Statistical Inference for the Population Landscape via Moment Adjusted Stochastic Gradients

Tengyuan Liang[*]

University of Chicago

Weijie Su[†]

University of Pennsylvania

Abstract. Modern statistical inference tasks often require iterative optimization methods to approximate the solution. Convergence analysis from optimization only tells us how well we are approximating the solution deterministically, but overlooks the sampling nature of the data. However, due to the randomness in the data, statisticians are keen to provide uncertainty quantification, or confidence, for the answer obtained after certain steps of optimization. Therefore, it is important yet challenging to understand the sampling distribution of the iterative optimization methods.

This paper makes some progress along this direction by introducing a new stochastic optimization method for statistical inference, the moment adjusted stochastic gradient descent. We establish non-asymptotic theory that characterizes the statistical distribution of the iterative methods, with good optimization guarantee. On the statistical front, the theory allows for model misspecification, with very mild conditions on the data. For optimization, the theory is flexible for both the convex and non-convex cases. Remarkably, the moment adjusting idea motivated from "error standardization" in statistics achieves similar effect as Nesterov's acceleration in optimization, for certain convex problems as in fitting generalized linear models. We also demonstrate this acceleration effect in the non-convex setting through experiments.

Key words and phrases: Non-asymptotic inference, discretized Langevin algorithm, stochastic gradient methods, Nesterov's acceleration, model misspecification, population landscape.

1. INTRODUCTION

Statisticians are interested in inferring properties about a population, based on independently sampled data. In the parametric regime, the inference problem boils down to constructing point estimates and confidence intervals for a finte number of unknown parameters. When the data-generation process is well-specified by the parametric family, elegant asymptotic theory has been

[*](e-mail: tengyuan.liang@chicagobooth.edu)

[†](e-mail: suw@wharton.upenn.edu)

established for maximum likelihood estimation (MLE) credited to Ronald Fisher in 1920s. This asymptotic theory is readily generalizable to the model mis-specification setting, for a properly-chosen risk function $\ell(\theta; z)$[1] and the corresponding empirical risk minimizer (ERM)

$$\hat{\theta}_{\mathrm{ERM}} \triangleq \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \ell(\theta; z_i) \qquad \text{empirical risk minimizer,}$$
$$\theta_* \triangleq \arg\min_{\theta} \mathop{\mathbb{E}}_{\mathbf{z}\sim P} \ell(\theta; \mathbf{z}) \qquad \text{population minimizer,}$$

with

$$\sqrt{N}\left(\hat{\theta}_{\mathrm{ERM}} - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{H}(\theta_*)^{-1}\mathbf{\Sigma}(\theta_*)\mathbf{H}(\theta_*)^{-1}\right).$$

Here θ is the parameter of the model, z_i's are i.i.d copies from the unknown distribution $\mathbf{z} \sim P$, $\mathbf{H}(\theta) \triangleq \mathbb{E}\,\mathrm{Hess}\left[\ell(\theta; \mathbf{z})\right]$, and $\mathbf{\Sigma}(\theta) \triangleq \mathbb{E}\left[\nabla_\theta \ell(\theta; \mathbf{z}) \otimes \nabla_\theta \ell(\theta; \mathbf{z})\right]$. Define the *population landscape* $L(\theta)$ as

$$L(\theta) \triangleq \mathop{\mathbb{E}}_{\mathbf{z}\sim P} \ell(\theta; \mathbf{z}). \tag{1.1}$$

One should notice that the elegant statistical theory for inference holds under rather mild regularity conditions, without requiring $L(\theta)$ being convex. However, it overlooks one important aspect: the optimization difficulty of the landscape on θ.

Optimization techniques are required to solve for the above estimators $\hat{\theta}$, as they rarely take closed-form. Global convergence and computational complexity is only well-understood when the sample analog $\frac{1}{N}\sum_{i=1}^{N} \ell(\theta; z_i)$ is convex. The optimization is done iteratively

$$\theta_{t+1} = \theta_t - \eta \mathbf{h}(\theta_t), \tag{1.2}$$

where $\mathbf{h}$ is based on the first and/or second order information, η is step-size. For the non-convex case, the convergence becomes less clear, but in practice people still employ iterative methods. Nevertheless, in either case, the available convergence results fall in short of the statistical aspect: after certain t iterations, one is interested in knowing the sampling distribution of θ_t, for uncertainty quantification of the optimization algorithm.

The goal of the present work is to combine the strength of the two worlds in inference and optimization: to characterize the statistical distribution of the iterative methods, with good optimization guarantee. Specifically, we study particular stochastic optimization methods for the (possibly non-convex) population landscape $L(\theta)$, and at the same time characterize the sampling distribution at each step, through establishing a non-asymptotic theory. We allow for model mis-specification, and require only mild moment conditions on the data generating process.

1.1 Motivation

Observe the simple fact that what one actually aims to optimize is the population objective $L(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\,\ell(\theta; \mathbf{z})$, not the sample version. Therefore, stochastic approximation pioneered by

[1] It is also called loss function in the statistical learning literature. In generalized methods of moment, $\mathbb{E}_{\mathbf{z}\sim P}\nabla_\theta \ell(\theta; \mathbf{z}) = 0$ is also called moment condition. The MLE can be also viewed as a special case with $\ell(\theta; \mathbf{z}) = -\log p_\theta(\mathbf{z})$ and the data-generation process being $P = P_{\theta_*}$.

Robbins and Monro [1951], Kiefer and Wolfowitz [1952] stands out as a natural optimization approach for the statistical inference problem. In modern practice, *Stochastic Gradient Descent* (SGD) with mini-batches of size n is widely used,

$$\theta_{t+1} = \theta_t - \eta \hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}), \tag{1.3}$$

where $\hat{\mathbb{E}}_n$ is the empirical expectation over n independently sampled mini-batch data at each step.

Our first motivation follows from the intuition that one can approximate the above step when n is not too small, which we will make rigorous in a moment. Define

$$\mathbf{b}(\theta) = \mathop{\mathbb{E}}_{\mathbf{z}\sim P} \nabla_\theta \ell(\theta, \mathbf{z}), \tag{1.4}$$

$$\mathbf{V}(\theta) = \{\mathrm{Cov}[\nabla_\theta \ell(\theta, \mathbf{z})]\}^{1/2}, \tag{1.5}$$

then observe the following approximation

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}) \\ &= \theta_t - \eta\, \mathbb{E}\, \nabla_\theta \ell(\theta_t, \mathbf{z}) + \eta \left[\mathbb{E}\, \nabla_\theta \ell(\theta_t, \mathbf{z}) - \hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}) \right] \\ &\approx \theta_t - \eta \mathbf{b}(\theta_t) + \sqrt{2\beta^{-1}\eta}\mathbf{V}(\theta_t)\mathbf{g}_t, \quad \text{with } \beta \triangleq \frac{2n}{\eta}, \end{aligned} \tag{1.6}$$

where $\mathbf{g}_t, t \geq 0$ are independent isotropic Gaussian vectors. The combination of n, η provides a stronger approximation guarantee at each iteration for large n, in contrast to the asymptotic normal approximation for the average of trajectory in Polyak and Juditsky [1992] as $t \to \infty$. The β^{-1} quantifies the "variance" injected each step (due to sampled mini-batches), or the "temperature" parameter: the larger the β is, the closer the distribution is concentrated near the deterministic steepest gradient descent updates. The scaling of the step-size η relates to Cauchy discretization of the Itô diffusion process (as $\eta \to 0$)

$$d\theta_t = -\mathbf{b}(\theta_t)dt + \sqrt{2\beta^{-1}}\mathbf{V}(\theta_t)dB_t.$$

Our second motivation comes from a classic "standardization" idea in statistics — we want to adjust the stochastic gradient vector at step t by $\mathbf{V}(\theta_t)$ so that the conditional noise (conditioned on θ_t) for each coordinate is independent and on the same scale,

$$\begin{aligned} \theta_{t+1} &= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}) \\ &\approx \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \mathbf{b}(\theta_t) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_t. \end{aligned} \tag{1.7}$$

This standardization trick is similar to the classic method of inverse-variance weighting, though with notable difference. The similarity lies in the fact that noisier gradient information is weighted less in both approaches and vice versa. However, the former scales weights proportional to inverse standard deviation, while the latter uses inverse standard variance instead.

To answer the inference question about $L(\theta)$ using the "moment adjusted" iterative method (1.7), one needs to know the sampling distribution of θ_t for a fixed t. One hopes to directly describe the distribution in a non-asymptotic fashion, instead of characterizing this distribution either through the asymptotic normal limit [Polyak and Juditsky, 1992] (passing over data once at a time) in the convex senario, or through the invariant distribution which could in theory take exponential time to converge for general non-convex $L(\theta)$ [Raginsky et al., 2017]. One thing to notice is that, at a fixed

time t, the distribution is distinct from Gaussian, for general $\mathbf{b}$ and $\mathbf{V}$. From an optimization angle, one would like the iterative algorithm to converge (to a local optima) fast. This is also important for the purpose of inference: given the distribution can be approximately characterized at each step, one hopes that the distribution will concentrate near a local minimum of the population landscape $L(\theta)$ within a reasonable time budget, before the error accumulates in the stochastic process and invalidates the approximation.

1.2 Contributions

We propose the *Moment Adjusted Stochastic Gradient descent* (MasGrad), an iterative optimization method that infers about the stationary points of the population landscape $L(\theta)$, namely $\{\theta : \|\nabla L(\theta)\| = 0\}$. The MasGrad is a simple variant of SGD that adjusts the descent direction using the square root of the covariance matrix $\mathbf{V}(\theta_t)$ (defined in (1.5)) of the gradient at the current location,

$$\text{MasGrad}: \quad \theta_{t+1} = \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}).$$

We summarize our main contributions in two perspectives. Other extensions will be discussed later along the paper. During the discussion, we use $\mathcal{O}_{\epsilon,\delta}(\cdot)$ to denote the order of magnitude for parameters ϵ, δ only, treating others as constants.

Inference. The distribution of MasGrad updates θ_t, with n independently sampled mini-batch data at each step, can be characterized in a non-asymptotic fashion. Informally, for any data-generating distribution $\mathbf{z} \sim P$ under mild conditions, the distribution of θ_t — denoted as $\mu(\theta_t)$ — satisfies,

$$D_{\text{TV}}(\mu(\theta_t), \nu_{t,\eta}) \le C\sqrt{\frac{t}{n}} \quad \Rightarrow \quad \mu(\theta_t) \xrightarrow{\mathcal{L}} \nu_{t,\eta}, \text{ converge in distribution as } n \to \infty.$$

Here $\nu_{t,\eta}$ is the distribution of ξ_t that follows the update initialized with $\xi_0 = \theta_0$

$$\xi_{t+1} = \xi_t - \eta \mathbf{V}(\xi_t)^{-1} \mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_t, \ \mathbf{g}_t \sim \mathcal{N}(0, \mathbf{I}_p) \text{ and } \beta = \frac{2n}{\eta}. \tag{1.8}$$

Remark that $\nu_{t,\eta}$ only depends on t, η, and the first and second moments $\mathbf{b}, \mathbf{V}$ of $\nabla \ell(\theta, \mathbf{z})$, regardless of the specific the data-generating distribution $\mathbf{z} \sim P$. The rigorous statement is deferred to Thm. 3.1, and further extensions to the continuous time analog are discussed in Section 3.

