}\|_{\psi_1(\mu;\mathcal{X})} = \sup_{v \in \mathbb{S}^{l-1}} \|\langle \mathcal{F}(Y), v\rangle\|_{\psi_1(\mu;\mathcal{X})}.$$

Definition A.6. *(ε-net) Let (T, d) be a metric space. Consider a set $K \subset T$ and a number $\varepsilon > 0$. A subset $\mathcal{N} \subset K$ is called an ε-net of K if every point in K is within distance ε of some point of $\mathcal{N}$, i.e.*

$$\forall x \in K \ \exists x_0 \in \mathcal{N} : \ d(x, x_0) \leq \varepsilon.$$

Definition A.7. *(Covering numbers) The smallest cardinality of an ε-net of a set K is called the covering number of K and is denoted $\mathcal{N}(K, d, \varepsilon)$. Equivalently, $\mathcal{N}(K, d, \varepsilon)$ is the smallest number of closed balls with centers in K and radii ε whose union covers K.*

Definition A.8. *(Normal cone) Let $\Theta \subset \mathbb{R}^l$ be a convex set with $\bar{\theta} \in \Theta$. The normal cone to Θ at $\bar{\theta}$ is*

$$N(\bar{\theta}; \Theta) := \left\{v \in \mathbb{R}^l \mid \langle v, \theta - \bar{\theta}\rangle \leq 0 \text{ for all } \theta \in \Theta\right\}.$$

By convention, we put $N(\bar{\theta}; \Theta) := \emptyset$ for $\bar{\theta} \notin \Theta$.

Let $T = \mathbb{R}^l$ with the Euclidean metric $d(x, y) = \|x - y\|_2$ in Definition A.6 and Definition A.7. To ease the notation , we omit the metric and define the ε-net and the covering numbers of the unit Euclidean ball $\mathbb{S}^{l-1}$ as $\mathcal{N}$ and $\mathcal{N}\left(\mathbb{S}^{l-1}, \varepsilon\right)$, respectively.

B Proof of Results in Section 3

B.1 Proof of Lemma 2

In the proof Lemma 2, we apply the Hermitian dilation to an $l_1 \times l_2$ rectangular matrix, thereby embedding it into an $(l_1 + l_2) \times (l_1 + l_2)$ matrix. To apply Lemma 1 to this dilated matrix, we must verify that it satisfies assumption (A1)-(A2). While assumption (A1) is easy to verify, for assumption (A2), we show that it follows from assumption (A3).

Proof. We construct the dilated matrices

$$\widetilde{\mathcal{F}}_j(z) := \mathcal{H}\left(\mathcal{F}_j(z)\right) = \begin{pmatrix} 0_{l_1 \times l_1} & \mathcal{F}_j(z) \\ \mathcal{F}_j^\top(z) & 0_{l_2 \times l_2} \end{pmatrix} \in \mathbb{R}^{(l_1+l_2)\times(l_1+l_2)}$$

by applying the Hermitian dilation to each $\mathcal{F}_j(z) \in \mathbb{R}^{l_1 \times l_2}$, $j = 1, \ldots, \hat{n}$. Since $\mathbb{E}_\mu\left[\mathcal{F}_j(z)\right] = 0$, we have $\mathbb{E}_\mu\left[\widetilde{\mathcal{F}}_j(z)\right] = \mathcal{H}\left(\mathbb{E}_\mu\left[\mathcal{F}_j(z)\right]\right) = 0$. Property (c) of $\mathcal{H}$ in Property A.1 and assumption (A3) implies that $\left\|\widetilde{\mathcal{F}}_j(z)\right\|_{\text{op}} = \|\mathcal{F}_j(z)\|_{\text{op}} \leq \|\mathcal{F}_j(z)\|_{\text{F}} \leq M'$. For each $\mathcal{F}_j(z) \in \mathbb{R}^{l_1 \times l_2}$,

$$\widetilde{\mathcal{F}}_j(z)^2 = \begin{pmatrix} \mathcal{F}_j(z)\mathcal{F}_j^\top(z) & 0_{l_1 \times l_2} \\ 0_{l_2 \times l_1} & \mathcal{F}_j^\top(z)\mathcal{F}_j(z) \end{pmatrix},$$

and

$$\mathbb{E}_\mu\left[\widetilde{\mathcal{F}}_j(z)^2\right] = \begin{pmatrix} \mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right] & 0_{l_1 \times l_2} \\ 0_{l_2 \times l_1} & \mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right] \end{pmatrix}.$$

Then with assumption (A3) we can get

$$\left\|\mathbb{E}_\mu\left[\widetilde{\mathcal{F}}_j(z)^2\right]\right\|_{\text{op}} = \max\left\{\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\text{op}}, \left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\text{op}}\right\}$$
$$\leq \max\left\{\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\text{F}}, \left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\text{F}}\right\} \leq W_j.$$

Applying $\widetilde{\mathcal{F}}_j(z)$, $j = 1, \ldots, \hat{n}$, to (8) in Lemma 1 and $\lambda_{\max}\left(\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right) = \left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}}$, we have for any $t' > 0$,

$$\mathbb{P}_\mu\left(\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}} \geq t'\right) \leq (l_1+l_2)^{2-\pi/4} \exp\left(\frac{-\pi^2 t'^2}{32\alpha(\lambda)\tau^2 + 32\beta(\lambda)M't'}\right), \tag{15}$$

where $\tau^2 = \sum_{j=1}^{\hat{n}} W_j$. Since $\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{F}} \leq \sqrt{l_1+l_2}\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}}$, it follows that

$$\left\{\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{F}} \geq t\right\} \subseteq \left\{\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}} \geq \frac{t}{\sqrt{l_1+l_2}}\right\}.$$

To upper bound the probability of the event $\left\{\left\|\sum_{j=1}^{\hat{n}} \mathcal{F}_j(z_j)\right\|_{\text{F}} \geq t\right\}$, it suffices to consider the event $\left\{\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{F}} \geq \sqrt{2}t\right\}$ invoking Property (b) of $\mathcal{H}$ in Property A.1, then we have

$$\left\{\left\|\sum_{j=1}^{\hat{n}} \mathcal{F}_j(z_j)\right\|_{\text{F}} \geq t\right\} = \left\{\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{F}} \geq \sqrt{2}t\right\} \subseteq \left\{\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}} \geq \frac{\sqrt{2}t}{\sqrt{l_1+l_2}}\right\}.$$

Letting $t' = \frac{\sqrt{2}t}{\sqrt{l_1+l_2}}$ in (15), we get for any $t > 0$,

$$\mathbb{P}_\mu\left(\left\|\sum_{j=1}^{\hat{n}} \mathcal{F}_j(z_j)\right\|_{\text{F}} \geq t\right)$$
$$\leq \mathbb{P}_\mu\left(\left\|\sum_{j=1}^{\hat{n}} \widetilde{\mathcal{F}}_j(z_j)\right\|_{\text{op}} \geq t'\right)$$
$$\leq (l_1+l_2)^{2-\pi/4} \exp\left(\frac{-\pi^2 t^2}{16(l_1+l_2)\alpha(\lambda)\tau^2 + 16\sqrt{2(l_1+l_2)}\beta(\lambda)M't}\right).$$

By assumption (A1) and Lemma 1, we get $\alpha(\lambda) = \frac{1+\lambda}{1-\lambda}$ and $\beta(\lambda) = \frac{8/\pi}{1-\lambda}$. Then we obtain for any $t > 0$,

$$\mathbb{P}_\mu\left(\left\|\sum_{j=1}^{\hat{n}} \mathcal{F}_j(z_j)\right\|_{\text{F}} \geq t\right) \leq (l_1+l_2)^{2-\pi/4} \exp\left(\frac{-\gamma\pi^3 t^2}{32\pi(l_1+l_2)\tau^2 + 128\sqrt{2(l_1+l_2)}M't}\right),$$

where $\gamma = 1 - \lambda$. The result is then proved, by letting $t = \hat{n}s$. □

B.2 Proof of Lemma 3

Proof of Lemma 3. By (E2) and Definition A.5, we have $\sup_{v \in \mathbb{S}^{l-1}} \|\langle \mathcal{F}_j(z), v\rangle\|_{\psi_1(\mu;\mathcal{X})} \leq B$, then for any vector e_p whose p-th component is 1 and all other components are 0, $p = 1, \ldots, l$, the projection $e_p^{\text{T}}\mathcal{F}_j(z) = \mathcal{F}_{j,p}(z)$ satisfies $\|\mathcal{F}_{j,p}(z)\|_{\psi_1(\mu;\mathcal{X})} \leq B$. $\mathcal{F}_{j,p}(z)$ denotes the p-th component of $\mathcal{F}_j(z)$. By Definition A.4, there exists a constant $C_0 > 0$ such that for any $t > 0$,

$$\mathbb{P}_\mu\left(|\mathcal{F}_{j,p}(z)| > t\right) \leq 2\exp\left(-\frac{t}{C_0 B}\right).$$

Thus

$$\begin{aligned}\mathbb{E}_\mu\left[(\mathcal{F}_{j,p}(z))^2\right] &= \int_0^\infty \mathbb{P}_\mu\left((\mathcal{F}_{j,p}(z))^2 > z\right)\mathrm{d}z = \int_0^\infty \mathbb{P}_\mu\left(|\mathcal{F}_{j,p}(z)| > \sqrt{z}\right)\mathrm{d}z \\ &\le \int_0^\infty 2\exp\left(-\frac{\sqrt{z}}{C_0 B}\right)\mathrm{d}z \overset{(z=u^2)}{=} 4\int_0^\infty u\exp\left(-\frac{u}{C_0 B}\right)\mathrm{d}u = 4C_0^2B^2.\end{aligned} \tag{16}$$

By Jensen's inequality, we have

$$\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\mathrm{F}} \le \mathbb{E}_\mu\left[\left\|\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right\|_{\mathrm{F}}\right] = \mathbb{E}_\mu\left[\left\|\mathcal{F}_j(z)\right\|_2^2\right].$$

Moreover, $\left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\mathrm{F}} = \left\|\mathbb{E}_\mu\left[\left\|\mathcal{F}_j(z)\right\|_2^2\right]\right\|_{\mathrm{F}} = \mathbb{E}_\mu\left[\left\|\mathcal{F}_j(z)\right\|_2^2\right]$. Then by (16) we can derive

$$\max\left\{\left\|\mathbb{E}_\mu\left[\mathcal{F}_j(z)\mathcal{F}_j^\top(z)\right]\right\|_{\mathrm{F}}, \left\|\mathbb{E}_\mu\left[\mathcal{F}_j^\top(z)\mathcal{F}_j(z)\right]\right\|_{\mathrm{F}}\right\} \le \mathbb{E}_\mu\left[\left\|\mathcal{F}_j(z)\right\|_2^2\right] \le 4lC_0^2B^2.$$

This completes the proof. □

B.3 Proof of Lemma 4

A crucial step in Lemma 3 is the application of an ε-net technique to discretize the unit Euclidean ball $\mathbb{S}^{l-1}$ thereby approximating the supremum of an infinite stochastic process by a maximum over a finite set. This enables the use of a union bound coupled with Markov's inequality to derive a tail bound for the L_2-norm of $\mathcal{F}_j(z)$.

Lemma B.1 comes from Vershynin (2018), which provides upper and lower bounds on the covering numbers of the Euclidean unit ball $\mathbb{S}^{l-1}$.