Optimization. Interestingly, in the strongly convex case such as in generalized linear models (GLMs), the "standardization" idea achieves the Nesterov's acceleration [Nesterov, 1983, 2013]. Informally, the number of iterations for an ϵ-minimizer for gradient descent requires

$$T_{\text{GD}} = \mathcal{O}_{\epsilon,\kappa}\left(\kappa \log \frac{1}{\epsilon}\right), \quad \text{for some } \kappa > 1.$$

We show that for GLMs under mild conditions, MasGrad reduces the number of iterations to

$$T_{\text{MasGrad}} = \mathcal{O}_{\epsilon,\kappa}\left(\sqrt{\kappa} \log \frac{1}{\epsilon}\right),$$

which matches Nesterov's acceleration in the strongly convex case. The formal statement is deferred to Section 4.

Combining the inference and optimization theory together, we present informally the results for both the *convex* and *non-convex* cases. Recall that $\theta \in \mathbb{R}^p$.

Convex. In the strongly convex case, MasGrad with a properly chosen step-size and the following choice of parameters

$$T = \mathcal{O}_\epsilon\left(\log\frac{1}{\epsilon}\right), \text{ and } \quad n = \mathcal{O}_{\epsilon,p}\left(\frac{p}{\epsilon}\right),$$

satisfies

$$\begin{aligned}
&\text{inference}: \quad D_{\mathrm{TV}}\left(\mu(\theta_T), \mu(\xi_T)\right) \le \mathcal{O}_\epsilon\left(\sqrt{\epsilon \log 1/\epsilon}\right), \\
&\text{optimization}: \quad \mathbb{E}\, L(\theta_T) - \min_\theta L(\theta) \le \epsilon, \ \ \mathbb{E}\, L(\xi_T) - \min_\theta L(\theta) \le \epsilon, \quad \text{where } \xi_T \sim \nu_{T,\eta}.
\end{aligned}$$

The formal result is stated in Thm. 4.1.

Non-convex. Under mild smoothness condition, MasGrad with a proper step-size and the following choice of parameters

$$T = \mathcal{O}_{\epsilon,\delta,p}\left(\frac{1 \vee p\delta^2}{\epsilon^2}\right), \text{ and } \quad n = \mathcal{O}_{\epsilon,\delta,p}\left(\frac{\delta^{-2} \vee p}{\epsilon^2}\right),$$

satisfies

$$\begin{aligned}
&\text{inference}: \quad D_{\mathrm{TV}}\left(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])\right) \le \mathcal{O}_\delta(\delta), \\
&\text{optimization}: \quad \mathbb{E} \min_{t \le T} \|\nabla L(\theta_t)\| \le \epsilon, \ \ \mathbb{E} \min_{t \le T} \|\nabla L(\xi_t)\| \le \epsilon, \quad \text{where } \xi_t \sim \nu_{t,\eta}, \text{ for } t \in [T],
\end{aligned}$$

where the evolution of ξ_t is defined in (1.8). The formal result is deferred to Thm. 5.1.

2. RELATIONS TO THE LITERATURE

In the case of a differentiable convex L, finding a minimum of this function is equivalent to solving $\nabla L(\theta) = 0$ for θ. This simple equivalence reveals that the vanilla SGD, which takes the form[2]

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \ell(\theta_t, z_t), \tag{2.1}$$

is an instance of stochastic approximation methods. This class of methods are iterative algorithms that attempt to solve fixed point equations (for example, $\nabla L(\theta) = 0$) provided noisy observations (for example, $\nabla_\theta \ell(\theta_t, z_t)$) [Robbins and Monro, 1951, Kiefer and Wolfowitz, 1952]. Using slowly diminishing step-sizes $\eta_t = O(1/t^\alpha)$ $(0 < \alpha < 1)$, Ruppert [1988] and Polyak [1990] showed that the average $\frac{1}{t}\sum_{i=1}^t \theta_i$ over trajectories of this recursive stochastic approximation algorithm attains optimal acceleration of convergence rate for a strongly convex L (see Polyak and Juditsky [1992] for more details).

In a different route, a fruitful line of research has focused on how to improve asymptotic convergence rate as $t \to \infty$ through pre-conditioning, a technique that involves approximating the unknown Hessian $\mathbf{H}(\theta) = \nabla^2 L(\theta)$ near the optimum θ^* (see, for instance, Bordes et al. [2009] and references therein). A popular example as such is AdaGrad [Duchi et al., 2011], which is a variant

[2]Recognize that $\nabla_\theta \ell(\theta_t, z_t)$ is an unbiased estimate of the population gradient as $\nabla_\theta L(\theta_t) = \mathbb{E}_{\mathbf{z}\sim P}[\nabla_\theta \ell(\theta_t, \mathbf{z})]$.

of SGD that adaptively determines learning rates for different coordinates by incorporating the geometric information of past iterates. In its simplest form, AdaGrad records previous gradient information through

$$G_t = \sum_{i=1}^{t} \nabla \ell(\theta_i, z_i) \otimes \nabla \ell(\theta_i, z_i),$$

and this procedure then updates iterates according to

$$\theta_{t+1} = \theta_t - \gamma G_t^{-\frac{1}{2}} \nabla \ell(\theta_t, z_t),$$

where $\gamma > 0$ is fixed. In large-scale learning tasks, evaluating $G_t^{-\frac{1}{2}}$ is computationally prohibitive and thus is often suggested to use $\mathrm{diag}(G_t)^{-\frac{1}{2}}$ instead. It should be, however, noted that the theoretical derivation of AdaGrad considers $G_t^{-\frac{1}{2}}$. AdaGrad is a flexible improvement on SGD and can easily extend to non-smooth optimization and non-Euclidean optimization such as mirror descent. With the geometric structure G_t learned from past gradients, AdaGrad assigns different learning rates to different components of the parameter, allowing infrequent features to take relatively larger learning rates. This adjustment is shown to speed up convergence dramatically in a wide range of empirical problems [Pennington et al., 2014].

Another closely related method is the natural gradient [Amari, 1998, 2012] raised first in the information geometry literature. When the parameter space enjoys certain structure, it has been shown that natural gradient outperforms the classic gradient descent both theoretically and empirically. To adapt the natural gradient to our setting, we relate the loss function to a generative model $\ell(\theta, z) = -\log p_\theta(z)$. The Riemannian structure of the parameter space (manifold) of the statistical model is defined by the Fisher information

$$\mathbf{I}(\theta) = \mathop{\mathbb{E}}_{\mathbf{z} \sim P} \left[\nabla_\theta \ell(\theta, \mathbf{z}) \otimes \nabla_\theta \ell(\theta, \mathbf{z}) \right].$$

The natural gradient can be viewed as the steepest descent induced by the Riemannian metric

$$\begin{aligned} \theta_{t+1} &= \arg\min_\theta \left[L(\theta_t) + \langle \nabla_\theta L(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta_t} \|\theta - \theta_t\|^2_{\mathbf{I}(\theta_t)} \right] \\ &= \theta_t - \eta_t \mathbf{I}(\theta_t)^{-1} \nabla_\theta L(\theta_t). \end{aligned}$$

It should be noted that there is an intimate connection between natural gradient descent and approximate second-order optimization method, as the Fisher information can be heuristically viewed as an approximation of the Hessian [Schraudolph, 2002, Martens, 2014]. The above heuristics sheds light on why in practice, natural gradient descent converges with fewer iterations compared to the classic gradient descent. However, in problems where the dimension is high, the per iteration computation for the inverse of the Fisher information can be burdensome.

Stochastic Gradient Langevin Dynamics (SGLD) has been an active research field in sampling and optimization in recent years [Welling and Teh, 2011, Dalalyan, 2017b, Bubeck et al., 2015, Raginsky et al., 2017, Mandt et al., 2017]. SGLD injects an additional $\sqrt{2\beta^{-1}\eta}$ level isotropic Gaussian noise to each step of SGD with step-size η, where β is the inverse temperature parameter. Besides similar optimization benefits as SGD such as convergence and chances of escaping stationary points, the injected randomness of SGLD provides an efficient way of sampling from the targeted invariant distribution of the continuous time diffusion process, which has been shown to be useful statistically in Bayesian sampling [Welling and Teh, 2011, Mandt et al., 2017]. In the current paper,

we take a distinct approach, we motivate and analyze a variant of SGD through lens of Langevin dynamics, from a frequentist point of view, and then present the optimization benefits as a by-product of the statistical motivation.

The approximation in Eqn. (1.6) relates the density evolution of θ_t to a discretized version of Itô diffusion process (as $\eta \to 0$)

$$d\theta_t = -\mathbf{b}(\theta_t)dt + \sqrt{2\beta^{-1}}\mathbf{V}(\theta_t)dB_t.$$

The invariant distribution $\pi(\theta)$ satisfies the following Fokker–Planck equation

$$\beta^{-1}\sum_{i,j}\frac{\partial^2}{\partial x_i x_j}(\pi \mathbf{a}_{ij}) + \sum_i \frac{\partial}{\partial x_i}(\pi \mathbf{b}_i) = 0$$

where $\mathbf{a}_{ij}(x) = (\mathbf{V}(x)\mathbf{V}(x)')_{ij}$. In general, the stationary distribution is hard to characterize unless both $\mathbf{V}$ and $\mathbf{b}$ take special simple forms. For example, when $\mathbf{b}(x)$ is linear and $\mathbf{V}(x)$ is independent of x as in [Mandt et al., 2017], the diffusion process reduces to Ornstein-Uhlenbeck process with multivariate Gaussian as the invariant distribution. Another simple case is when $\mathbf{V}(x) = \mathbf{I}$, the diffusion process is also referred to as Langevin dynamics, with the Gibbs measure $\pi(\theta) \propto \exp(-\beta L(\theta))$ as the unique invariant distribution [Welling and Teh, 2011, Dalalyan, 2017b, Raginsky et al., 2017].

3. STATISTICAL INFERENCE VIA LANGEVIN

Recall the *Moment Adjusted Stochastic Gradient descent* (MasGrad) we introduced, which adjusts the descent direction using the root of the covariance matrix at the current location,

$$\text{MasGrad}: \quad \theta_{t+1} = \theta_t - \eta \mathbf{V}(\theta_t)^{-1}\hat{\mathbb{E}}_n \nabla_\theta \ell(\theta_t, \mathbf{z}). \tag{3.1}$$

Here we present the simplest version of the algorithm, assuming that $\mathbf{V}(\theta)$ can be evaluated at any given θ. We will explain why MasGrad (3.1) produces recursive updates whose statistical distribution can be characterized in the current section. We would like to mention that MasGrad at the same time achieves significant acceleration (compared to SGD) in optimization, when $L(\theta)$ is strongly convex, which we defer the discussions to Section 4. For the general non-convex case, we provide non-asymptotic theory for inference and optimization in Section 5.