Lemma B.1. *(Corollary 4.2.11, (Vershynin, 2018)) The covering number of the unit Euclidean ball $\mathbb{S}^{l-1}$ satisfy the following for any $\varepsilon > 0$:*

$$\left(\frac{1}{\varepsilon}\right)^l \le \mathcal{N}\left(\mathbb{S}^{l-1}, \varepsilon\right) \le \left(\frac{2}{\varepsilon} + 1\right)^l.$$

Proof of Lemma 4. Let $v^* \in \mathbb{S}^{l-1}$ be a unit vector such that $v^{*\top}\mathcal{F}_j(z) = \sup_{v\in\mathbb{S}^{l-1}} v^\top\mathcal{F}_j(z) = \left\|\mathcal{F}_j(z)\right\|_2$. Fix $\varepsilon \in (0,1)$, and let $\mathcal{N} \subset \mathbb{S}^{l-1}$ be a finite ε-net. Then there exists $u \in \mathcal{N}$ satisfying $\|v^* - u\|_2 < \varepsilon$. By the Cauchy–Schwarz inequality, we have

$$u^\top\mathcal{F}_j(z) = \left(u - v^* + v^*\right)^\top\mathcal{F}_j(z) \ge \left\|\mathcal{F}_j(z)\right\|_2 - \left\|u - v^*\right\|_2\left\|\mathcal{F}_j(z)\right\|_2 \ge (1-\varepsilon)\left\|\mathcal{F}_j(z)\right\|_2.$$

It follows that if $\left\|\mathcal{F}_j(z)\right\|_2 > t$, there exists $u \in \mathcal{N}$ such that $u^\top\mathcal{F}_j(z) > (1-\varepsilon)t$. Therefore, we obtain the inclusion

$$\left\{\left\|\mathcal{F}_j(z)\right\|_2 > t\right\} \subseteq \bigcup_{u\in\mathcal{N}}\left\{u^\top\mathcal{F}_j(z) > (1-\varepsilon)t\right\}. \tag{17}$$

From (17) we get

$$\begin{aligned}\mathbb{P}_\mu\left(\left\|\mathcal{F}_j(z)\right\|_2 > t\right) &\le \sum_{u\in\mathcal{N}}\mathbb{P}_\mu\left(u^\top\mathcal{F}_j(z) > (1-\varepsilon)t\right) \\ &\le \mathcal{N}\left(\mathbb{S}^{l-1}, \varepsilon\right)\max_{u\in\mathcal{N}}\mathbb{P}_\mu\left(u^\top\mathcal{F}_j(z) > (1-\varepsilon)t\right),\end{aligned} \tag{18}$$

where $\mathcal{N}\left(\mathbb{S}^{l-1}, \varepsilon\right)$ is the covering numbers of the unit Euclidean ball $\mathbb{S}^{l-1}$. Since $\|\mathcal{F}_j\|_{\psi_1(\mu;\mathcal{X})} \le B$, by Definition A.5 it holds that $\left\|v^\top\mathcal{F}_j(z)\right\|_{\psi_1(\mu;\mathcal{X})} \le B$ for all $v \in \mathbb{S}^{l-1}$. Applying Markov's inequality gives

$$\mathbb{P}_\mu\left(\left|v^\top\mathcal{F}_j(z)\right| > t\right) \le \exp\left(-\frac{t}{B}\right)\mathbb{E}_\mu\left[\exp\left(\frac{\left|v^\top\mathcal{F}_j(z)\right|}{B}\right)\right] \le 2\exp\left(-\frac{t}{B}\right).$$

Then for any $u \in \mathcal{N}$,

$$\mathbb{P}_\mu\left(u^\top\mathcal{F}_j(z) > (1-\varepsilon)t\right) \le 2\exp\left(-\frac{(1-\varepsilon)t}{B}\right). \tag{19}$$

Substituting (19) into (18) and applying Lemma B.1 yields the desired inequality. □

B.4 PROOF OF LEMMA 5

Proof. To deal with the possibly unbounded stochastic gradient $G(\theta; z)$, we firstly fix $M_1 > 0$ and consider the truncation function $\bar{G}(\theta; z) = G(\theta; z) \cdot \min\left(1, \frac{M_1}{\|G(\theta;z)\|_2}\right)$ and $\hat{G}(\theta; z) = G(\theta; z) - \bar{G}(\theta; z)$. If $\{Z_j\}_{j=1}^{\hat{n}}$ is a non-stationary Markov chain with stationary distribution μ_θ and absolute spectral gap $\gamma(P_\theta)$, then for any event $A \in \sigma(Z_1, \cdots, Z_{\hat{n}})$, we have from the Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{P}_\nu(A) &= \int_{\mathcal{X}} \mathbb{P}(A|Z_1 = z)\nu(\mathrm{d}z) = \int_{\mathcal{X}} \mathbb{P}(A|Z_1 = z)\frac{\mathrm{d}\nu}{\mathrm{d}\mu_\theta}\mu_\theta(\mathrm{d}z) \\ &\leq \sqrt{\int_{\mathcal{X}} \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu_\theta}\right)^2 \mathbb{P}(A|Z_1 = z)\mu_\theta(\mathrm{d}z) \int_{\mathcal{X}} \mathbb{P}(A|Z_1 = z)\mu_\theta(\mathrm{d}z)} \\ &\leq C\sqrt{\mathbb{P}_{\mu_\theta}(A)}.\end{aligned} \tag{20}$$

Hence, we only need to consider the stationary case $\nu = \mu_\theta$, i.e., $\{z_i\}_{i=1}^{\hat{n}}$ is drawn from a stationary Markov chain.

Note that

$$\begin{aligned}&\mathbb{P}_{\mu_\theta}\left(\left\{\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} G(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[G(\theta; z)\right]\right\|_2 \geq s\right\} \cap \left\{\forall j \in [\hat{n}], \|G(\theta; z_j)\|_2 \leq M_1\right\}\right) \\ &= \mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} \bar{G}(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta; z) + \hat{G}(\theta; z)\right]\right\|_2 \geq s\right) \\ &\leq \mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} \bar{G}(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta; z)\right]\right\|_2 + \left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta; z)\right]\right\|_2 \geq s\right) \\ &= \mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} \bar{G}(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta; z)\right]\right\|_2 \geq s - \left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta; z)\right]\right\|_2\right),\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_{\mu_\theta}\left(\exists\, j \in [\hat{n}],\, \|G(\theta; z_j)\|_2 > M_1\right) &\leq \sum_{j=1}^{\hat{n}} \mathbb{P}_{\mu_\theta}\left(\|G(\theta; z_j)\|_2 > M_1\right) \\ &\leq 2\hat{n}\left(\frac{2}{\varepsilon} + 1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right),\end{aligned}$$

where in the last inequality we use (H2) and Lemma 4.

Then, we have

$$\begin{aligned}&\mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} G(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[G(\theta; z)\right]\right\|_2 \geq s\right) \\ &\leq \mathbb{P}_{\mu_\theta}\left(\left\{\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} G(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[G(\theta; z)\right]\right\|_2 \geq s\right\} \cap \left\{\forall j \in [\hat{n}], \|G(\theta; z_j)\|_2 \leq M_1\right\}\right) \\ &\quad + \mathbb{P}_{\mu_\theta}\left(\exists\, j \in [\hat{n}],\, \|G(\theta; z_j)\|_2 > M_1\right) \\ &\leq \mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} \bar{G}(\theta; z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta; z)\right]\right\|_2 \geq s - \left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta; z)\right]\right\|_2\right) \\ &\quad + 2\hat{n}\left(\frac{2}{\varepsilon} + 1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right).\end{aligned} \tag{21}$$

Next we give a bound for $\mathbb{E}_{\mu_\theta}\left[\left\|\hat{G}(\theta;z)\right\|_2\right]$. Since $\hat{G}(\theta;z) \le G(\theta;z)\mathbf{1}_{\{\|G(\theta;z)\|_2 > M_1\}}$, we find

$$\begin{aligned}
\left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta;z)\right]\right\|_2 &\le \mathbb{E}_{\mu_\theta}\left[\left\|\hat{G}(\theta;z)\right\|_2\right] \le \mathbb{E}_{\mu_\theta}\left[\left\|G(\theta;z)\mathbf{1}_{\{\|G(\theta;z)\|_2 > M_1\}}\right\|_2\right] \\
&= \int_0^\infty \mathbb{P}_{\mu_\theta}\left(\left\|G(\theta;z)\mathbf{1}_{\{\|G(\theta;z)\|_2 > M_1\}}\right\|_2 > t\right) dt \\
&= \int_0^{M_1} \mathbb{P}_{\mu_\theta}\left(\|G(\theta;z)\|_2 > M_1\right) dt + \int_{M_1}^\infty \mathbb{P}_{\mu_\theta}\left(\|G(\theta;z)\|_2 > t\right) dt \\
&\le 2M_1\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right) \\
&\qquad + 2\frac{B}{(1-\varepsilon)}\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right) \\
&= \left(2M_1 + 2\frac{B}{(1-\varepsilon)}\right)\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right),
\end{aligned}$$

where in the last inequality we use Lemma 4. Therefore, when $s \ge 2\left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta;z)\right]\right\|_2$, by Lemma 2, Lemma 3 and assumption (H1) it holds that

$$\begin{aligned}
&\mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}}\bar{G}(\theta;z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta;z)\right]\right\|_2 \ge s - \left\|\mathbb{E}_{\mu_\theta}\left[\hat{G}(\theta;z)\right]\right\|_2\right) \\
&\le \mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}}\bar{G}(\theta;z_j) - \mathbb{E}_{\mu_\theta}\left[\bar{G}(\theta;z)\right]\right\|_2 \ge \frac{s}{2}\right) \\
&\le (l+1)^{2-\pi/4}\exp\left(\frac{-\gamma^\star\hat{n}\pi^3 s^2}{512\pi(l^2+l)C_0^2B^2 + 256\sqrt{2(l+1)}M_1 s}\right).
\end{aligned} \tag{22}$$

For any fixed $s \ge \frac{(\log N_3)^2}{(N_1-N_2)^2\gamma^\star\hat{n}}$, $N_1 = \frac{(1-\varepsilon)C_1}{\sqrt{B}}$, $N_2 = \frac{\pi^3}{512C_1\sqrt{2B(l+1)}}$ and $N_3 = \frac{2\hat{n}(\frac{2}{\varepsilon}+1)^l}{(l+1)^{2-\frac{\pi}{4}}}$, we take $M_1 = C_1\sqrt{\gamma^\star\hat{n}sB}$ and $C_1 = \frac{\pi^{\frac{3}{2}}}{(2l+2)^{\frac{1}{4}}(256(1-\varepsilon))^{\frac{1}{2}}}$. Then, by (21) and (22) we obtain

$$\begin{aligned}
&\mathbb{P}_{\mu_\theta}\left(\left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}}G(\theta;z_j) - \mathbb{E}_{\mu_\theta}\left[G(\theta;z)\right]\right\|_2 \ge s\right) \\
&\le (l+1)^{2-\pi/4}\exp\left(\frac{-\gamma^\star\hat{n}\pi^3 s^2}{512\pi(l^2+l)C_0^2B^2 + 256\sqrt{2(l+1)}M_1 s}\right) \\
&\qquad + 2\hat{n}\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)M_1}{B}\right) \\
&\le (l+1)^{2-\pi/4}\exp\left(\frac{-\gamma^\star\hat{n}\pi^3 s^2}{2\max\{512\pi(l^2+l)C_0^2B^2, 256\sqrt{2(l+1)}C_1\sqrt{\gamma^\star\hat{n}sB}s\}}\right) \\
&\qquad + 2\hat{n}\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)C_1\sqrt{\gamma^\star\hat{n}sB}}{B}\right) \\
&\le (l+1)^{2-\pi/4}\exp\left(\frac{-\gamma^\star\hat{n}\pi^3 s^2}{1024\pi(l^2+l)C_0^2B^2}\right) + (l+1)^{2-\pi/4}\exp\left(\frac{-\pi^3\sqrt{\gamma^\star\hat{n}s}}{512C_1\sqrt{2B(l+1)}}\right) \\
&\qquad + 2\hat{n}\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)C_1\sqrt{\gamma^\star\hat{n}s}}{\sqrt{B}}\right) \\
&\le (l+1)^{2-\pi/4}\exp\left(\frac{-\gamma^\star\hat{n}\pi^2 s^2}{1024(l^2+l)C_0^2B^2}\right) + 2(l+1)^{2-\pi/4}\exp\left(\frac{-\pi^3\sqrt{\gamma^\star\hat{n}s}}{512C_1\sqrt{2B(l+1)}}\right),
\end{aligned}$$

where the last inequality holds from $(l+1)^{2-\pi/4}\exp\left(\frac{-\pi^3\sqrt{\gamma^\star\hat{n}s}}{512C_1\sqrt{2B(l+1)}}\right) \geq 2\hat{n}\left(\frac{2}{\varepsilon}+1\right)^l\exp\left(-\frac{(1-\varepsilon)C_1\sqrt{\gamma^\star\hat{n}s}}{\sqrt{B}}\right)$ as long as $s \geq \frac{(\log N_3)^2}{(N_1-N_2)^2\gamma^\star\hat{n}}$. Using (20), this completes the proof. □