3.1 Inference via discretized diffusion approximation

As we have heuristically outlined in Eqn. (1.6), the MasGrad can be approximated by the following discretized Langevin diffusion,

$$\text{Discretized diffusion}: \quad \xi_{t+1} = \xi_t - \eta \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_t. \tag{3.2}$$

In this section, we establish non-asymptotic bounds on the distance between the distribution of MasGrad process $\mathcal{L}(\theta_t, t \in [T])$ and discretized diffusion process $\mathcal{L}(\xi_t, t \in [T])$.

The proof is based on the entropic central limit theorem (CLT) [Barron, 1986, Bobkov et al., 2013, 2014]. The classic CLT based on convergence in distribution is too weak for our purpose: we need to translate the non-asymptotic bounds at each step to the whole stochastic process. It turns out that the entropic CLT couples naturally with the chain-rule property of relative entropy, which together provides non-asymptotic characterization on closeness of the distributions for the stochastic processes.

Let's state the mild assumptions before introducing the theorem. Define $\forall i$,

$$X_i(\theta) = \mathbf{V}(\theta)^{-1}\left[\mathop{\mathbb{E}}_{\mathbf{z}\sim P} \nabla_\theta \ell(\theta, \mathbf{z}) - \nabla_\theta \ell(\theta, z_i)\right].$$

It is clear that $\mathbb{E}\, X_i(\theta) = 0$ and $\mathrm{Cov}[X_i(\theta)] = \mathbf{I}_p$, and X_i's are i.i.d. vector-valued random variables.

ASSUMPTION 3.1 (Entropic distance). *Assume that X has bounded entropic distance to the Gaussian distribution, in the following sense*

$$\sup_\theta D_{\mathrm{KL}}\left(\mu(X(\theta))||\mu(\mathbf{g})\right) \le D, \quad \textit{where } \mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_p). \tag{3.3}$$

ASSUMPTION 3.2 (Finite $(4+\delta)$-th moments). *Assume that $\mathbb{E}\|X(\theta)\|^{4+\delta} < \infty$, for all θ.*

THEOREM 3.1 (Non-asymptotic bound for inference). *Let $\mu(\theta_t, t \in [T])$ denote $\mathcal{L}(\theta_t, t \in [T])$, the joint distribution of MasGrad process, and $\mu(\xi_t, t \in [T])$ be the joint distribution of the discretized diffusion process in (3.2). Assume $\theta_0 = \xi_0$. Under the Assumptions 3.1 and 3.2, the following bound holds,*

$$D_{\mathrm{TV}}\left(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])\right) \le C\sqrt{\frac{T}{n} + o\left(\frac{T(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)}, \tag{3.4}$$

where C is some constant that depends on the $(4+\delta)$-moments in the Assumption 3.2.

REMARK 3.1. The above theorem characterize the sampling distribution of MasGrad – θ_t, using a measure that only depends on the first and second moments of $\nabla\ell(\theta, \mathbf{z})$, namely $\mathbf{b}$ and $\mathbf{V}$, regardless of the specific the data-generating distribution $\mathbf{z} \sim P$. Observe that the distribution closeness is in a strong total variation distance sense, for the two stochastic processes $\{\theta_t, t \in [T]\}$ and $\{\xi_t, t \in [T]\}$. If we dig in to the proof, one can easily obtain the following marginal result

$$D_{\mathrm{TV}}\left(\mu(\theta_T), \mu(\xi_T)\right) \le \sqrt{2D_{\mathrm{KL}}\left(\mu(\theta_T)||\mu(\xi_T)\right)} \le \sqrt{2D_{\mathrm{KL}}\left(\mu(\theta_t, t \in [T])||\mu(\xi_t, t \in [T])\right)},$$

where the last inequality follows from the chain-rule of relative entropy. Therefore, one can as well prove

$$D_{\mathrm{TV}}\left(\mu(\theta_T), \mu(\xi_T)\right) \le C\sqrt{\frac{T}{n}}.$$

REMARK 3.2. One important fact about the Thm. 3.1 is that it holds for any step-size η, which provides us additional freedom of choosing the optimal step-size for the optimization purpose. This theorem is stated in the fix dimensional setting when p does not changes with n. Remark in addition that the Gaussian approximation at each step still holds with high probability, in the moderate dimensional setting when $p = o(\frac{\log n}{\log\log n})$, as shown in the non-asymptotic bound in the above Thm. 3.1.

For the purpose of statistical inference, one can always approximately characterize the distribution of MasGrad using Thm. 3.1. As an additional benefit, the result naturally provides us an algorithmic way of sampling this target universal distribution $\mu(\xi_t)$. For some particular tasks, it remains of theoretical interest to analytically characterize the distribution of MasGrad using the continuous time Langevin diffusion and its invariant distribution. In the next section, we will analyze the discrepancy between the discretized diffusion to the continuous time analog.

3.2 Continuous time Langevin diffusion

In this section we will provide non-asymptotic bounds on the closeness of the discretized and the continuous time Langevin diffusion, in terms of both the Wasserstein-2 distance, and the entropic distance. Let us introduce few notations within this section. Denote $\mathbf{h}(x) = \mathbf{V}(x)^{-1}\mathbf{b}(x)$, and let's define two processes θ_t and ξ_t with the same initial position ξ_0 as follows

$$\text{Continuous}: \quad \theta_t = \theta_0 - \int_0^t \mathbf{h}(\theta_s)ds + \sqrt{2\beta^{-1}}\int_0^t dB_s, \tag{3.5}$$

$$\text{Interpolation}: \quad \xi_t = \theta_0 - \int_0^t \mathbf{h}(\xi_{\lfloor s/\eta \rfloor \eta})ds + \sqrt{2\beta^{-1}}\int_0^t dB_s. \tag{3.6}$$

Here θ_t is the continuous time Langevin diffusion, and ξ_t is an interpolation of the discretization process ξ_k in Eqn. (3.2): for any integer k, the marginal distribution of $\xi_{k\eta}$ is the same as ξ_k, and is well-defined for any $t \in [(k-1)\eta, k\eta]$. Under the following standard assumptions, we can establish Lemma 3.1 for Wasserstein distance and Lemma 3.2 for relative entropy.

ASSUMPTION 3.3 (Lipschitz). *Assume that $\mathbf{h}(\cdot)$ is ℓ-Lipschitz,*

$$\|\mathbf{h}(x) - \mathbf{h}(y)\| \le \ell\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

ASSUMPTION 3.4 (Boundedness). *Assume that $\mathbf{h}(\cdot)$ is M-bounded,*

$$\|\mathbf{h}(x)\| \le M, \quad \forall x \in \mathbb{R}^d.$$

ASSUMPTION 3.5 (Expansiveness). *Assume that $x \mapsto x - \eta\mathbf{h}(x)$ is δ-expansive,*

$$\| (x - \eta\mathbf{h}(x)) - (y - \eta\mathbf{h}(y)) \| \le \delta\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

LEMMA 3.1 (Wasserstein). *Let $W_2(\mu, \nu)$ denote the Wasserstein-2 distance,*

$$W_2(\mu, \nu) \triangleq \left\{ \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^2 d\gamma(x, y) \right\}^{1/2}, \quad \Gamma(\mu, \nu) \text{ are all couplings of } \mu, \nu.$$

Under Assumptions 3.3, 3.4 and 3.5, the Wasserstein-2 distance between the $\xi_{k\eta}$ in (3.6) and $\theta_{k\eta}$ in (3.5) satisfies

$$W_2\left(\mu(\xi_{k\eta}), \mu(\theta_{k\eta})\right) \le \left(\frac{2\ell^2 M^2}{3}\eta^4 + 2\ell^2 p \cdot \beta^{-1}\eta^3 \right)^{1/2} \cdot \sum_{i=0}^{k-1} \delta^i.$$

REMARK 3.3. Let's make few remarks to dissect the non-asymptotic upper bound in Lemma 3.1. Here we borrow Lemma 3.7 in Hardt et al. [2015], which accounts for the expansiveness of updates induced by the vector field $\mathbf{h}$, as follows:

1. If $\mathbf{h}$ is ℓ-smooth, then $x - \eta\mathbf{h}(x)$ is $(1 + \eta\ell)$-expansive;
2. If in addition $\mathbf{h} = \nabla U$, where U is a convex, then for $\eta \le \frac{2}{\ell}$, $x - \eta\mathbf{h}(x)$ is 1-expansive;
3. If in addition U is α-strongly convex, then for $\eta \le \frac{2}{\alpha+\ell}$, $x - \eta\mathbf{h}(x)$ is $(1 - \frac{\eta\alpha\ell}{\alpha+\ell})$-expansive.

First, let's focus on the dependence of η and k in the Wasserstein bound. Plug in $\beta = \frac{2n}{\eta}$, we discuss the three cases for the expansiveness parameter δ.

1. Smooth non-convex: $\delta = 1 + \eta\ell$, we have $W_2\left(\mu(\xi_{k\eta}), \mu(\theta_{k\eta})\right) \le \mathcal{O}(\eta e^{\ell k\eta})$.
2. Convex: $\delta = 1$, the Wasserstein-2 distance reads $W_2\left(\mu(\xi_{k\eta}), \mu(\theta_{k\eta})\right) \le O(k\eta^2)$.
3. Strongly convex: $\delta = 1 - \frac{\eta\alpha\ell}{\alpha+\ell}$, we have $W_2\left(\mu(\xi_{k\eta}), \mu(\theta_{k\eta})\right) \le O(\eta\frac{\alpha+\ell}{\alpha\ell})$.

In the convex and strongly convex cases, the Wasserstein bound depends on k in a desirable weak manner, as one utilizes the non-expansiveness of the vector fields $\mathbf{h}$. In the most general smooth non-convex case, the Wasserstein bound $\mathcal{O}(\eta e^{\ell k\eta})$ agrees with the Grönwall's inequality [Borkar and Mitter, 1999] on the exponential dependence ($e^{k\eta}$ for effective time scaling $k\eta$). This undesirable exponential dependence motivates us to also present the non-asymptotic bound using a different notion — the relative entropy via Girsanov formula in Lemma 3.2.

LEMMA 3.2 (Relative entropy). *Under Assumptions 3.3 and 3.4, the relative entropy between stochastic processes $\{\xi_t, 0 \le t \le k\eta\}$ in (3.6) and $\{\theta_t, 0 \le t \le k\eta\}$ in (3.5) satisfies,*

$$D_{\mathrm{KL}}\left(\mu(\theta_t, 0 \le t \le k\eta) || \mu(\xi_t, 0 \le t \le k\eta)\right) \le \left(\frac{\ell^2 M^2}{6}\beta\eta^3 + \frac{\ell^2 p}{2}\eta^2\right) \cdot k.$$

REMARK 3.4. Let's explain the pros and cons of the upper bound in Lemma 3.2. On the one hand, the bound on relative entropy reads $O((n+p)k\eta^2)$ when plug in $\beta = \frac{2n}{\eta}$, which results in better dependence on k for the general non-convex case. On the other hand, the bound is not as desirable as in Lemma 3.1 for two reasons. First, notions like total variation or entropic distance can be very strong, which can be easily seen in the extreme case when $n \to \infty$ and $\beta = \frac{2n}{\eta} \to \infty$ — the distribution of the discretized diffusion and the continuous time analog are two δ measures with total variation distance 1, even though we know they are path-wise close. The Wasserstein distance captures the path-wise closeness. Second, bound in Lemma 3.2 fails to provide more detailed characterization when the vector field $\mathbf{h}(x)$ enjoys the non-expansive property as in Lemma 3.1.