B.5 Proof of Theorem 1

Proof. We follow the notation in the proof of Lemma 5. Denote $A_2 = \log(C(l+1)^{2-\frac{\pi}{4}})$, $R_1 = \sqrt{\frac{2048(l^2+l)C_0^2B^2}{\gamma^\star\hat{n}\pi^2}}$, $A_3 = \log(2C(l+1)^{2-\frac{\pi}{4}})$, and $R_2 = \frac{1024C_1}{\pi^3}\sqrt{\frac{2B(l+1)}{\gamma^\star\hat{n}}}$. Moreover, s_0, N_1, N_2 and N_3 are identical to those appearing in Lemma 5. Then, we only need to bound $\mathbb{E}\left[\Delta^m\right]$ under the condition

$$\mathbb{P}_\nu(\Delta \geq s) \leq \exp\left(A_2 - \frac{s^2}{R_1^2}\right) + \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right), \quad \forall s \geq s_0,$$

where $\Delta = \left\|g(\theta;S) - \nabla f(\theta)\right\|_2$. Note that

$$\begin{aligned}
&\frac{1}{m}\mathbb{E}_\nu\left[\Delta^m\right] \\
&= \int_0^\infty s^{m-1}\mathbb{P}_\nu(\Delta \geq s)\mathrm{d}s \\
&\leq \int_0^{s_0} s^{m-1}ds + \int_{s_0}^\infty s^{m-1}\min\left\{1, \exp\left(A_2 - \frac{s^2}{R_1^2}\right) + \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right)\right\}\mathrm{d}s \\
&\leq \frac{s_0^m}{m} + I_1 + I_2.
\end{aligned} \tag{23}$$

where $I_1 = \int_{s_0}^\infty s^{m-1}\min\left\{1, \exp\left(A_2 - \frac{s^2}{R_1^2}\right)\right\}\mathrm{d}s$, $I_2 = \int_{s_0}^\infty s^{m-1}\min\left\{1, \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right)\right\}\mathrm{d}s$.

We first give a bound for I_1. When $s_0 \geq R_1\sqrt{A_2}$, it follows that

$$\begin{aligned}
I_1 &= \int_{s_0}^\infty s^{m-1}\exp\left(A_2 - \frac{s^2}{R_1^2}\right)\mathrm{d}s \\
&= \frac{R_1^m}{2}\int_{s_0}^\infty \left(\frac{s}{R_1}\right)^{m-2}\exp\left(A_2 - \frac{s^2}{R_1^2}\right)\mathrm{d}\left(\frac{s^2}{R_1^2}\right) \\
&= \frac{R_1^m}{2}\int_{\frac{s_0^2}{R_1^2}}^\infty t^{\frac{m-2}{2}}\exp\left(A_2 - t\right)\mathrm{d}t \\
&\leq \frac{R_1^m}{2}\int_{A_2}^\infty t^{\frac{m-2}{2}}\exp\left(A_2 - t\right)\mathrm{d}t \\
&\leq \frac{R_1^m}{m}\left(\left(A_2 + \frac{m}{2}\right)^{\frac{m}{2}} - A_2^{\frac{m}{2}}\right),
\end{aligned} \tag{24}$$

where the last inequality is due to the fact $\int_A^\infty t^{k-1}\exp(A-t)\mathrm{d}t \leq (A+k-1)^{k-1} \leq \frac{((A+k)^k - A^k)}{k}$ for any integer $k \geq 1$.

When $0 \le s_0 < R_1\sqrt{A_2}$, it follows that

$$\begin{aligned} I_1 &= \int_{s_0}^{\infty} s^{m-1} \min\left\{1, \exp\left(A_2 - \frac{s^2}{R_1^2}\right)\right\} \mathrm{d}s \\ &= \int_{s_0}^{R_1\sqrt{A_2}} s^{m-1} \mathrm{d}s + \int_{R_1\sqrt{A_2}}^{\infty} s^{m-1} \exp\left(A_2 - \frac{s^2}{R_1^2}\right) \mathrm{d}s \\ &\le \int_{s_0}^{\infty} s^{m-1} \mathrm{d}s + \int_{R_1\sqrt{A_2}}^{\infty} s^{m-1} \exp\left(A_2 - \frac{s^2}{R_1^2}\right) \mathrm{d}s \\ &= \frac{s_0^m}{m} + \frac{R_1^m}{2} \int_{R_1\sqrt{A_2}}^{\infty} \left(\frac{s}{R_1}\right)^{m-2} \exp\left(A_2 - \frac{s^2}{R_1^2}\right) \mathrm{d}\left(\frac{s^2}{R_1^2}\right) \\ &= \frac{s_0^m}{m} + \frac{R_1^m}{2} \int_{A_2}^{\infty} t^{\frac{m-2}{2}} \exp\left(A_2 - t\right) \mathrm{d}t \\ &\le \frac{s_0^m}{m} + \frac{R_1^m}{m} \left(\left(A_2 + \frac{m}{2}\right)^{\frac{m}{2}} - A_2^{\frac{m}{2}}\right), \end{aligned} \tag{25}$$

where the last inequality is due to the fact $\int_A^{\infty} t^{k-1} \exp(A-t)\mathrm{d}t \le (A+k-1)^{k-1} \le \frac{((A+k)^k - A^k)}{k}$ for any integer $k \ge 1$.
Combining (24) and (25), we obtain that for any $s_0 \ge 0$,

$$I_1 \le \frac{s_0^m}{m} + \frac{R_1^m}{m} \left(\left(A_2 + \frac{m}{2}\right)^{\frac{m}{2}} - A_2^{\frac{m}{2}}\right). \tag{26}$$

Similarly, we give a bound for I_2. When $s_0 \ge R_2^2 A_3^2$, if follows that

$$\begin{aligned} I_2 &= \int_{s_0}^{\infty} s^{m-1} \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right) \mathrm{d}s \\ &= 2R_2^{2m} \int_{s_0}^{\infty} \left(\frac{\sqrt{s}}{R_2}\right)^{2m-1} \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right) \mathrm{d}\left(\frac{\sqrt{s}}{R_2}\right) \\ &= 2R_2^{2m} \int_{\frac{\sqrt{s_0}}{R_2}}^{\infty} t^{2m-1} \exp\left(A_3 - t\right) \mathrm{d}t \\ &\le 2R_2^{2m} \int_{A_3}^{\infty} t^{2m-1} \exp\left(A_3 - t\right) \mathrm{d}t \\ &\le \frac{R_2^{2m}}{m} \left(\left(A_3 + 2m\right)^{2m} - A_3^{2m}\right). \end{aligned} \tag{27}$$

When $0 \le s_0 < R_2^2 A_3^2$, it follows that

$$\begin{aligned} I_2 &= \int_{s_0}^{\infty} s^{m-1} \min\left\{1, \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right)\right\} \mathrm{d}s \\ &= \int_{s_0}^{R_2^2 A_3^2} s^{m-1} \mathrm{d}s + \int_{R_2^2 A_3^2}^{\infty} s^{m-1} \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right) \mathrm{d}s \\ &\le \int_{s_0}^{\infty} s^{m-1} \mathrm{d}s + \int_{R_2^2 A_3^2}^{\infty} s^{m-1} \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right) \mathrm{d}s \\ &= \frac{s_0^m}{m} + 2R_2^{2m} \int_{R_2^2 A_3^2}^{\infty} \left(\frac{\sqrt{s}}{R_2}\right)^{2m-1} \exp\left(A_3 - \frac{\sqrt{s}}{R_2}\right) \mathrm{d}\left(\frac{\sqrt{s}}{R_2}\right) \\ &= \frac{s_0^m}{m} + 2R_2^{2m} \int_{A_3}^{\infty} t^{2m-1} \exp\left(A_3 - t\right) \mathrm{d}t \\ &\le \frac{s_0^m}{m} + \frac{R_2^{2m}}{m} \left(\left(A_3 + 2m\right)^{2m} - A_3^{2m}\right). \end{aligned} \tag{28}$$

Combining (27) and (28), we obtain that for any $s_0 \ge 0$,

$$I_2 \le \frac{s_0^m}{m} + \frac{R_2^{2m}}{m} \left(\left(A_3 + 2m\right)^{2m} - A_3^{2m}\right). \tag{29}$$

Combining (23), (26) and (29), we obtain

$$\begin{aligned}\mathbb{E}_\nu\left[\Delta^m\right] &\leq s_0^m + mI_1 + mI_2 \\ &\leq 3s_0^m + R_1^m\left(\left(A_2+\frac{m}{2}\right)^{\frac{m}{2}} - A_2^{\frac{m}{2}}\right) + R_2^{2m}\left((A_3+2m)^{2m} - A_3^{2m}\right) \\ &\leq 3s_0^m + R_1^m\left(\left(A_2+\frac{m}{2}\right)^{\frac{m}{2}}\right) + R_2^{2m}\left((A_3+2m)^{2m}\right).\end{aligned} \tag{30}$$

When $m=1$, we get

$$\begin{aligned}&\mathbb{E}_\nu\left[\|g(\theta;S)-\nabla f(\theta)\|_2\right] \\ &\leq \frac{3(\log N_3)^2}{(N_1-N_2)^2\gamma^\star\hat{n}} + \sqrt{\frac{2048(\log(C(l+1)^{2-\frac{\pi}{4}})+\frac{1}{2})(l^2+l)C_0^2B^2}{\gamma^\star\hat{n}\pi^2}} \\ &\qquad + \frac{(1024\sqrt{2})^2(\log(2C(l+1)^{2-\frac{\pi}{4}})+2)^2(l+1)C_1^2B}{\gamma^\star\hat{n}\pi^6} \\ &= \mathcal{O}\left(\sqrt{\frac{\log(C(l+1)^{2-\frac{\pi}{4}})(l^2+l)B^2}{\gamma^\star\hat{n}}}\right).\end{aligned}$$

When $m=2$, we get

$$\begin{aligned}&\mathbb{E}_\nu\left[\|g(\theta;S)-\nabla f(\theta)\|_2^2\right] \\ &\leq \frac{3(\log N_3)^4}{(N_1-N_2)^4(\gamma^\star)^2\hat{n}^2} + \frac{2048(\log(C(l+1)^{2-\frac{\pi}{4}})+1)(l^2+l)C_0^2B^2}{\gamma^\star\hat{n}\pi^2} \\ &\qquad + \frac{(1024\sqrt{2})^4(\log(2C(l+1)^{2-\frac{\pi}{4}})+4)^4(l+1)^2C_1^4B^2}{(\gamma^\star)^2\hat{n}^2\pi^{12}} \\ &= \mathcal{O}\left(\frac{\log(C(l+1)^{2-\frac{\pi}{4}})(l^2+l)B^2}{\gamma^\star\hat{n}}\right).\end{aligned}$$

This completes the proof. □

C Proof of Results in Section 4

C.1 Key Lemmas in Section 4

Lemma C.2 establishes a relaxed triangle inequality for the sub-exponential norm of random variables. Lemma C.3 provides a uniformly upper bound for the second-order moment of the gradient estimator when the samples are generated from the non-stationary Markov chains. Lemma C.4 provides a bound for the size of $\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)$. Lemma C.5 establishes a recurrence relation for the deviation $\|m_t - \nabla f(\theta_t)\|^2$. By leveraging this recursion, we can bridge the gap between the momentum error and the stochastic gradient error $\|g_t - \nabla f(\theta_t)\|^2$ in the proof of Theorem 2. This transformation enables us to invoke the variance bounds established in Theorem 1 to conclude the final convergence analysis. Lemma C.6 characterizes the relationship between the L_2 norm $\|\cdot\|$ and the weighted norm $\|\cdot\|_{D_t^{-1}}$, which helps us give a convergence result in the sense of the L_2 norm $\|\cdot\|$.

Lemma C.2. *Let $\{Z_i\}_{i=1}^{\hat{n}}$ be a collection of random variables (not necessarily independent) where each $Z_i : \mathcal{X} \to \mathbb{R}$, there exists a constant $C' \geq 1$ such that*

$$\left\|\sum_{i=1}^{\hat{n}} Z_i\right\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq C' \sum_{i=1}^{\hat{n}} \|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}.$$

.

Proof of Lemma C.2. Denote $\tilde{S} = \sum_{i=1}^{\hat{n}} \|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}$. By Definition A.4, $\left\|\sum_{i=1}^{\hat{n}} Z_i\right\|_{\psi_1(\mu_\theta;\mathcal{X})}$ is the smallest K that satisfies $\mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{|\sum_{i=1}^{\hat{n}} Z_i|}{\|\sum_{i=1}^{\hat{n}} Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right)\right] \leq 2$. Hence it suffices to prove that

there exists a constant $C' \geq 1$ satisfying

$$\mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{\left|\sum_{i=1}^{\hat{n}} Z_i\right|}{C'\tilde{S}}\right)\right] \leq 2. \tag{31}$$

By the triangle inequality, we have $\mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{|\sum_{i=1}^{\hat{n}} Z_i|}{C'\tilde{S}}\right)\right] \leq \mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{\sum_{i=1}^{\hat{n}}|Z_i|}{C'\tilde{S}}\right)\right]$. Define $\alpha_i = \frac{\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}{\tilde{S}}$, $i = 1, \ldots, \hat{n}$. Hence, $\frac{\sum_{i=1}^{\hat{n}}|Z_i|}{C'\tilde{S}} = \sum_{i=1}^{\hat{n}} \alpha_i \frac{|Z_i|}{C'\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}$. Then we get

$$\exp\left(\frac{\sum_{i=1}^{\hat{n}}|Z_i|}{C'\tilde{S}}\right) = \exp\left(\sum_{i=1}^{\hat{n}} \alpha_i \frac{|Z_i|}{C'\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right) \leq \sum_{i=1}^{\hat{n}} \alpha_i \exp\left(\frac{|Z_i|}{C'\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right). \tag{32}$$

Taking expectation on both sides of (32), we obtain

$$\mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{\sum_{i=1}^{\hat{n}}|Z_i|}{C'\tilde{S}}\right)\right] \leq \sum_{i=1}^{\hat{n}} \alpha_i \, \mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{|Z_i|}{C'\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right)\right].$$

By Definition A.4, $\mathbb{E}_{\mu_\theta}\left[\frac{|Z_i|}{\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right] \leq 2$. As $C' \geq 1$, we obtain

$$\mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{|Z_i|}{C'\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right)\right] \leq \mathbb{E}_{\mu_\theta}\left[\exp\left(\frac{|Z_i|}{\|Z_i\|_{\psi_1(\mu_\theta;\mathcal{X})}}\right)\right] \leq 2.$$

Then we can get (31), which can induce the result. □

Lemma C.3. *Suppose that all the conditions in Lemma 5 are satisfied. The gradient estimator $g(\theta; S) = \frac{1}{\hat{n}} \sum_{j=n_b+1}^{n_b+\hat{n}} G(\theta; z_j)$ has the following bounds that*

$$\mathbb{E}_\nu\left[\|g(\theta; S)\|_2^2\right] \leq B',$$

where $B' = \frac{8\sqrt{2}CC'^2B^2(\frac{2}{\varepsilon}+1)^{\frac{l}{2}}}{(1-\varepsilon)^2}$, C is defined in Lemma 5, C' is defined in Lemma C.2, B is defined in assumption (H2) and ε is defined in Lemma 4.

Proof of Lemma C.3. For any unit vector $u \in \mathbb{S}^{\ell-1}$, we have $\langle u, g(\theta; S)\rangle = \frac{1}{\hat{n}} \sum_{j=1}^{\hat{n}} \langle u, G(\theta; z_i)\rangle$. Let $Z'_j = \langle u, G(\theta; z_i)\rangle$, we have $\left\|\sum_{j=1}^{\hat{n}} Z'_j\right\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq C' \sum_{j=1}^{\hat{n}} \|Z'_j\|_{\psi_1(\mu_\theta;\mathcal{X})}$ according to Lemma C.2. Moreover, $\|Z'_j\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq B$ for each j by assumption (H2) and Definition A.5. Consequently,

$$\|\langle u, g(\theta; S)\rangle\|_{\psi_1(\mu_\theta;\mathcal{X})} = \left\|\frac{1}{\hat{n}}\sum_{j=1}^{\hat{n}} Z'_j\right\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq \frac{C'}{\hat{n}} \sum_{j=1}^{\hat{n}} \|Z'_j\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq \frac{C'}{\hat{n}} \hat{n} B = C'B.$$

We can take the supremum over all unit vector $u \in \mathbb{S}^{\ell-1}$ to obtain $\|g(\theta; S)\|_{\psi_1(\mu_\theta;\mathcal{X})} \leq C'B$. According to Lemma 4, we have

$$\mathbb{P}_{\mu_\theta}\left(\|g(\theta; S)\|_2 > t\right) = \mathbb{P}_{\mu_\theta}\left(\sup_{v\in\mathbb{S}^{l-1}} v^\top g(\theta; S) > t\right) \leq 2\left(\frac{2}{\varepsilon}+1\right)^l \exp\left(-\frac{(1-\varepsilon)t}{C'B}\right). \tag{33}$$

By (20) we know $\mathbb{P}_\nu\left(\|g(\theta; S)\|_2 > t\right) \leq C\sqrt{\mathbb{P}_{\mu_\theta}\left(\|g(\theta; S)\|_2 > t\right)}$. Finally, by using (33), we can calculate the bound on the second moment of $\|g(\theta; S)\|$:

$$\begin{aligned}\mathbb{E}_\nu\left[\|g(\theta; S)\|_2^2\right] &\leq \int_0^\infty \mathbb{P}_\nu\left(||g(\theta; S)||_2^2 > z\right) \mathrm{d}z = \int_0^\infty \mathbb{P}_\nu\left(||g(\theta; S)||_2 > \sqrt{z}\right) \mathrm{d}z \\ &\leq \int_0^\infty \sqrt{2}C\left(\frac{2}{\varepsilon}+1\right)^{\frac{l}{2}} \exp\left(-\frac{(1-\varepsilon)\sqrt{z}}{2C'B}\right) \mathrm{d}z = \frac{8\sqrt{2}CC'^2B^2\left(\frac{2}{\varepsilon}+1\right)^{\frac{l}{2}}}{(1-\varepsilon)^2}.\end{aligned}$$

This completes the proof. □

Lemma C.4. *Let θ_{t+1} be given in* (6)*. Then, for any $\theta_t \in \Theta$, $m_t \in \mathbb{R}^l$ and $\alpha_t > 0$, we have*

$$\left\langle m_t,\ \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\rangle_{D_t^{-1}} \geq \left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2_{D_t^{-1}} + \frac{1}{\alpha_t}(h(\theta_{t+1}) - h(\theta_t)).$$

Proof of Lemma C.4. By the optimality condition of (6), we have

$$0 \in m_t + \frac{1}{\alpha_t} D_t(\theta_{t+1} - \theta_t) + \partial h(\theta_{t+1}) + N(\theta_{t+1}; \Theta).$$

For any $p \in \partial h(\theta_{t+1})$, we have

$$\left\langle m_t + \frac{1}{\alpha_t} D_t(\theta_{t+1} - \theta_t) + p,\ \theta - \theta_{t+1} \right\rangle \geq 0, \quad \forall \theta \in \Theta. \tag{34}$$

Letting $\theta = \theta_t$ in (34), by the convexity of h, we further obtain

$$\left\langle m_t + \frac{1}{\alpha_t} D_t(\theta_{t+1} - \theta_t),\ \theta_t - \theta_{t+1} \right\rangle \geq \langle p,\ \theta_{t+1} - \theta_t\rangle \geq h(\theta_{t+1}) - h(\theta_t).$$

By (7) we have

$$\left\langle m_t,\ \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\rangle_{D_t^{-1}} \geq \left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2_{D_t^{-1}} + \frac{1}{\alpha_t}(h(\theta_{t+1}) - h(\theta_t)).$$

This completes the proof. □

Lemma C.5. *The following recursion holds*

$$\begin{aligned}\|m_t - \nabla f(\theta_t)\|^2 \leq (1-\beta_t)\|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{4}{\beta_t}\|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 \\ + 2\beta_t \|g_t - \nabla f(\theta_t)\|^2.\end{aligned}$$

Proof of Lemma C.5. Observe that

$$\begin{aligned}&m_t - \nabla f(\theta_t)\\ &= (1-\beta_t)m_{t-1} + \beta_t g_t + (1-\beta_t)\nabla f(\theta_{t-1}) - (1-\beta_t)\nabla f(\theta_{t-1}) - (1-\beta_t)\nabla f(\theta_t) - \beta_t \nabla f(\theta_t)\\ &= (1-\beta_t)\left(m_{t-1} - \nabla f(\theta_{t-1})\right) + (1-\beta_t)\left(\nabla f(\theta_{t-1}) - \nabla f(\theta_t)\right) + \beta_t\left(g_t - \nabla f(\theta_t)\right).\end{aligned}$$

Then

$$\begin{aligned}&\|m_t - \nabla f(\theta_t)\|^2\\ &= (1-\beta_t)^2\|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + (1-\beta_t)^2\|\nabla f(\theta_{t-1}) - \nabla f(\theta_t)\|^2 + \beta_t^2\|g_t - \nabla f(\theta_t)\|^2\\ &\quad + \underbrace{2(1-\beta_t)^2\langle m_{t-1} - \nabla f(\theta_{t-1}),\ \nabla f(\theta_{t-1}) - \nabla f(\theta_t)\rangle}_{\text{(I)}}\\ &\quad + \underbrace{2\beta_t(1-\beta_t)\langle m_{t-1} - \nabla f(\theta_{t-1}),\ g_t - \nabla f(\theta_t)\rangle}_{\text{(II)}}\\ &\quad + \underbrace{2\beta_t(1-\beta_t)\langle \nabla f(\theta_{t-1}) - \nabla f(\theta_t),\ g_t - \nabla f(\theta_t)\rangle}_{\text{(III)}}.\end{aligned}$$

We now deal with (I), (II) and (III) by using the Young's inequality.