4. OPTIMIZATION AND ACCELERATION

In this section, we will demonstrate that the "moment adjusting" idea motivated from standardizing the error from an inference perspective achieves similar effect as acceleration in convex optimization. We will investigate *Generalized Linear Models* (GLMs) as the main example. Later, we will also discuss the case with non-smooth regularization. It should be noted that using first-order information to achieve acceleration was first established in the seminal work by Nesterov [1983, 2013] based on the ingenious notion of estimating sequence.

4.1 Convergence to optima

Let us first state a simple convergence lemma for MasGrad, for general smooth convex function $L(\theta)$ in the noiseless setting ($\beta = \infty$). This lemma will be helpful for presenting the main theorem for inference and optimization in the convex case, as well as in the study of GLMs.

LEMMA 4.1 (Convergence: noiseless). *Let $L(w) : \mathbb{R}^p \to \mathbb{R}$ be a smooth convex function. Recall $\mathbf{b}(w) = \nabla L(w)$, and denote $\mathbf{H}(w)$ as the Hessian matrix of L. $\mathbf{V}(w) \in \mathbb{R}^{p\times p}$ is a positive definite matrix. Assume that*

$$\alpha \triangleq \min_{v,w} \lambda_{\min}\left(\mathbf{V}(w)^{-1/2}\mathbf{H}(v)\mathbf{V}(w)^{-1/2}\right) > 0,$$
$$\gamma \triangleq \max_{v,w} \lambda_{\max}\left(\mathbf{V}(w)^{-1/2}\mathbf{H}(v)\mathbf{V}(w)^{-1/2}\right) > 0.$$

The deterministic updates $w_{t+1} = w_t - \eta \mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)$, *with step-size* $\eta = 1/\gamma$, *satisfies*

$$L(w_{t+1}) - \min_w L(w) \le \left(1 - \frac{\alpha}{\gamma}\right)\left(L(w_t) - \min_w L(w)\right).$$

REMARK 4.1. If we define the condition number of MasGrad as

$$\kappa_{\text{MasGrad}} = \frac{\max_{w,v} \lambda_{\max}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right)}{\min_{w,v} \lambda_{\min}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right)}, \quad \kappa_{\text{GD}} = \frac{\max_v \lambda_{\max}\left(\mathbf{H}(v)\right)}{\min_v \lambda_{\min}\left(\mathbf{H}(v)\right)}, \tag{4.1}$$

compared to the condition number in gradient descent. To obtain a solution such that $L(w_t) - \min_w L(w) \le \epsilon$, one need the number of iterations being

$$t = \kappa_{\text{MasGrad}} \cdot \log \frac{L(w_0) - \min_w L(w)}{\epsilon}.$$

Now we are ready to state the theory for inference and optimization using MasGrad, in the strongly convex case.

THEOREM 4.1 (MasGrad: convex). *Let* $L(w)$ *and* α, γ *be the same as in Lemma 4.1. Consider the MasGrad updates* θ_t *in (3.1) with step-size* $\eta = 1/\gamma$, *and the corresponding discretized diffusion* ξ_t,

$$\xi_{t+1} = \xi_t - \eta \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_t, \quad \text{where } \beta = \frac{2n}{\eta}.$$

Then for any precision $\epsilon > 0$, *one can choose*

$$T = \frac{\gamma}{\alpha}\log\frac{2(L(\theta_0) - \min_\theta L(\theta))}{\epsilon}, \text{ and } n = \frac{4p\max_\theta \|\mathbf{V}(\theta)\|}{\alpha\epsilon}, \tag{4.2}$$

such that

$$(1) \quad D_{\text{TV}}\left(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])\right) \le \mathcal{O}_\epsilon\left(\sqrt{\epsilon\log(1/\epsilon)}\right),$$
$$(2) \quad \mathbb{E}\, L(\theta_t) - \min_\theta L(\theta) \le \epsilon, \ \ \mathbb{E}\, L(\xi_t) - \min_\theta L(\theta) \le \epsilon,$$

with in total $\mathcal{O}_\epsilon(\epsilon^{-1}\log 1/\epsilon)$ *independent data samples.*

REMARK 4.2. Using Lemma 4.1, for all $t > 0$, one can prove

$$\mathbb{E}\, L(\xi_t) - \min_\theta L(\theta) \le \left(1 - \frac{\alpha}{\gamma}\right)^t (L(\theta_0) - \min_\theta L(\theta)) + \max_\theta \|\mathbf{V}(\theta)\| \cdot \frac{\gamma}{\alpha}\beta^{-1}p.$$

If $\beta = \frac{2n}{\eta}$ and T, n are chosen as in (4.2), we know that $\mathbb{E}\, L(\xi_T) - L(\theta_*) \le \epsilon$. Recall the result we establish in Thm. 3.1, the total variation distance between MasGrad and the discretize diffusion in this case is bounded by $\sqrt{T/n} = \mathcal{O}_\epsilon\left(\sqrt{\epsilon\log(1/\epsilon)}\right)$, and the total number of samples used is of the order $nT = \mathcal{O}_{\epsilon,p}(p/\epsilon\log(1/\epsilon))$. This result can be contrasted with the classical asymptotic normality for MLE or ERM: to achieve an ϵ-minimizer,

$$\epsilon \ge L(\hat{\theta}_N) - L(\theta_*) \asymp \|\hat{\theta}_N - \theta_*\|^2 \asymp \frac{p}{N} \Leftrightarrow N = \mathcal{O}_{\epsilon,p}(p/\epsilon),$$

the asymptotic sampling complexity is $\mathcal{O}_{\epsilon,p}(p/\epsilon)$. Similar calculations also hold with the Ruppert–Polyak average on stochastic approximation with a carefully chosen decreasing step-size. As we can see, our result holds non-asymptotically, and it achieves both the optimization and inference goal, with an additional log factor of samples.

4.2 Acceleration: GLMs

Now let's take GLMs as an example to describe the effect of acceleration. We will first use an illustrating toy example to show the intuition in an informal way, and then present the rigorous acceleration result for GLMs.

Toy example, non-rigorous. Consider $y_i = \langle x_i, \theta_* \rangle + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for $i \in [N]$. Let's focus on the fixed design case (where the expectation is only over $\mathbf{y}$), the loss $\ell(\theta; (x, y)) = \frac{1}{2}(\langle x, \theta \rangle - y)^2$. Denote $X \in \mathbb{R}^{N \times p}$, then we have

$$\mathbf{b}(\theta) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i^T\theta - y_i)x_i\right] = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T(\theta - \theta_*) = \frac{1}{N}X^T X(\theta - \theta_*),$$

$$\mathbf{V}(w) = \left[\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T \sigma^2\right]^{1/2} = \sigma[\frac{1}{N}X^T X]^{1/2},$$

and the Hessian is $\mathbf{H}(w) = X^T X/N$. Therefore, in this case, we have

$$\kappa_{\text{MasGrad}} = \sqrt{\kappa_{\text{GD}}}.$$

Apply Lemma 4.1, one achieves the same effect as Nesterov's acceleration in the strongly convex case [Nesterov, 2013]. Remark that the above analysis is to demonstrate the intuition, and is not rigorous — as MasGrad only makes sense with the random design.

Generalized linear models, random design, mis-specified model. Now let's provide a rigorous and unified treatment for the generalized linear models. Consider the generalized linear model where the response $\mathbf{y}$ follows from the exponential family parametrize by θ, ϕ

$$f(y; \theta, \phi) = b(y, \phi)e^{\frac{y\theta - c(\theta)}{d(\phi)}}$$

where $\mu = \mathbb{E}[\mathbf{y}|\mathbf{x} = x] = c'(\theta)$, $c''(\theta) > 0$, and the natural parameter satisfies the linear relationship $\theta = \theta(\mu) = x^T w$. In this case, we choose the loss function according to the negative log-likelihood

$$\ell(w; (x, y)) = -y_i x_i^T w + c(x_i^T w).$$

Remark that in the Bernoulli model (logistic regression), one has

$$c(\theta) = \log(1 + e^\theta), \text{ where } x_i^T w = \theta = \log\frac{\mu}{1 - \mu}.$$

In the Poisson model (Poisson regression),

$$c(\theta) = e^\theta, \text{ where } x_i^T w = \theta = \log\mu.$$

In the Gaussian model (linear regression),

$$c(\theta) = \frac{1}{2}\theta^2, \text{ where } x_i^T w = \theta = \mu.$$

We are interested in inference even when the model can be *mis-specified*. We consider the statistical learning setting where $z_i = (x_i, y_i) \sim P = P_{\mathbf{x}} P_{\mathbf{y}|\mathbf{x}}, i \in [N]$ i.i.d. from some unknown joint distribution P. We are trying to infer the parameters w through fitting the data using a parametric exponential family, however, we allow the flexibility that the exponential family model for

$P(\mathbf{y}|\mathbf{x} = x)$ can be mis-specified. Specifically, the true regression function $m_*(x) = \mathbb{E}(\mathbf{y}|\mathbf{x} = x)$ may not be $c'(x^T w)$ for all w, namely, may not be realized by any model in the exponential family model class. We have the population landscape $L(w)$

$$L(w) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim P} \left[-\mathbf{y}\mathbf{x}^T w + c(\mathbf{x}^T w)\right]. \tag{4.3}$$

Define the conditional variance $\xi(x) = \mathrm{Var}(\mathbf{y}|\mathbf{x} = x) \in \mathbb{R}$ and the bias $\beta(\mathbf{x}, w) \triangleq c'(\mathbf{x}^T w) - m_*(\mathbf{x}) \in \mathbb{R}$, one can calculate,

$$\mathbf{b}(w) = \mathbb{E}\left[-\mathbf{y}\mathbf{x} + c'(\mathbf{x}^T w)\mathbf{x}\right] = \mathbb{E}\left[(c'(\mathbf{x}^T w) - m_*(\mathbf{x}))\mathbf{x}\right],$$
$$\mathbf{V}(w) = \left(\mathbb{E}[\xi(\mathbf{x})^2 \mathbf{x}\mathbf{x}^T] + \mathrm{Cov}[\beta(\mathbf{x}, w)\mathbf{x}]\right)^{1/2}, \quad \mathbf{H}(w) = \mathbb{E}\left[c''(\mathbf{x}^T w)\mathbf{x}\mathbf{x}^T\right].$$

We have the following acceleration result for GLMs.