For (I): $2(1-\beta_t)^2 \langle m_{t-1} - \nabla f(\theta_{t-1}),\ \nabla f(\theta_{t-1}) - \nabla f(\theta_t)\rangle$

$$\leq (1-\beta_t)^2 \frac{\beta_t}{2(1-\beta_t)} \|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{2(1-\beta_t)^3}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2$$

$$= \frac{1}{2}\beta_t(1-\beta_t) \|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{2(1-\beta_t)^3}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2$$

$$\leq \frac{1}{2}\beta_t(1-\beta_t) \|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{2(1-\beta_t)^2}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 .$$

For (II): $2\beta_t(1-\beta_t) \langle m_{t-1} - \nabla f(\theta_t),\ g_t - \nabla f(\theta_t)\rangle$

$$\leq \frac{1}{2}\beta_t(1-\beta_t) \|m_{t-1} - \nabla f(\theta_t)\|^2 + 2\beta_t(1-\beta_t) \|g_t - \nabla f(\theta_t)\|^2 .$$

For (III): $2\beta_t(1-\beta_t) \langle \nabla f(\theta_{t-1}) - \nabla f(\theta_t),\ g_t - \nabla f(\theta_t)\rangle$

$$\leq \beta_t(1-\beta_t)\frac{(1-\beta_t)}{\beta_t^2} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 + \beta_t(1-\beta_t)\frac{\beta_t^2}{(1-\beta_t)} \|g_t - \nabla f(\theta_t)\|^2$$

$$= \frac{(1-\beta_t)^2}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 + \beta_t^3 \|g_t - \nabla f(\theta_t)\|^2$$

$$\leq \frac{(1-\beta_t)^2}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 + \beta_t^2 \|g_t - \nabla f(\theta_t)\|^2 .$$

Finally, we obtain

$$\begin{aligned}
&\|m_t - \nabla f(\theta_t)\|^2 \\
&= (1-\beta_t) \|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{4(1-\beta_t)^2}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 + 2\beta_t \|g_t - \nabla f(\theta_t)\|^2 \\
&\leq (1-\beta_t) \|m_{t-1} - \nabla f(\theta_{t-1})\|^2 + \frac{4}{\beta_t} \|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 + 2\beta_t \|g_t - \nabla f(\theta_t)\|^2 .
\end{aligned}$$

This completes the proof. □

Lemma C.6. *R is a random variable with distribution $\mathbb{P}(R = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1} \alpha_t}$. The following inequality holds*

$$\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \leq \frac{\sqrt{B'} + \epsilon}{\sum_{t=0}^{T-1} \alpha_t} \sum_{t=0}^{T-1} \alpha_t \mathbb{E}\left[\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2_{D_t^{-1}}\right].$$

Proof of Lemma C.6. We initiate the proof by introducing Hölder's inequality for arbitrary random variables X and Y such that

$$\mathbb{E}\left[|XY|\right] \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \left(\mathbb{E}\left[|Y|^q\right]\right)^{1/q}, \tag{35}$$

which p and q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Letting $p = \frac{3}{2}$, $q = 3$, $X = \left(\frac{\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2}{\sqrt{\|v_t\|_1} + \epsilon}\right)^{\frac{2}{3}}$ and $Y = \left(\sqrt{\|v_t\|_1} + \epsilon\right)^{\frac{2}{3}}$ in (35), we get

$$\mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^{\frac{4}{3}}\right] \leq \left(\mathbb{E}\left[\frac{\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2}{\sqrt{\|v_t\|_1} + \epsilon}\right]\right)^{\frac{2}{3}} \left(\mathbb{E}\left[\left(\sqrt{\|v_t\|_1} + \epsilon\right)^2\right]\right)^{\frac{1}{3}}. \tag{36}$$

To bound the first term in the right-hand side of (36), we find

$$\begin{aligned}
\frac{\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2}{\sqrt{\|v_t\|_1} + \epsilon} &= \sum_{j=1}^{l} \frac{\left|\mathcal{G}_j^{D_t}(\theta_t, m_t, \alpha_t)\right|^2}{\sqrt{\sum_{j=1}^{l} v_{tj}} + \epsilon} \\
&\leq \sum_{j=1}^{l} \frac{1}{\sqrt{v_{tj}} + \epsilon} \left|\mathcal{G}_j^{D_t}(\theta_t, m_t, \alpha_t)\right|^2 = \left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2_{D_t^{-1}},
\end{aligned} \tag{37}$$

where $\mathcal{G}_j^{D_t}(\theta_t, m_t, \alpha_t)$ is the j-th component of $\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)$. From Lemma C.5 we observe that

$$\begin{aligned}\mathbb{E}\left[\|v_t\|_1\right] &= \mathbb{E}\left[\sum_{j=1}^{l} v_{t,j}\right] = \mathbb{E}\left[\sum_{j=1}^{l}\left((1-\rho_t)v_{t-1,j} + \rho_t g_{t,j}^2\right)\right] \\ &= (1-\rho_t)\mathbb{E}\left[\sum_{j=1}^{l} v_{t-1,j}\right] + \rho_t \mathbb{E}\left[\sum_{j=1}^{l} g_{t,j}^2\right] \\ &= (1-\rho_t)\mathbb{E}\left[\|v_{t-1}\|_1\right] + \rho_t \mathbb{E}\left[\|g_t\|^2\right] \\ &\le (1-\rho_t)\mathbb{E}\left[\|v_{t-1}\|_1\right] + \rho_t B',\end{aligned}$$

and

$$\mathbb{E}\left[\|v_0\|_1\right] = \mathbb{E}\left[\left\|\rho_0 g_0^2\right\|_1\right] \le \mathbb{E}\left[\left\|g_0^2\right\|_1\right] = \mathbb{E}\left[\sum_{j=1}^{l} g_{0,j}^2\right] = \mathbb{E}\left[\|g_0\|_2^2\right] \le B'.$$

Then by induction we can prove that $\mathbb{E}[\|v_t\|_1] \le B'$. For the second term of (36), we have by Jensen's inequality

$$\begin{aligned}\mathbb{E}\left[(\sqrt{\|v_t\|_1} + \epsilon)^2\right] &= \mathbb{E}\left[\|v_t\|_1\right] + 2\epsilon\mathbb{E}\left[\sqrt{\|v_t\|_1}\right] + \epsilon^2 \le \mathbb{E}\left[\|v_t\|_1\right] + 2\epsilon\sqrt{\mathbb{E}\left[\|v_t\|_1\right]} + \epsilon^2 \\ &\le B' + 2\epsilon\sqrt{B'} + \epsilon^2 = (\sqrt{B'} + \epsilon)^2.\end{aligned} \tag{38}$$

Combining (36), (37) and (38), we can get

$$\mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^{\frac{4}{3}}\right] \le \left(\mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|_{D_t^{-1}}^2\right]\right)^{\frac{2}{3}} \left(\sqrt{B'} + \epsilon\right)^{\frac{2}{3}}. \tag{39}$$

By (39) and Jensen's inequality, we conclude that

$$\begin{aligned}\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} &= \left(\frac{1}{\sum_{t=0}^{T-1}\alpha_t}\sum_{t=0}^{T-1}\alpha_t \mathbb{E}\left[\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \\ &\le \frac{1}{\sum_{t=0}^{T-1}\alpha_t}\sum_{t=0}^{T-1}\alpha_t\left(\mathbb{E}\left[\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \\ &\le \frac{\sqrt{B'} + \epsilon}{\sum_{t=0}^{T-1}\alpha_t}\sum_{t=0}^{T-1}\alpha_t \mathbb{E}\left[\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|_{D_t^{-1}}^2\right],\end{aligned}$$

thereby completing the proof. □

C.2 Proof of Theorem 2

Proof. Denote $\delta_t = m_t - \nabla f(\theta_t)$. For $f \in \mathcal{C}_L^{1,1}(\Theta)$, we have

$$\begin{aligned}f(\theta_{t+1}) &\le f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \\ &= f(\theta_t) + \left\langle \nabla f(\theta_t), -\alpha_t D_t^{-1}\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\rangle + \frac{L}{2}\left\|\alpha_t D_t^{-1}\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|^2 \\ &= f(\theta_t) - \alpha_t\left\langle \nabla f(\theta_t), \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\rangle_{D_t^{-1}} + \frac{L\alpha_t^2}{2}\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|_{D_t^{-2}}^2.\end{aligned} \tag{40}$$

For any $u \in \mathbb{R}^l$, we have

$$\begin{aligned}\|u\|_{D_t^{-2}}^2 &= \langle u, D_t^{-2}u\rangle = \langle D_t^{-1}u, D_t^{-1}u\rangle \le \lambda_{\max}(D_t^{-1})\langle u, D_t^{-1}u\rangle \\ &= \lambda_{\max}(D_t^{-1})\|u\|_{D_t^{-1}}^2 \le \frac{1}{\epsilon}\|u\|_{D_t^{-1}}^2.\end{aligned} \tag{41}$$

Combing (40) and (41), we obtain

$$f(\theta_{t+1}) \le f(\theta_t) - \alpha_t\left\langle \nabla f(\theta_t), \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\rangle_{D_t^{-1}} + \frac{L\alpha_t^2}{2\epsilon}\left\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\right\|_{D_t^{-1}}^2. \tag{42}$$

From Lemma C.4, we have

$$h(\theta_{t+1}) - h(\theta_t) \leq \alpha_t \left\langle m_t, \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\rangle_{D_t^{-1}} - \alpha_t \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2. \tag{43}$$

Summing up (42) and (43), we get

$$\begin{aligned}
&\Phi(\theta_{t+1}) \\
&\leq \Phi(\theta_t) + \alpha_t \left\langle m_t - \nabla f(\theta_t), \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\rangle_{D_t^{-1}} + \left(\frac{L\alpha_t^2}{2\epsilon} - \alpha_t \right) \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 \\
&\leq \Phi(\theta_t) + \alpha_t \left\langle D_t^{-\frac{1}{2}} \delta_t, D_t^{-\frac{1}{2}} \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\rangle + \left(\frac{L\alpha_t^2}{2\epsilon} - \alpha_t \right) \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 \\
&\leq \Phi(\theta_t) + \alpha_t \|\delta_t\|_{D_t^{-1}} \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}} + \left(\frac{L\alpha_t^2}{2\epsilon} - \alpha_t \right) \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 \\
&\leq \Phi(\theta_t) + \frac{\alpha_t}{2} \|\delta_t\|_{D_t^{-1}}^2 + \frac{\alpha_t}{2} \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 + \left(\frac{L\alpha_t^2}{2\epsilon} - \alpha_t \right) \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2.
\end{aligned} \tag{44}$$

For any $u \in \mathbb{R}^l$, we have

$$\|u\|_{D_t^{-1}}^2 = \left\langle u,\ D_t^{-1} u \right\rangle \leq \lambda_{max}(D_t^{-1}) \left\langle u,\ u \right\rangle \leq \frac{1}{\epsilon} \left\langle u,\ u \right\rangle = \frac{1}{\epsilon} \|u\|^2. \tag{45}$$

Combing (44) and (45), we have

$$\begin{aligned}
\left(\frac{1}{2}\alpha_t - \frac{L\alpha_t^2}{2\epsilon} \right) \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 &\leq \Phi(\theta_t) - \Phi(\theta_{t+1}) + \frac{\alpha_t}{2} \|\delta_t\|_{D_t^{-1}}^2 \\
&\leq \Phi(\theta_t) - \Phi(\theta_{t+1}) + \frac{\alpha_t}{2\epsilon} \|\delta_t\|^2.
\end{aligned} \tag{46}$$

Summing up (46) for $t = 0, \ldots, T-1$ and noting that $\alpha_t \leq \frac{\epsilon}{2L}$, we obtain

$$\sum_{t=0}^{T-1} \alpha_t \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2 \leq 4(\Phi(\theta_0) - \Phi^{\text{opt}}) + 2 \sum_{t=0}^{T-1} \frac{\alpha_t}{\epsilon} \|\delta_t\|^2. \tag{47}$$

By Lemma C.5, we have

$$\begin{aligned}
&\beta_t \mathbb{E}\left[\|\delta_{t-1}\|^2 \right] \\
&\leq \mathbb{E}\left[\|\delta_{t-1}\|^2 \right] - \mathbb{E}\left[\|\delta_t\|^2 \right] + 2\beta_t \mathbb{E}\left[\|g_t - \nabla f(\theta_t)\|^2 \right] + \frac{4}{\beta_t} \mathbb{E}\left[\|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2 \right].
\end{aligned} \tag{48}$$

Summing up (48) for $t = 1, \ldots, T$, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \beta_{t+1} \mathbb{E}\left[\|\delta_t\|^2 \right] &= \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|\delta_{t-1}\|^2 \right] \\
&\leq \mathbb{E}\left[\|\delta_0\|^2 \right] + 4 \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2}{\beta_t} \right] \\
&\quad + 2 \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|g_t - \nabla f(\theta_t)\|^2 \right].
\end{aligned} \tag{49}$$

Note that

$$\begin{aligned}
&\epsilon \|\theta_t - \theta_{t+1}\|^2 \\
&= \langle \theta_t - \theta_{t+1},\ \epsilon I(\theta_t - \theta_{t+1}) \rangle \leq \langle \theta_t - \theta_{t+1},\ D_t(\theta_t - \theta_{t+1}) \rangle \\
&= \|\theta_t - \theta_{t+1}\|_{D_t}^2 = \left\| \alpha_t D_t^{-1} \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t}^2 = \alpha_t^2 \left\| \mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t) \right\|_{D_t^{-1}}^2.
\end{aligned} \tag{50}$$

Moreover, by (2), (6), (50) and $\frac{\alpha_t}{\epsilon} = c\beta_{t+1}$, we get

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2}{\beta_t}\right] \\
&\leq \sum_{t=1}^{T} L^2 \mathbb{E}\left[\frac{\|\theta_t - \theta_{t-1}\|^2}{\beta_t}\right] = \sum_{t=0}^{T-1} \frac{L^2}{\beta_{t+1}} \mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] \\
&\leq \sum_{t=0}^{T-1} \frac{L^2 \alpha_t^2}{\epsilon \beta_{t+1}} \mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right] \\
&= c^2 L^2 \epsilon \sum_{t=0}^{T-1} \beta_{t+1} \mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right].
\end{aligned} \tag{51}$$

Combing (47), (49), (51) and $\frac{\alpha_t}{\epsilon} = c\beta_{t+1}$, we get

$$\begin{aligned}
&\sum_{t=0}^{T-1} \alpha_t \mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right] \\
&\leq 4\left(\Phi(\theta_0) - \Phi^{\text{opt}}\right) + 2c \sum_{t=0}^{T-1} \beta_{t+1} \mathbb{E}\left[\|\delta_t\|^2\right] \\
&\leq 4\left(\Phi(\theta_0) - \Phi^{\text{opt}}\right) + 2c\mathbb{E}\left[\|\delta_0\|^2\right] + 4c \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|g_t - \nabla f(\theta_t)\|^2\right] \\
&\qquad + 8c \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\nabla f(\theta_t) - \nabla f(\theta_{t-1})\|^2}{\beta_t}\right] \\
&\leq 4\left(\Phi(\theta_0) - \Phi^{\text{opt}}\right) + 2c\mathbb{E}\left[\|\delta_0\|^2\right] + 4c \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|g_t - \nabla f(\theta_t)\|^2\right] \\
&\qquad + 8c^3 L^2 \epsilon \sum_{t=0}^{T-1} \beta_{t+1} \mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right].
\end{aligned} \tag{52}$$

By Lemma C.3 and Theorem 1, we get

$$\begin{aligned}
\mathbb{E}\|\delta_0\|^2 = \mathbb{E}\left[\|m_0 - \nabla f(\theta_0)\|^2\right] &= \mathbb{E}\left[\|\beta_0 g_0 - g_0 + g_0 - \nabla f(\theta_0)\|^2\right] \\
&\leq 2(1-\beta_0)^2 \mathbb{E}\left[\|g_0\|^2\right] + 2\mathbb{E}\left[\|g_0 - \nabla f(\theta_0)\|^2\right] \\
&\leq 2B' + \mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_0}\right).
\end{aligned} \tag{53}$$

Since $c \leq \frac{1}{4L}$ and $\frac{\alpha_t}{\epsilon} = c\beta_{t+1}$, we obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1} \alpha_t \mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right] \\
&\leq 8\left(\Phi(\theta_0) - \Phi^{\text{opt}}\right) + 4c\mathbb{E}\left[\|\delta_0\|^2\right] + 8c \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|g_t - \nabla f(\theta_t)\|^2\right].
\end{aligned} \tag{54}$$

This, combines with (53), (54), Lemma C.6 and Theorem 1, yields

$$\begin{aligned}
&\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \\
&\leq \frac{\sqrt{B'}+\epsilon}{\sum_{t=0}^{T-1}\alpha_t}\sum_{t=0}^{T-1}\alpha_t\mathbb{E}\left[\|\mathcal{G}^{D_t}(\theta_t, m_t, \alpha_t)\|^2_{D_t^{-1}}\right] \\
&\leq \left(\sqrt{B'}+\epsilon\right)\frac{8\left(\Phi(\theta_0)-\Phi^{\mathrm{opt}}\right)+4c\mathbb{E}\|\delta_0\|^2+8c\sum_{t=1}^{T}\beta_t\mathbb{E}\left[\|g_t-\nabla f(\theta_t)\|^2\right]}{\sum_{t=0}^{T-1}\alpha_t} \\
&\leq \left(\sqrt{B'}+\epsilon\right)\frac{8\left(\Phi(\theta_0)-\Phi^{\mathrm{opt}}\right)+8cB'+4c\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_0}\right)+8\sum_{t=1}^{T}\frac{\alpha_{t-1}}{\epsilon}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_t}\right)}{\sum_{t=0}^{T-1}\alpha_t} \\
&\leq \left(\sqrt{B'}+\epsilon\right)\frac{8\left(\Phi(\theta_0)-\Phi^{\mathrm{opt}}\right)+8cB'+4c\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_0}\right)+8\sum_{t=0}^{T-1}\frac{\alpha_t}{\epsilon}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n_{t+1}}\right)}{\sum_{t=0}^{T-1}\alpha_t}.
\end{aligned} \tag{55}$$

This completes the proof. □

C.3 Proof of Corollary 2

Proof. Given the total number of calls to the stochastic first-order oracle $\bar{N}$ and the number n of calls to the $\mathcal{SFO}$ at each iteration, our algorithm can perform at most $T = \lfloor \bar{N}/n \rfloor$ iterations. Obviously, $T \geq \bar{N}/(2n)$. By Corollary 1 we have

$$\begin{aligned}
&\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \\
&\leq \left(\sqrt{B'}+\epsilon\right)\left(\frac{2LD}{\epsilon T}+\frac{8cL}{\epsilon T}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n}\right)+\frac{8}{\epsilon}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n}\right)\right) \\
&\leq \frac{4L\left(\sqrt{B'}+\epsilon\right)Dn}{\epsilon\bar{N}}+\frac{16cL\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\bar{N}\gamma^\star}\mathcal{O}\left(1\right)+\frac{8\left(\sqrt{B'}+\epsilon\right)}{\epsilon}\mathcal{O}\left(\frac{\sigma^2}{\gamma^\star n}\right) \\
&\leq \frac{4L\left(\sqrt{B'}+\epsilon\right)D}{\epsilon}\mathcal{O}\left(\frac{1}{\sqrt{\bar{N}}}\right)+\frac{16cL\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\gamma^\star}\mathcal{O}\left(\frac{1}{\bar{N}}\right)+\frac{8\sigma^2\left(\sqrt{B'}+\epsilon\right)}{\epsilon\gamma^\star}\mathcal{O}\left(\frac{1}{\sqrt{\bar{N}}}\right) \\
&\leq \mathcal{O}\left(\frac{D_1}{\bar{N}}\right)+\mathcal{O}\left(\frac{D_2}{\sqrt{\bar{N}}}\right),
\end{aligned}$$

where $D_1 = \frac{16cL\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\gamma^\star}$, $D_2 = \frac{4L\left(\sqrt{B'}+\epsilon\right)D}{\epsilon}+\frac{8\sigma^2\left(\sqrt{B'}+\epsilon\right)}{\epsilon\gamma^\star}$ with D and B' being defined as those in Theorem 2. The proof is completed. □

C.4 Proof of Corollary 3

Proof. Given the total number of calls to the stochastic first-order oracle $\bar{N}$ and the number n of calls to the $\mathcal{SFO}$ at each iteration, our algorithm can perform at most $T = \lfloor \bar{N}/n \rfloor$ iterations. Obviously,

$T \geq \bar{N}/(2n)$. By Theorem 2 and $\sum_{t=0}^{T-1} \alpha_t = \mathcal{O}\left(\sqrt{nT}\right)$, we have

$$
\begin{aligned}
&\left(\mathbb{E}\left[\|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^{\frac{4}{3}}\right]\right)^{\frac{3}{2}} \\
&\leq \mathcal{O}\left(\frac{\left(\sqrt{B'}+\epsilon\right)D}{\sqrt{nT}}\right) + \mathcal{O}\left(\frac{4\left(\sqrt{B'}+\epsilon\right)c\sigma^2}{\gamma^\star n\sqrt{nT}}\right) + \mathcal{O}\left(\frac{8\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\gamma^\star n}\right) \\
&\leq \mathcal{O}\left(\frac{\sqrt{2}\left(\sqrt{B'}+\epsilon\right)D}{\sqrt{\bar{N}}}\right) + \mathcal{O}\left(\frac{4\sqrt{2}\left(\sqrt{B'}+\epsilon\right)c\sigma^2}{\gamma^\star n\sqrt{\bar{N}}}\right) + \mathcal{O}\left(\frac{8\left(\sqrt{B'}+\epsilon\right)\sigma^2}{\epsilon\gamma^\star n}\right) \\
&= \mathcal{O}\left(\frac{\left(\sqrt{B'}+\epsilon\right)\left(D\epsilon\gamma^\star+\sigma^2\right)}{\epsilon\gamma^\star\sqrt{\bar{N}}}\right) + \mathcal{O}\left(\frac{\left(\sqrt{B'}+\epsilon\right)c\sigma^2}{\gamma^\star\bar{N}}\right) \\
&= \mathcal{O}\left(\frac{D_3}{\bar{N}}\right) + \mathcal{O}\left(\frac{D_4}{\sqrt{\bar{N}}}\right),
\end{aligned}
$$

where $D_3 = \frac{\left(\sqrt{B'}+\epsilon\right)c\sigma^2}{\gamma^\star}, D_4 = \frac{\left(\sqrt{B'}+\epsilon\right)\left(D\epsilon\gamma^\star+\sigma^2\right)}{\epsilon\gamma^\star}$ with D and B' being defined as those in Theorem 2. The proof is completed. □