THEOREM 4.2 (Acceleration). *Consider the condition number defined in (4.1) for MasGrad and GD, assume that there exists constant $C > 1$ such that for any x, w, v,*

$$0 < \max\left\{\frac{\xi(x)^2 + \beta(x, w)^2}{c''(x^T v)}, \frac{c''(x^T v)}{\xi(x)^2}\right\} < C^{1/3}.$$

Then for the optimization problem associated with GLMs defined in (4.3), the following holds

$$\kappa_{\mathrm{MasGrad}} < C\sqrt{\kappa_{\mathrm{GD}}}.$$

REMARK 4.3. The above theorem together with Lemma 4.1 states that in the noiseless setting, the time complexity for MasGrad accelerates to $\mathcal{O}\left(\sqrt{\kappa_{\mathrm{GD}}}\log 1/\epsilon\right)$ in contrast to the complexity of GD – $\mathcal{O}\left(\kappa_{\mathrm{GD}}\log 1/\epsilon\right)$, which is crucial when the condition number is large.

4.3 Non-smooth regularization

In this section, we extend the acceleration result to inference problems with non-smooth regularization. We will investigate regression models with non-smooth regularizer $h(\cdot)$ motivated in modern applications, including ℓ_1-regularized sparse regression, and matrix trace regression with nuclear norm regularization. The main results are based on a *Moment Adjusted Proximal Gradient descent* (MaProx). Again, we will present the convergence result in a general form.

In general, we consider when the population loss function can be decomposed into

$$L(w) = g(w) + h(w) \tag{4.4}$$

where $g(w)$ is a smooth and convex function in w, and $h(w)$ is a non-smooth regularizer that is convex. In the case of sparse regression,

$$\ell(w; (x_i, y_i)) = \frac{1}{2}(x_i^T w - y_i)^2 + \lambda\|w\|_1$$
$$L(w) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim P}\left[\frac{1}{2}(\mathbf{x}^T w - \mathbf{y})^2\right] + \lambda\|w\|_1 := g(w) + h(w),$$

where $g(\cdot)$ is smooth convex and $h(w) = \lambda\|w\|_1$ is convex but non-smooth. In the case of low rank matrix trace regression, $X_i, W \in \mathbb{R}^{p\times q}$

$$\ell(W;(X_i,y_i)) = \frac{1}{2}(\langle X_i, W\rangle - y_i)^2 + \lambda\|W\|_*$$
$$L(W) = \mathop{\mathbb{E}}_{(\mathbf{X},\mathbf{y})\sim P}\left[\frac{1}{2}(\langle \mathbf{X}, W\rangle - \mathbf{y})^2\right] + \lambda\|W\|_* := g(W) + h(W),$$

where $g(\cdot)$ is smooth convex and $h(W) = \lambda\|W\|_*$ is convex but non-smooth.

Now we will show the role of moment matrix $\mathbf{V}$ in "speeding up" the convergence of proximal gradient descent. Here we focus on an easier case when $\mathbf{V}(w)$ does not depend on w[3]. Consider the moment adjusted proximal function

$$\text{prox}_{\eta,\mathbf{V}}(w) = \arg\min_u \left[\frac{1}{2\eta}\|u - w\|_{\mathbf{V}}^2 + h(u)\right]. \tag{4.5}$$

One can see that

$$\text{MaProx:}\quad w_{t+1} = \text{prox}_{\eta,\mathbf{V}}(w_t - \eta\mathbf{V}^{-1}\nabla g(w_t)) \tag{4.6}$$

implements moment adjusting gradient (using implicit updates) because w_{t+1} satisfies the implicit equation

$$w_{t+1} = w_t - \eta\mathbf{V}^{-1}(\nabla g(w_t) + \partial h(w_{t+1})),$$

in comparison to the sub-gradient step (explicit updates)

$$w_{t+1} = w_t - \eta\mathbf{V}^{-1}(\nabla g(w_t) + \partial h(w_t)).$$

Remark the classic proximal gradient is when $\mathbf{V}$ being the identity matrix.

THEOREM 4.3 (Moment Adjusted Proximal). *Consider $L(w) = g(w)+h(w)$ where g is a smooth convex function and h is a non-smooth regularizer. Denote $\mathbf{H}$ as the Hessian of g, and define*

$$\alpha \triangleq \min_v\ \lambda_{\min}\left(\mathbf{V}^{-1/2}\mathbf{H}(v)\mathbf{V}^{-1/2}\right) > 0, \quad \gamma \triangleq \max_v\ \lambda_{\max}\left(\mathbf{V}^{-1/2}\mathbf{H}(v)\mathbf{V}^{-1/2}\right) > 0.$$

Consider MaProx updates with step-size $\eta = 1/\gamma$ and adjusting matrix $\mathbf{V}$,

$$w_{t+1} = \text{prox}_{\eta,\mathbf{V}}(w_t - \eta\mathbf{V}^{-1}\nabla g(w_t)),$$

then if

$$T \geq \frac{\gamma}{\alpha}\log\left(\frac{\alpha}{2\epsilon}\|w_0 - w_*\|_{\mathbf{V}}^2 + 1\right),$$

we have

$$L(w_T) - \min_w L(w) \leq \epsilon.$$

[3]As is in the linear regression fixed design case, where $\mathbf{V}(w) = \left(\mathbb{E}[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T]\right)^{1/2}$ does not depend on w.

REMARK 4.4. Remark that as in the GLM case, the moment adjusted idea speed up the computation as the number of proximal steps scales with adjusted condition number

$$\kappa_{\text{MaProx}} = \frac{\max_v \lambda_{\max}\left(\mathbf{V}^{-1/2}\mathbf{H}(v)\mathbf{V}^{-1/2}\right)}{\min_w \lambda_{\min}\left(\mathbf{V}^{-1/2}\mathbf{H}(v)\mathbf{V}^{-1/2}\right)} \approx \sqrt{\kappa_{rmGD}}.$$

However, to be fair, it can harder to implement each proximal step for a non-diagonal $\mathbf{V}$. Motivated from the diagonalizing idea in AdaGrad [Duchi et al., 2011], one can substitute $\mathbf{V}$ by $\text{diag}(\mathbf{V})$ to save the per-round computation.

In the limit case when $\alpha = 0$, i.e., the function g is convex but not strongly convex, our complexity scales with

$$\lim_{\alpha \downarrow 0} \frac{\gamma}{\alpha} \log\left(\frac{\alpha}{2\epsilon}\|w_0 - w_*\|_{\mathbf{V}}^2 + 1\right) = \frac{\gamma\|w_0 - w_*\|_{\mathbf{V}}^2}{2\epsilon},$$

which means the convergence scales with $\mathcal{O}(1/T)$. Remark that one can use explicit sub-gradient updates. However, the convergence is considerably slower, which scales with $\mathcal{O}^*(1/\sqrt{T})$.

We would like to conclude this section by discussing the connections between pre-conditioning methods and our moment adjusting method. Pre-conditioning considers performing a linear transformation $\xi = A^{-1}\theta$ on the original parameter space on θ. In other words, consider $\tilde{L}(\xi) \triangleq L(A\xi)$, and perform the updates on ξ yields

$$\xi_{t+1} = \xi_t - \eta\nabla_\xi \tilde{L}(\xi) = \xi_t - \eta A\mathbf{b}(A\xi_t) \;\Rightarrow\; \theta_{t+1} = \theta_t - \eta A^2\mathbf{b}(\theta_t),$$

Therefore, in the noiseless case, moment adjusting method is equivalent to pre-conditioning when the moment matrix $\mathbf{V}(\theta)$ is a constant matrix w.r.t. θ. However, in Langevin diffusion when the isotropic Gaussian noise is presented, the connection becomes more subtle — as $\mathbf{V}^{-1}(\theta)\mathbf{b}(\theta)$ may not be the gradient vector field for any function. The moment adjusting idea motivated from standardizing noise in statistics is different from the pre-conditioning idea in optimization. We would also like to point out that a nice idea using Hessian information to speed up the Langevin diffusion for sampling from log-concave distribution has been considered in Dalalyan [2017b].

Remark that we use the moment matrix at the current point θ_t (time varying) instead of the optimal point θ_* (which is unknown). We also use the matrix root instead of the covariance matrix itself. In the case when the model is well-specified, and the loss function chosen to be the negative log-likelihood, the $V(\theta_*)$ is the root of the Fisher information matrix.

5. NON-CONVEX INFERENCE

In this section, we will study the non-asymptotic inference and optimization for stationary points for a smooth non-convex population landscape $L(\theta)$, via our proposed MasGrad.

5.1 Convergence to stationary points

First we will state a theorem that quantifies how well our proposed MasGrad achieves both the inference and optimization goal.

THEOREM 5.1 (MasGrad: non-convex). *Let $L(w) : \mathbb{R}^p \to \mathbb{R}$ be a smooth function. Recall $\mathbf{b}(w) = \nabla L(w)$, and $\mathbf{H}(w)$ being the Hessian matrix of L. $\mathbf{V}(w) \in \mathbb{R}^{p\times p}$ is a positive definite matrix. Assume*

$$\gamma \triangleq \max_{v,w}\; \lambda_{\max}\left(\mathbf{V}(w)^{-1/2}\mathbf{H}(v)\mathbf{V}(w)^{-1/2}\right) > 0.$$

Consider the MasGrad updates θ_t in (3.1) with step-size $\eta = 1/\gamma$, and the corresponding discretized diffusion ξ_t,

$$\xi_{t+1} = \xi_t - \eta \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_t, \quad \textit{where } \beta = \frac{2n}{\eta}.$$

Then for any precision $\epsilon, \delta > 0$, one can choose

$$T = \frac{2\gamma(L(\theta_0) - \min_\theta L(\theta)) + p\delta^2}{\epsilon^2} \cdot (\max_\theta \|\mathbf{V}(\theta)\| \vee 1), \ \textit{and} \ \ n = \frac{T}{\delta^2}, \tag{5.1}$$

such that

$$\begin{aligned} &(1) \quad D_{\mathrm{TV}}\left(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])\right) \le \mathcal{O}_\delta(\delta), \\ &(2) \quad \mathbb{E}\min_{t \le T} \|\nabla L(\theta_t)\| \le \epsilon, \ \ \mathbb{E}\min_{t \le T} \|\nabla L(\xi_t)\| \le \epsilon, \end{aligned}$$

with in total $\mathcal{O}_{\epsilon,\delta}(\epsilon^{-4}\delta^{-2})$ independent data samples.

Remark 5.1. We would like to contrast the optimization part of the above theorem to the complexity result of classic SGD. To obtain an ϵ-stationary point w such that in expectation $\|\nabla L(w)\| \le \epsilon$, SGD needs $\mathcal{O}_\epsilon(\epsilon^{-4})$ for non-convex smooth functions (with step size $\eta_t = \min\{1/\gamma, 1/\sqrt{t}\}$). Here we show that one can achieve this accuracy with the same dependence on ϵ with MasGrad, while being able to make statistical inference at the same time. And the additional price we pay for δ-closeness in distribution for statistical inference is a factor of δ^{-2}.