C.5 Proof of Corollary 4

Proof. Denote the right hand side of Corollary 2 as $C(T)$. Let $\zeta = \|\mathcal{G}^{D_R}(\theta_R, m_R, \alpha_R)\|^2$. We have $\mathbb{E}\left[|\zeta|^{\frac{2}{3}}\right] \leq C(T)^{\frac{2}{3}}$. By Chebyshev's inequality, we have

$$
\mathbb{P}\left(|\zeta|^{\frac{2}{3}} > \frac{C(T)^{\frac{2}{3}}}{\delta^{\frac{2}{3}}}\right) \leq \frac{\mathbb{E}\left[|\zeta|^{\frac{2}{3}}\right]}{C(T)^{\frac{2}{3}}/\delta^{\frac{2}{3}}} \leq \delta^{\frac{2}{3}}.
$$

Namely, $\mathbb{P}\left(|\zeta| > \frac{C(T)}{\delta}\right) \leq \delta^{\frac{2}{3}}$. Therefore, $\mathbb{P}\left(|\zeta| \leq \frac{C(T)}{\delta}\right) \geq 1-\delta^{\frac{2}{3}}$. This completes the proof. □

D Additional Experimental Details and Results

D.1 Experimental Setup and Hyperparameters of Table 2

Experiments on MNIST were conducted on an NVIDIA RTX 3090 GPU (approx. 1 hour for 55 epochs), while CIFAR-10 and CelebA tasks utilized RTX 4090 (approx. 10 hours for 100 epochs and 30 hours for 90 epochs, respectively).

MCMC Sampling: We set the step size $\eta = 10$ and $\sigma_t = 0.005$ in (5). The number of steps k is set to 60 and the batch size n is set to 128.
MNIST dataset: We set $\alpha_t = \frac{\alpha}{\sqrt{t}}$, $\alpha = 0.015$, $\beta_t = 0.1$ for $t = 1$ and $\beta_t = \min(1 - \frac{\alpha}{\sqrt{t-1}\cdot\epsilon}, 0.1)$ for $t > 1$, $\rho_t = \frac{0.001+0.999r}{t^r}$, $r = 0.75$, $\epsilon =$ 1e-8 for both cases with and without L_1 regularization. The L_1 regularization parameter $\lambda_{\text{reg}} =$ 1e-9.

For experiments on the CIFAR-10 and CelebA datasets, α_t, β_t, and ρ_t are set following Du & Mordatch (2019).
CIFAR-10 dataset: We set $\alpha_t =$ 1e-4, $\beta_t = 1$, $\rho_t = 0.001$, and $\epsilon =$ 1e-8 for both cases with and without L_1 regularization. The L_1 regularization parameter $\lambda_{\text{reg}} =$ 4e-11.
CelebA dataset: We set $\alpha_t =$ 1e-4, $\beta_t = 1$, $\rho_t = 0.001$, and $\epsilon =$ 1e-8 for both cases with and without L_1 regularization. The L_1 regularization parameter $\lambda_{\text{reg}} =$ 1.5e-10.

D.2 Explanation for Selecting the Regularization Parameter λ_{reg}

We evaluated candidate values for λ_{reg} ranging from 1e-12 to 1e-9 (e.g., 1e-9, 1.5e-10, 4e-11, 5e-11, 1e-12) and selected the value that yielded the lowest FID score for the computed solution. The optimal

choice of $\lambda_{\text{reg}} = 4\text{e-}11$ is relatively small, which is naturally determined by the relative magnitudes of the data-fitting term $f(\theta)$ and the regularization term $h(\theta)$ in our composite optimization framework (1).

Table 4: Proximal operator statistics ($\lambda_{\text{reg}} = 4\text{e-}11$ vs 0)

Metric	Batch 0	Batch 50	Batch 100
Mean(v_t)	5.38e-15	2.68e-13	5.17e-13
Threshold (max/mean)	4e-7/2.16e-7	3.65e-7/2.19e-8	3.42e-7/1.81e-8
Weight mag. (min/mean)	2.18e-13/9.56e-3	1.79e-10/9.79e-3	1.75e-10/9.90e-3
# Zeros ($\lambda_{\text{reg}} = 4\text{e-}11$)	6590	114	80
# Zeros ($\lambda_{\text{reg}} = 0$)	0	0	0

To clarify the active mechanism of the proximal operator under a small regularization parameter ($\lambda_{\text{reg}} = 4\text{e-}11$), Table 4 summarizes the operator statistics at early iteration steps (Batch 0, 50, and 100), where # Zeros indicates the number of weight variables newly truncated to exact zero at each step. As shown in Table 4, despite the small magnitude of λ_{reg}, the proximal operator actively sets 6,590 weights to zero at the initial step (Batch 0) and continues to truncate 114 and 80 weights at Batches 50 and 100, respectively. In contrast, standard optimization without L_1 regularization ($\lambda_{\text{reg}} = 0$) produces zero sparse entries. These statistics demonstrate that the proximal operator is not numerically identity and effectively modifies the optimization trajectory.

D.3 Failures of PPG and MCMC-SGD as baselines

We evaluate two extra baselines: PPG (for the L_1-regularized model (14) with learning rate 1e-4 and regularization strength 1e-9) and MCMC-SGD (for the unregularized model (13) with learning rate 1e-4). As shown in Table 5, both baselines suffer from severe model collapse early in the training process of the EBM on the MNIST dataset; this explains why they are excluded as primary baselines in Table 2 of Section 5. In addition, we also experimented with alternative learning rates of 1e-3, 1e-5, and 1e-6. These configurations yield results similar to those of 1e-4, where both algorithms similarly suffer from severe model collapse during the initial training phase. FID scores are evaluated using 60,000 generated images and 60,000 real images.

Table 5: FID of PPG and MCMC-SGD along epochs.

Epoch	PPG(Atchadé et al., 2017)	MCMC-SGD(Li et al., 2023)
5	399.37 ± 1.18	399.92 ± 1.70
10	399.58 ± 1.35	400.26 ± 2.12
15	399.63 ± 1.46	400.89 ± 2.49
20	399.78 ± 1.68	401.46 ± 2.93
25	399.88 ± 1.62	402.19 ± 3.76
30	400.13 ± 1.79	402.92 ± 4.47
35	400.18 ± 1.78	403.82 ± 5.41
40	400.38 ± 1.86	404.69 ± 6.53
45	400.48 ± 2.19	405.75 ± 7.81
50	400.62 ± 2.05	406.91 ± 9.28
55	400.84 ± 2.30	408.03 ± 10.7

D.4 Ablation Experiments

On the MNIST dataset, we evaluated MCMC-ProxAdam with both the dynamic forms of $\alpha_t, \beta_t, \rho_t$ defined in Appendix D.1 and their fixed-value counterparts. The results are summarized in Table 6. These empirical results clearly indicate that the dynamic parameter setting achieves faster convergence compared to fixed parameters, thereby justifying our choice of dynamic hyperparameter schedules in Table 2. FID scores are evaluated using 60,000 generated images and 60,000 real images.

Furthermore, we conducted a similar ablation study for MCMC-SubAdam by comparing the dynamic parameter schedule in Appendix D.1 against its fixed-parameter counterpart, with the results summarized in Table 7. The empirical evidence consistently indicates that the dynamic parameter

Table 6: Ablation experiments for MCMC-Adam on model without L_1 regularization, using different settings of momentum parameter β_t ($\alpha_t = 1\text{e-}4$, $\rho_t = 0.001$, $\epsilon = 1\text{e-}8$), and comparison with our MCMC-ProxAdam for L_1-regularized model.

Epoch	MCMC-ProxAdam (Dynamic) $\lambda_{\text{reg}} = 10^{-9}$	MCMC-Adam $\beta_t = 1$ $\lambda_{\text{reg}} = 0$	MCMC-Adam $\beta_t = 0.9$ $\lambda_{\text{reg}} = 0$	MCMC-Adam $\beta_t = 0.5$ $\lambda_{\text{reg}} = 0$	MCMC-Adam $\beta_t = 0.1$ $\lambda_{\text{reg}} = 0$	MCMC-Adam (Dynamic) $\lambda_{\text{reg}} = 0$
5	59.73 ± 8.31	117.47 ± 21.69	115.87 ± 22.50	114.34 ± 21.67	108.10 ± 6.20	59.72 ± 10.68
10	56.04 ± 13.42	111.89 ± 20.37	108.44 ± 16.09	104.89 ± 11.17	88.14 ± 7.80	49.99 ± 5.47
15	46.47 ± 5.79	89.48 ± 7.52	91.67 ± 4.69	94.89 ± 6.18	104.96 ± 43.02	44.75 ± 3.63
20	39.34 ± 5.43	72.84 ± 9.90	76.35 ± 14.25	71.35 ± 2.60	87.08 ± 17.78	41.94 ± 3.72
25	35.49 ± 8.08	64.64 ± 4.89	64.12 ± 5.57	64.33 ± 9.68	65.73 ± 10.15	35.14 ± 7.80
30	31.25 ± 6.15	62.52 ± 11.26	63.30 ± 7.34	64.31 ± 4.10	63.78 ± 11.15	35.15 ± 3.68
35	31.07 ± 4.63	53.56 ± 4.62	53.08 ± 6.92	49.43 ± 8.12	51.65 ± 5.67	33.16 ± 2.46
40	33.11 ± 8.41	49.41 ± 5.97	50.12 ± 6.12	49.49 ± 7.50	59.37 ± 11.44	36.10 ± 9.80
45	33.43 ± 3.64	55.41 ± 10.51	54.06 ± 10.49	52.52 ± 8.95	56.10 ± 8.05	31.68 ± 3.90
50	29.34 ± 2.23	50.16 ± 7.15	51.86 ± 7.06	50.86 ± 4.42	49.40 ± 6.27	29.84 ± 3.35
55	25.99 ± 1.85	47.39 ± 14.45	49.33 ± 16.93	50.39 ± 13.36	45.85 ± 10.44	27.08 ± 2.92

schedule accelerates algorithmic convergence. Notably, since MCMC-SubAdam is also applied to solve the L_1-regularized problem (14), contrasting our MCMC-ProxAdam with MCMC-SubAdam can additionally serve as an ablation study of the proximal operator. FID scores are evaluated using $60{,}000$ generated images and $60{,}000$ real images.

Table 7: Comparison of our MCMC-ProxAdam with MCMC-SubAdam (abbreviated as SubAdam below; $\alpha_t = 1\text{e-}4$, $\rho_t = 0.001$, $\epsilon = 1\text{e-}8$) on the same L_1 model with $\lambda_{\text{reg}} = 10^{-9}$, together with ablation experiments on momentum parameter β_t and proximal operator.