The result can also be compared to Thm. 4.1 (the strongly convex case). In both cases, statistically, we have shown that the discretized diffusion ξ_t tracks the non-asymptotic distribution of MasGrad θ_t, as long as the data generating process satisfies weak moment condition and bounded entropic distance to Gaussian. The distribution of ξ_t is universal regardless of the specific data generating distribution. In terms of optimization, to obtain an ϵ-minimizer, the discretized diffusion approximation to MasGrad — with the proper step-size η, and inverse temperature $\beta = 2n/\eta$ — achieves the acceleration in the strongly convex case, and enjoys the same dependence on ϵ as SGD in the non-convex case.

5.2 Why local inference

For a general non-convex landscape, let us discuss why we focus on inference about local optima, or more precisely stationary points. Our Thm. 5.1 can be read as, within reasonable number of steps, the MasGrad converges to a population stationary point, and the distribution is well-described by the discretized Langevin diffusion. One can argue that the random perturbation introduced by the isotropic Gaussian noise in Langevin diffusion makes the process hard to converge to a typical saddle point. Therefore, intuitively, the MasGrad will converge to a distribution that is well concentrated near a certain local optima (depends on the initialization) as the temperature parameter $\beta^{-1} = \eta/2n$ is small. In this asymptotic low temperature regime, the Eyring-Kramer Law states that the transiting time from one local optimum to another local optimum, or the exiting time from a certain local optimum, is very long — roughly $e^{\beta h}$ where h is the depth of the basin of the local optimum. Therefore, a reasonable and tangible goal is to establish statistical inference for population local optima, for a particular initialization.

6. NUMERICAL EXPERIMENTS

6.1 Linear model

The first numerical example is the simple linear regression, as in Fig. 1. Here we generate two plots as a proof of concept. The top one summarizes the trajectory of several methods for inference — our proposed *MasGrad*, the discretized diffusion approximation *diff_MasGrad*, as well as classic *SGD*, and the diffusion approximation *diff_SGD* — with the confidence intervals (95% coverage) at each time step t. In this convex setting, we can solve for the global optimum, which is labeled as the *truth*. Here the mini-batch size is $n = 50$. We run 100 independent chains to calculate the confidence intervals at each step. We look at the low dimensional case $p = 4$, and the four subfigures (on top) each corresponds to one coordinate of the parameter $w_i, i \in [p]$. The x-axis is t, the time of the evolution, and y-axis is the value of the parameter w. Remark that, *MasGrad* and *diff_MasGrad*, are path-wise close in terms of distribution, which verifies our statistical theory in Thm. 3.1. Similar fact holds for *GD* and *diff_GD*. Remark that in this simulation, the condition number of the empirical Gram matrix is 30.98, and the first and third coordinates have very small population eigenvalues, which explains why in those coordinates *MasGrad* has significant acceleration compared to *SGD* as shown in the figure. To be fair, at each time step, both MasGrad and SGD sample same amount of data, and the step-size is chosen as in Thm. 4.2. All four chains start with the same random initialization.

To examine the optimization side of the story, we plot the logarithm of the ℓ_2-error according to time t, for *diff_MasGrad* and *diff_SGD*, in the bottom plot. Remark that the error bar quantifies the confidence interval for the log error. In theory, we should expect that the slope of MasGrad is twice as the slope of SGD. In simulation, it seems that the acceleration is slightly better than what the theory predicts.

We would like to remark that compared to GD, which different coordinates make uneven process (fast process in the second and fourth coordinates, but slow on the others), MasGrad adaptively adjust the relative step-size on each coordinate to achieve synced progress. This effect has also been observed in AdaGrad and natural gradient descent.

Let us provide the full details of the experiment. In the experiment, we generate a larger number of samples as the population (so that we can evaluate $\mathbf{V}$ easily), then use bootstrap to sample from this population at each step. The population minimizer can be solved using least squares. Here each row of the "population" data matrix $X \in \mathbb{R}^{500\times 4}$ is sampled from a multivariate Gaussian independently, with a covariance matrix Σ that has condition number 30.98. Each step we independently subsample $n = 50$ rows with replacement. The response is generated from a well-specified linear model with additive standard Gaussian noise. The step-size is through calculating the smoothness parameter γ as in Thm. 4.2.

6.2 Logistic model

Fig. 2 illustrates the acceleration for inference in logistics regression. The figure should be read the same way as in the linear case. In this case, we sample a much larger number of samples ($N = 500$) and the use GLM package in R to fit the global optimum. Then for MasGrad and SGD, we generate bootstrap subsamples ($n = 25$) to make stochastic descents at each iteration. Again, we run 100 independent chains to calculate the confidence interval at each step. In this case, there is no theoretically optimal way of choosing the step-size, so we choose the same step-size ($\eta = 0.2$) for both MasGrad and SGD.

Statistically, the *MasGrad* and *diff_MasGrad* are close in distribution when $t < 100$, and they both reach a stationary distribution after around 50 steps, simultaneously for all $p = 4$ coordinates.

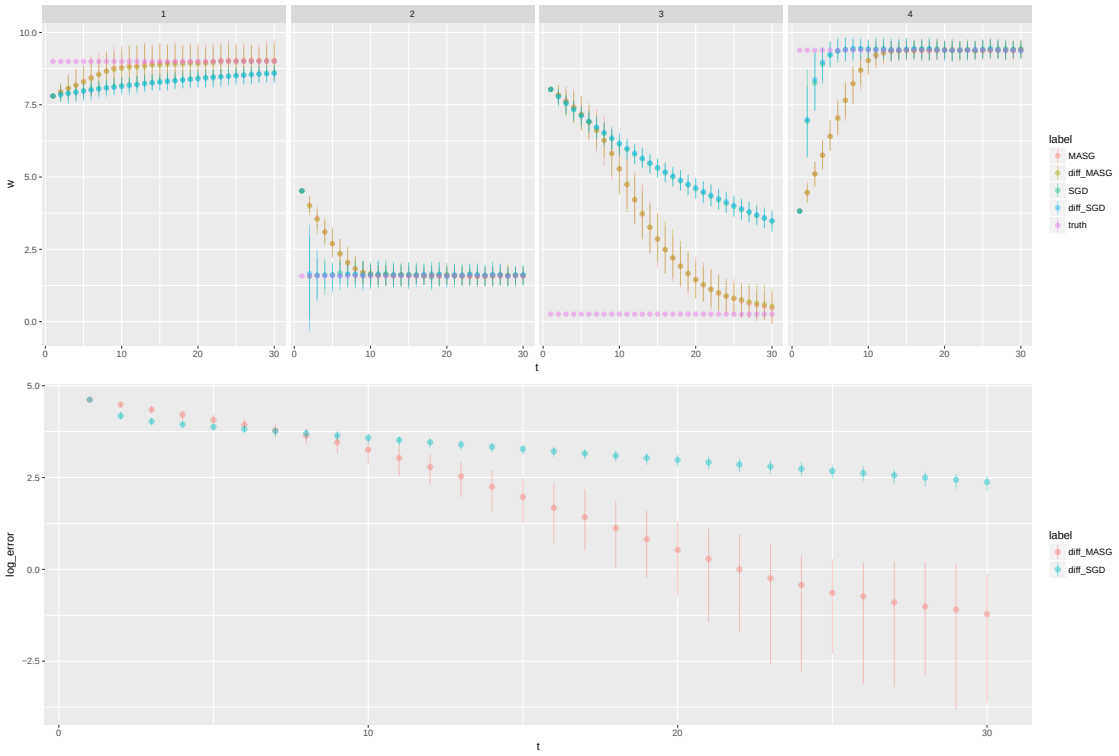

FIG 1. *Linear regression*

Then the distribution fluctuates around stationarity. However, *GD* and *diff_GD* makes much slower progress, and they haven't reach the global optimum in 100 steps.

For optimization, empirically, the acceleration in the log error plot seems to be better than what the theoretical results predict. Remark that the confidence intervals are on the scale of log error, therefore, it is negative-skewed.

Again we will provide the full details of the experiment. We fix a step-size $\eta = 0.2$ (other step-sizes essentially provide similar results), which implies the inverse temperature is $\beta = 2n/\eta = 250$. The data matrix $X \in \mathbb{R}^{500\times 4}$ is generated from multivariate Gaussian with identity covariance. The response is generated from a well-specified logistic model with each coordinate of w_* uniformly sampled between $[1, 2]$.

6.3 Gaussian mixture

In this section we showcase inference via MasGrad for the Gaussian mixture model. We will consider a simple setting: the data $z_i \in \mathbb{R}^n, 1 \le i \le [N]$ generated from a mixture of p Gaussians, with mean $[\theta_1, \theta_2, \ldots, \theta_p] \triangleq \theta$ respectively, and variance σ^2. The goal is to infer the unknown mean vector $\theta \in \mathbb{R}^p$. The problem is non-convex due to the mixture nature: the maximum likelihood is multimodal, as we can shuffle the coordinates of θ to obtain equivalent class of local optima.

$$\ell(\theta; z) = -\log\left(\sum_{i=1}^{p} q_i \phi(z - \theta_i)\right), \quad \text{s.t. } \sum_{i=1}^{p} q_i = 1,$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ denotes the density function for Gaussian. Here in simulations we consider the case when the mixture probability $q_i, i \in [p]$ is known and uniform for the simplicity that we

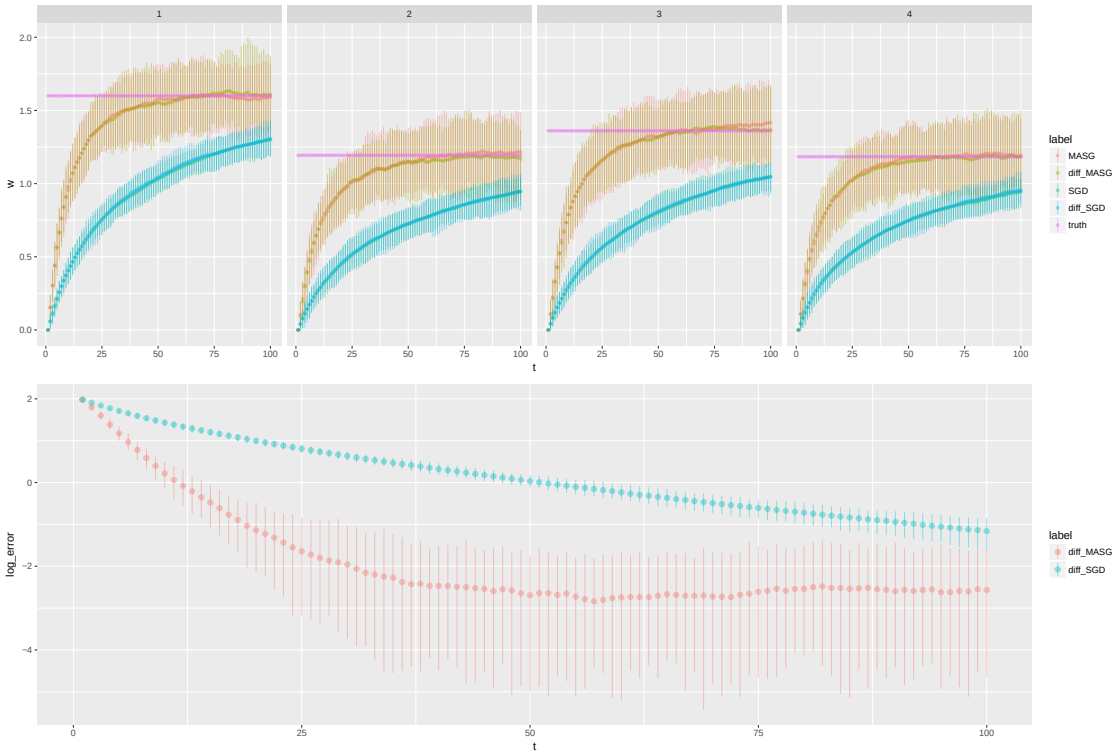

FIG 2. *Logistic regression*

can apply the MasGrad without equality constraints[4], and we have a clear picture of the global optima due to symmetry.