Epoch	MCMC-ProxAdam (Dynamic)	SubAdam $\beta_t = 1$	SubAdam $\beta_t = 0.9$	SubAdam $\beta_t = 0.5$	SubAdam $\beta_t = 0.1$	SubAdam (Dynamic)
5	59.73 ± 8.31	118.01 ± 22.34	115.92 ± 23.04	115.11 ± 22.49	108.14 ± 6.20	61.74 ± 9.35
10	56.04 ± 13.42	112.78 ± 20.44	107.85 ± 13.73	105.98 ± 11.11	88.24 ± 7.86	52.56 ± 7.59
15	46.47 ± 5.79	90.42 ± 7.28	92.25 ± 5.51	94.87 ± 6.56	104.87 ± 43.36	43.62 ± 8.26
20	39.34 ± 5.43	73.14 ± 10.16	76.29 ± 14.34	72.10 ± 2.78	87.08 ± 16.59	41.35 ± 4.42
25	35.49 ± 8.08	63.75 ± 5.15	63.41 ± 5.66	62.45 ± 8.09	64.94 ± 9.58	35.08 ± 7.85
30	31.25 ± 6.15	64.75 ± 13.15	64.77 ± 8.17	63.35 ± 5.10	62.39 ± 6.88	31.15 ± 4.50
35	31.07 ± 4.63	51.11 ± 4.17	52.96 ± 4.46	53.00 ± 7.39	51.82 ± 4.01	34.28 ± 3.54
40	33.11 ± 8.41	47.96 ± 6.17	50.68 ± 6.65	52.07 ± 8.05	55.84 ± 9.06	34.79 ± 4.29
45	33.43 ± 3.64	55.28 ± 8.36	54.84 ± 9.95	52.57 ± 9.28	49.82 ± 9.45	30.79 ± 5.66
50	29.34 ± 2.23	51.84 ± 7.02	52.23 ± 8.00	52.07 ± 3.91	53.34 ± 4.81	30.78 ± 3.19
55	25.99 ± 1.85	48.18 ± 17.39	48.73 ± 15.57	46.50 ± 12.16	45.38 ± 13.62	28.16 ± 2.10

D.5 Sensitivity Study

On the MNIST dataset, we conduct a sensitivity analysis for MCMC-ProxAdam with respect to the burn-in n_b and the spectral gap of the underlying chain. The algorithm is evaluated under the dynamic parameter schedule, configured identically to the parameter setup inAppendix D.1. Since the spectral gap is generally unavailable in practice, we use different Langevin noise levels σ_t as an empirical sensitivity analysis rather than as a direct estimator or proxy of the spectral gap. FID scores are evaluated using $60{,}000$ generated images and $60{,}000$ real images.

As shown in Table 8 and Table 9, as the number of epochs increases, the mean FID significantly decreases across all parameter configurations, and the standard deviations shrink, indicating robust convergence and stability.

According to Table 8, the MCMC-ProxAdam algorithm demonstrates robust convergence on the L_1 model ($\lambda_{\text{reg}} = 1\text{e-}9$) with respect to the batch size n_b. As the number of epochs increases from 5 to 55, the mean errors for all n_b values (ranging from 59 to 99) decrease significantly, accompanied by a general reduction in standard deviations, indicating stable convergence. Specifically, in the early training stages (Epochs 5-15), smaller n_b values (e.g., 59 and 69) yield lower mean errors and faster convergence rates. In the later stages (Epochs 45-55), the performance gap among different

Table 8: Sensitivity analysis of MCMC-ProxAdam with respect to n_b on L_1 model with λ_{reg} = 1e-9.

Epoch	n_b=59	n_b=69	n_b=79	n_b=89	n_b=99
5	59.73 ± 8.31	66.86 ± 13.20	66.43 ± 12.24	70.21 ± 10.94	71.74 ± 16.43
10	56.04 ± 13.42	59.46 ± 11.04	57.35 ± 6.67	62.15 ± 8.41	65.90 ± 11.52
15	46.47 ± 5.79	45.87 ± 6.00	47.06 ± 4.19	50.14 ± 11.17	50.36 ± 9.23
20	39.34 ± 5.43	42.60 ± 5.24	41.90 ± 8.58	43.17 ± 7.01	43.42 ± 3.80
25	35.49 ± 8.08	36.14 ± 6.95	39.89 ± 10.68	42.94 ± 7.58	40.58 ± 6.39
30	31.25 ± 6.15	35.67 ± 4.16	35.60 ± 5.43	39.16 ± 5.19	40.35 ± 6.48
35	31.07 ± 4.63	32.92 ± 2.20	37.39 ± 5.31	40.66 ± 2.44	38.32 ± 3.68
40	33.11 ± 8.41	35.02 ± 7.85	36.62 ± 3.95	39.45 ± 4.65	41.38 ± 7.47
45	33.43 ± 3.64	33.18 ± 4.24	34.78 ± 4.80	36.85 ± 2.71	38.54 ± 4.73
50	29.34 ± 2.23	32.12 ± 1.87	32.08 ± 1.72	36.13 ± 3.85	36.19 ± 1.36
55	25.99 ± 1.85	30.90 ± 3.99	31.74 ± 3.12	31.63 ± 2.89	33.81 ± 3.16

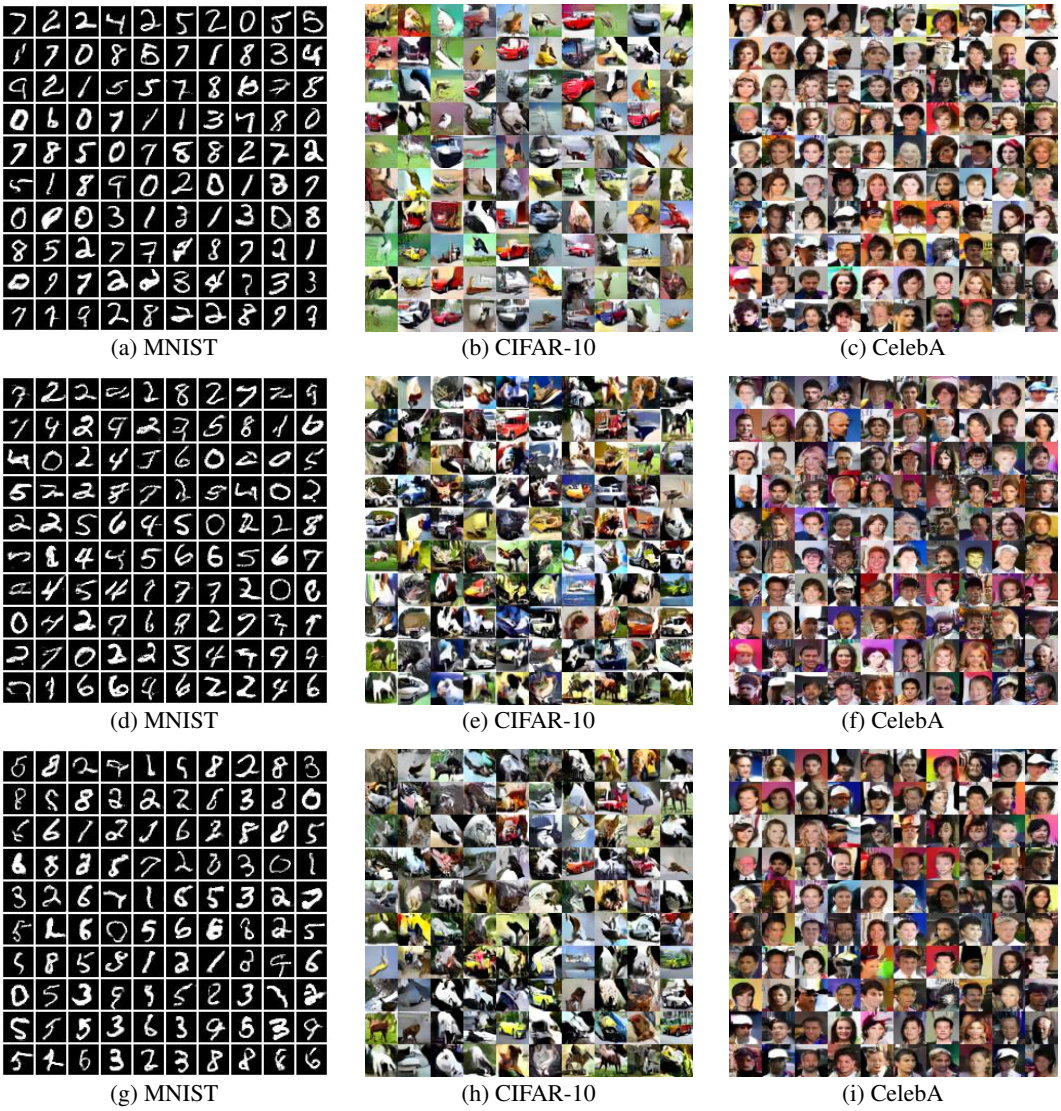
Table 9: Sensitivity analysis of MCMC-ProxAdam with respect to σ_t on L_1 model with λ_{reg} = 1e-9.

Epoch	$\sigma = 0.001$	$\sigma = 0.003$	$\sigma = 0.005$	$\sigma = 0.007$	$\sigma = 0.01$
5	61.17 ± 7.78	60.89 ± 12.09	59.73 ± 8.31	57.92 ± 8.76	59.15 ± 10.80
10	58.58 ± 5.95	55.32 ± 11.96	56.04 ± 13.42	51.64 ± 5.46	48.69 ± 5.05
15	46.35 ± 12.07	44.65 ± 5.35	46.47 ± 5.79	43.78 ± 6.64	41.92 ± 7.14
20	41.39 ± 7.97	42.74 ± 7.06	39.34 ± 5.43	38.62 ± 5.44	39.67 ± 6.15
25	42.07 ± 10.20	36.99 ± 2.33	35.49 ± 8.08	33.24 ± 5.72	34.70 ± 7.87
30	39.62 ± 7.00	34.38 ± 6.87	31.25 ± 6.15	34.18 ± 5.75	32.17 ± 6.12
35	39.07 ± 8.19	33.28 ± 3.23	31.07 ± 4.63	32.58 ± 5.19	32.85 ± 2.57
40	40.60 ± 9.04	35.83 ± 7.89	33.11 ± 8.41	33.31 ± 4.98	34.48 ± 8.14
45	35.16 ± 7.37	32.52 ± 3.93	33.43 ± 3.64	31.62 ± 5.93	34.02 ± 6.10
50	33.17 ± 1.52	29.81 ± 4.48	29.34 ± 2.23	30.60 ± 2.91	31.69 ± 3.27
55	29.53 ± 3.02	27.41 ± 1.61	25.99 ± 1.85	27.00 ± 0.84	29.79 ± 4.36

configurations narrows. Overall, while $n_b = 59$ achieves the best final error performance (25.99 ± 1.85), the model's sensitivity to n_b regarding final accuracy is relatively limited.

According to Table 9, the algorithm exhibits certain sensitivity to the parameter σ_t. As epochs increase, the mean errors for all σ values decrease significantly. In the early training stages (Epochs 5-10), larger σ values (e.g., 0.007 and 0.01) show faster initial descent. However, in the later stages (Epoch 55), a moderate $\sigma = 0.005$ achieves the best final performance (25.99 ± 1.85). In contrast, a very small σ (0.001) converges relatively slowly (29.53 ± 3.02), while a very large σ (0.01) results in higher final mean errors and greater fluctuation (29.79 ± 4.36). Therefore, a σ value around 0.005 is the optimal choice for balancing convergence speed and final accuracy in this configuration.

D.6 GENERATED SAMPLES



(a) MNIST (b) CIFAR-10 (c) CelebA

(d) MNIST (e) CIFAR-10 (f) CelebA

(g) MNIST (h) CIFAR-10 (i) CelebA

Figure 1: Corresponding generative samples for EBMs with different optimizers across three datasets: (a)-(c) samples for MCMC-ProxAdam; (d)-(f) samples for Adam; (g)-(i) samples for Sub-Adam.