Fig. 3 illustrates the acceleration for inference in the Gaussian mixture model. Here we run two simulations, according to the difficulty or separability of the problem defined as SNR $\triangleq \min_{i\neq j} |\theta_i - \theta_j|/\sigma$. The top one is for the easy case with SNR $= 3.3$ and the bottom one for the hard case with SNR $= 1$. In both simulations, $\theta = (1, 2, 3) \in \mathbb{R}^3$, and we choose a random initial point to start the chains. The plot is presented as before. At each iteration, we subsample $n = 20$ data points to calculate the decent direction, and the step-size is fixed to be $\eta = 0.05$.

Remark that there are many population local optima (at least $3! = 6$), and both *MasGrad* and *diff_MasGrad* seem to be able to find a good local optimum relatively quickly (which concentrates near a permutation of $1, 2, 3$ for each coordinate), compared to *SGD* and *diff_SGD*. The acceleration effect in both cases seems to be apparent. Again, we want to emphasize that the convergence for each coordinate in MasGrad seems to happen around the same number of iterations, which is not true for SGD.

6.4 Shallow neural nets

In this section we run MasGrad on a 2-layer ReLU neural network, as a proof of concept for non-convex models. Define the ReLU activation $\sigma(x) = \max(x, 0)$, a two layer neural network (with k hidden units) represents a function

$$f_w(x) = \sigma(W_2\sigma(W_1 x)), \quad \text{where } x \in \mathbb{R}^d,\ w = \{W_1 \in \mathbb{R}^{k\times d}, W_2 \in \mathbb{R}^{1\times k}\}.$$

[4]When the mixture probability is also unknown, one will need to consider adding a proper barrier function before applying the gradient method.

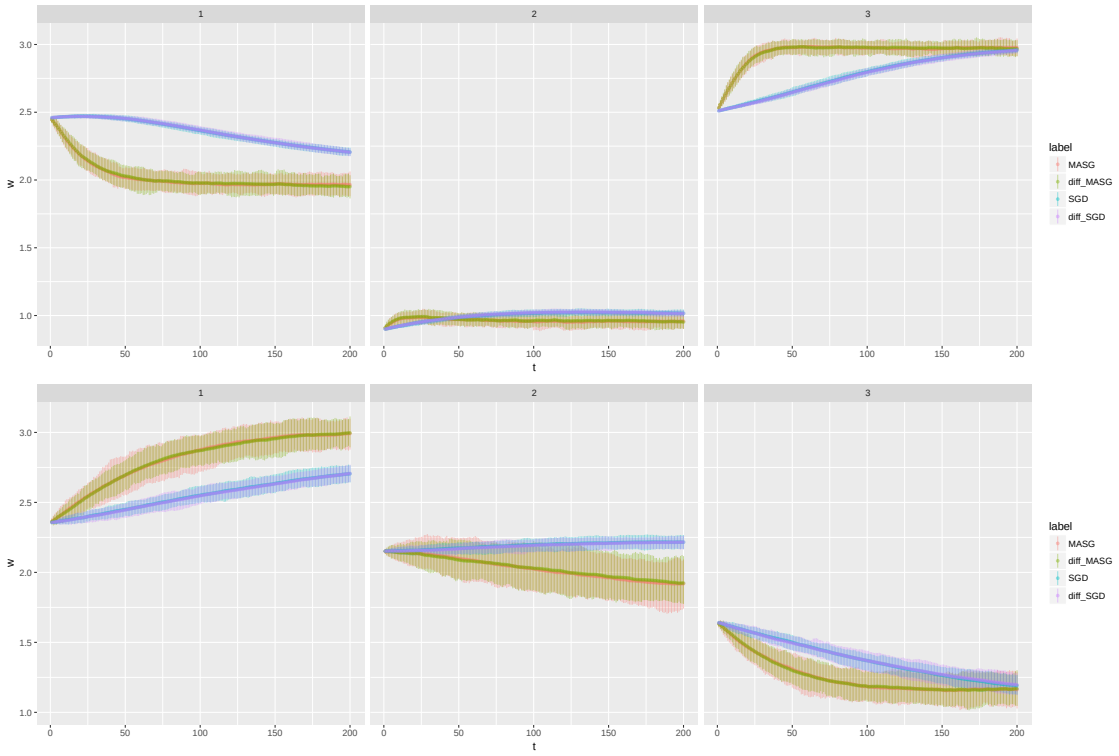

FIG 3. *Gaussian mixture*

In our experiment, we use the square loss

$$\ell(w;(x,y)) = \frac{1}{2}(y - f_w(x))^2.$$

The gradients can be calculated through back-propagation. In the experiment we generate from a well-specified model with very small additive Gaussian noise. However, due to the presense of the hidden layer, the problem is non-convex with many local optima. To break the ReLU scaling invariance (i.e., $\{cW_1, 1/cW_2\}$ is equivalent to $\{W_1, W_2\}$, for the purpose of letting stationary points more separable), we add a non-programmable constant in each layer in the experiment, namely $f_w(x) = \sigma(1 + W_2\sigma(\mathbf{1} + W_1 x))$. In this case, it is harder to calculate the global optimum, instead, we run 50 experiments with random initializations to explore the population landscape, in order to compare the *diff_MasGrad* and *SGD*.

For each experiment (as illustrated in the top figure in Fig. 4), we randomly initialize the weights using standard Gaussians. Because we generate the data from a well specified model, we also present the true parameter in the plot. Here we choose $n = 30$, and each step we subsample with replacement from $N = 300$ data points. The step-size is fixed to be $\eta = 0.1$, which implies the inverse temperature being $\beta = 600$. As usual, we run 100 independent chains with the same initial points for *diff_MasGrad* and *SGD* to calculate the confidence interval. As anticipated, the distribution is rather non-Gaussian (for instance, in coordinate 2 and 6). We run the chain for 100 steps, and then evaluate the population loss function for the two methods. Out of the 50 experiments, $45/50 = 90\%$ time the population loss returned by *diff_MasGrad* is much smaller than that of the *SGD*. The bottom figure in Fig. 4 plots the histogram (dotplot using ggplot2 [Wickham, 2009]) of the population error (test accuracy). Remarkably, the *diff_MasGrad* seems to be able to converge

to "better" local optima most of the time. There could be several explanations: first, MasGrad uses better local geometry (similar to natural gradient) so that it induces better implicit regularization; second, MasGrad as an optimization method accelerates the chain so that it mixes to a local optima faster, compared to SGD which may not yet converge within a certain time budget.

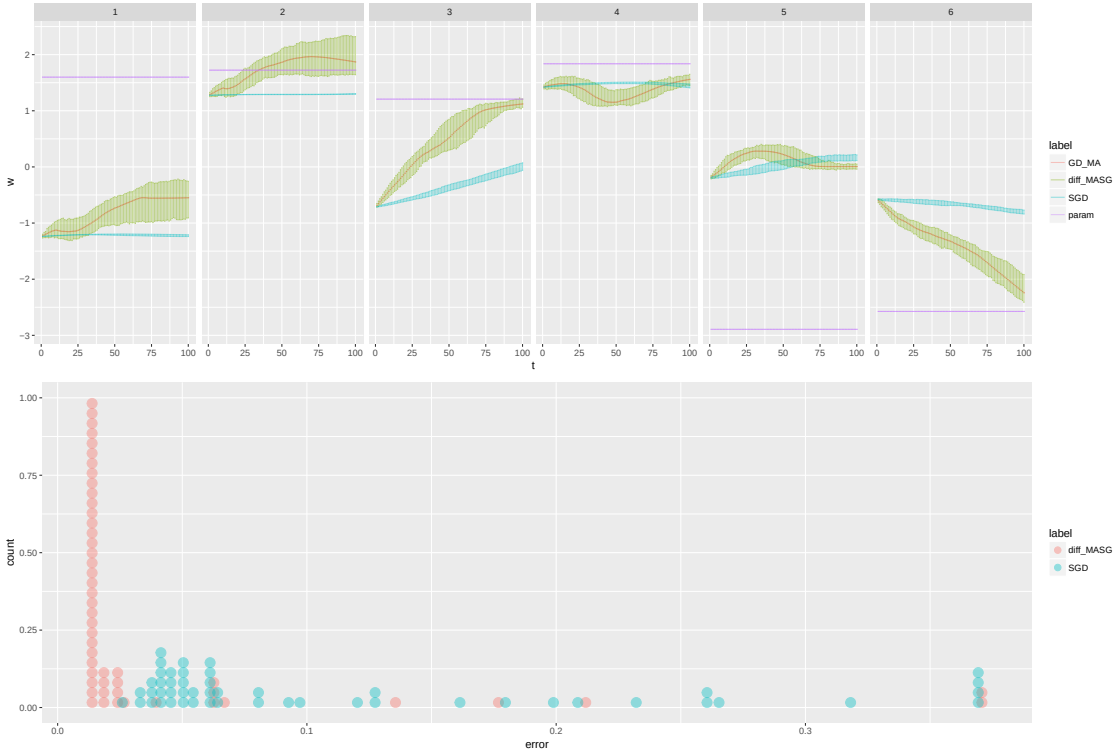

FIG 4. *Shallow neural nets*

7. FURTHER DISCUSSIONS

In this section, we will briefly discuss the issue with the unknown $\mathbf{V}(\theta_t)$. Note that in the fixed dimension setting, one can estimate the covariance matrix of the gradient $\nabla\ell(\theta;\mathbf{z})$ using the empirical version with N independent samples, when N is large. Let us be more careful in this statement: (1) When the population landscape is convex, then the global optimum of $\hat{L}_N(\theta)$ and $L(\theta)$ are within $1/\sqrt{N}$. We can always treat $\hat{L}_N(\theta)$ as the population version and at each step, we bootstrap subsamples of size n to evaluate the stochastic gradients, adjusted using the empirical covariance $\hat{\mathbf{V}}_N$ calculated using N data points. Intuitively, when $\eta < \mathcal{O}(n/N)$ (so that $\beta > N$), we know the MasGrad will concentrate near the optimum of $\hat{L}_N(\theta)$ with better accuracy than $1/\sqrt{N}$. (2) In the non-convex case, things become unclear. However, under stronger conditions such as strongly Morse [Mei et al., 2016], i.e., when there are nice one-to-one correspondence between the stationary points of $\hat{L}_N(\theta)$ and $L(\theta)$, one may still using the bootstrap idea above with $\hat{\mathbf{V}}_N$. (3) Computation of $\hat{\mathbf{V}}_N$ and its inverse could be burdensome, one may want to calculate a diagonalized version of $\hat{\mathbf{V}}_N$ as done in AdaGrad [Duchi et al., 2011]. (4) To have fully rigorous non-asymptotic theory as the case when $\mathbf{V}$ is known, one may require involved tools from self-normalized processes [Peña et al., 2008] to establish a similar version of entropic CLT for multivariate self-normalized processes, where we standardize $\hat{\mathbb{E}}_n[\nabla\ell(\theta,\mathbf{z})]$ by the empirical covariance matrix $\hat{\mathbf{V}}_n$ calculated

based on the same samples. To the best of our knowledge, this is an ambitious and challenging goal that is beyond the scope and focus of the current paper.

8. TECHNICAL PROOFS

PROOF OF THEOREM 3.2. The MasGrad updates can be represented as

$$\theta_{t+1} = \theta_t - \eta \mathbf{V}(\theta_t)^{-1}\mathbf{b}(\theta_t) + \sqrt{2\beta^{-1}\eta}\frac{\sum_{i=1}^n X_i(\theta_t)}{\sqrt{n}}. \tag{8.1}$$

Denote $S_n(X,\theta_t) = \frac{\sum_{i=1}^n X_i(\theta_t)}{\sqrt{n}}$. Under the Assumptions 3.1 and 3.2, Thm. 6.1 in Bobkov et al. [2013] (with $(4+\delta)$-moment condition) implies at each step t,

$$D_{\mathrm{KL}}\left(\mu(S_n(X,\theta_t))||\mu(\mathbf{g}_t)|\theta_t\right) = \frac{C}{n} + o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{\frac{4+\delta-2}{2}}}\right) = \frac{C}{n} + o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right), \tag{8.2}$$

conditioned on θ_t, or some constant $C > 0$.

Apply the chain-rule for relative entropy, we know that

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mu(\theta_t, t\in[T])||\mu(\xi_t, t\in[T])\right)\\
&= D_{\mathrm{KL}}\left(\mu(\theta_t, t\in[T-1])||\mu(\xi_t, t\in[T-1])\right) + \int D_{\mathrm{KL}}\left(\mu(S_n(X,\theta_{T-1})||\mu(\mathbf{g}_{T-1})|\theta_{T-1}\right) d\mu(\theta_t, t\in[T-1])\\
&\le D_{\mathrm{KL}}\left(\mu(\theta_t, t\in[T-1])||\mu(\xi_t, t\in[T-1])\right) + \frac{C}{n} + o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)\\
&\le \ldots \le D_{\mathrm{KL}}\left(\mu(\theta_0)||\mu(\xi_0)\right) + \frac{CT}{n} + o\left(\frac{T(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right),
\end{aligned}$$

where the second step uses the fact for $a, b > 0$, $D_{\mathrm{KL}}(\mu(X)||\mu(Y)) = D_{\mathrm{KL}}(\mu(a+bX)||\mu(a+bY))$, therefore

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mu\left(\theta_{T-1} - \eta\mathbf{V}(\theta_t)^{-1}\mathbf{b}(\theta_{T-1}) + \sqrt{2\beta^{-1}\eta}\frac{\sum_{i=1}^n X_i(\theta_{T-1})}{\sqrt{n}}\right)||\,\mu\left(\theta_{T-1} - \eta\mathbf{V}(\theta_t)^{-1}\mathbf{b}(\theta_{T-1}) + \sqrt{2\beta^{-1}\eta}\mathbf{g}_{T-1}\right)\middle|\,\theta_{T-1}\right)\\
&= D_{\mathrm{KL}}\left(\mu\left(\sum_{i=1}^n X_i(\theta_{T-1})/\sqrt{n}\right)||\mu(\mathbf{g}_{T-1})|\theta_{T-1}\right)\\
&= D_{\mathrm{KL}}\left(\mu(S_n(X,\theta_{T-1})||\mu(\mathbf{g}_{T-1})|\theta_{T-1}\right).
\end{aligned}$$

Apply the Pinsker's inequality that for any random variables X, Y,

$$\frac{1}{2}D_{TV}(\mu(X),\mu(Y))^2 \le D_{\mathrm{KL}}(\mu(X)||\mu(Y)),$$

we finish the proof. □

PROOF OF LEMMA 3.1. The proof is motivated from [Dalalyan, 2017a]. We will show that the proof extends to more general vector fields $\mathbf{h}$ using the notion of expansiveness [Hardt et al., 2015], without requiring $\mathbf{h}$ to be the gradient of a strongly convex function. Another difference is that we are tracking the difference between the Cauchy discretization ξ_t and the Langevin diffusion θ_t,

instead of characterizing the distance of ξ_t to the invariant measure. In addition, we generalize the proof to review the explicit dependence on the inverse temperature β.

Consider θ_t and ξ_t defined using the same Brownian motion B_t, then we have

$$\begin{aligned}
\|\xi_{k\eta} - \theta_{k\eta}\| &\le \|[\xi_{(k-1)\eta} - \eta \mathbf{h}(\xi_{(k-1)\eta})] - [\theta_{(k-1)\eta} - \eta \mathbf{h}(\theta_{(k-1)\eta})]\| \\
&\quad + \left\| \int_{(k-1)\eta}^{k\eta} \left[\mathbf{h}(\theta_t) - \mathbf{h}(\theta_{(k-1)\eta}) \right] dt \right\| \\
\left(\mathbb{E} \|\xi_{k\eta} - \theta_{k\eta}\|^2 \right)^{1/2} &\le \left(\mathbb{E} \|[\xi_{(k-1)\eta} - \eta \mathbf{h}(\xi_{(k-1)\eta})] - [\theta_{(k-1)\eta} - \eta \mathbf{h}(\theta_{(k-1)\eta})]\|^2 \right)^{1/2} \\
&\quad + \underbrace{\left(\mathbb{E} \left\| \int_{(k-1)\eta}^{k\eta} \left[\mathbf{h}(\theta_t) - \mathbf{h}(\theta_{(k-1)\eta}) \right] dt \right\|^2 \right)^{1/2}}_{\text{defined as } \Delta} \\
&\le \delta \, \mathbb{E} \left(\|\xi_{(k-1)\eta} - \theta_{(k-1)\eta}\|^2 \right)^{1/2} + \Delta
\end{aligned}$$

where the first two steps use triangle inequality, on $\mathbb{R}^p$ and ℓ_2 space associated with $\mathbb{E}$ respectively. The last step uses the following fact about the δ-expansiveness,

$$\|[\xi_{(k-1)\eta} - \eta \mathbf{h}(\xi_{(k-1)\eta})] - [\theta_{(k-1)\eta} - \eta \mathbf{h}(\theta_{(k-1)\eta})]\| \le \delta \|\xi_{(k-1)\eta} - \theta_{(k-1)\eta}\|.$$

For the term Δ,

$$\begin{aligned}
\Delta^2 &= \mathbb{E} \sum_{i=1}^{d} \left| \int_{(k-1)\eta}^{k\eta} \left[\mathbf{h}(\theta_t) - \mathbf{h}(\theta_{(k-1)\eta}) \right]_i dt \right|^2 \\
&\le \mathbb{E} \sum_{i=1}^{d} \eta \int_{(k-1)\eta}^{k\eta} | \left[\mathbf{h}(\theta_t) - \mathbf{h}(\theta_{(k-1)\eta}) \right]_i |^2 dt \quad \text{by Cauchy-Schwartz} \\
&= \eta \int_{(k-1)\eta}^{k\eta} \mathbb{E} \|\mathbf{h}(\theta_t) - \mathbf{h}(\theta_{(k-1)\eta})\|^2 dt \\
&\le \eta \ell^2 \int_{(k-1)\eta}^{k\eta} \mathbb{E} \|\theta_t - \theta_{(k-1)\eta}\|^2 dt \quad \text{by } \ell\text{-Lipschitz} \\
&= \eta \ell^2 \int_{(k-1)\eta}^{k\eta} \mathbb{E} \left\| - \int_{(k-1)\eta}^{t} \mathbf{h}(\theta_s) ds + \sqrt{2\beta^{-1}} (B_t - B_{(k-1)\eta}) \right\|^2 dt
\end{aligned}$$

$$\begin{aligned}
&\le \eta \ell^2 \int_{(k-1)\eta}^{k\eta} \left\{ 2 \, \mathbb{E} \left\| - \int_{(k-1)\eta}^{t} \mathbf{h}(\theta_s) ds \right\|^2 + 2 \, \mathbb{E} \left\| \sqrt{2\beta^{-1}} (B_t - B_{(k-1)\eta}) \right\|^2 \right\} dt \\
&\le 2\eta \ell^2 \int_{(k-1)\eta}^{k\eta} (t - (k-1)\eta) \int_{(k-1)\eta}^{t} \mathbb{E} \|h(\theta_s)\|^2 ds dt + 2\eta \ell^2 \int_{(k-1)\eta}^{k\eta} 2\beta^{-1} p (t - (k-1)\eta) dt \\
&\le 2\eta \ell^2 \int_{(k-1)\eta}^{k\eta} (t - (k-1)\eta)^2 M^2 dt + 2\ell^2 p \beta^{-1} \eta^3 \quad \text{by } M\text{-boundedness} \\
&\le \frac{2}{3} \ell^2 M^2 \eta^4 + 2\ell^2 p \beta^{-1} \eta^3.
\end{aligned}$$

Then going back to the original equation we are trying to bound

$$\left(\mathbb{E}\left\|\xi_{k\eta}-\theta_{k\eta}\right\|^2\right)^{1/2} \leq \left(\frac{2}{3}\ell^2 M^2\eta^4 + 2\ell^2 p\beta^{-1}\eta^3\right)^{1/2} \cdot \sum_{i=0}^{k-1}\delta^i. \tag{8.3}$$

□

Proof of Lemma 3.2. The proof follows from calculations as in Dalalyan [2017b], Raginsky et al. [2017]. The continuous-time interpolation enjoys the same distribution as $\xi_{k\eta}$ for all k. One can apply Girsanov formula to calculate the relative entropy

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\mu(\theta_t, 0\leq t\leq k\eta)||\mu(\xi_t, 0\leq t\leq k\eta)\right)\\
&= \frac{\beta}{4}\int_0^{k\eta}\mathbb{E}\|\mathbf{h}(\xi_t)-\mathbf{h}(\xi_{\lfloor t/\eta\rfloor\eta})\|^2 dt\\
&= \frac{\beta}{4}\sum_{i=0}^{k-1}\int_{i\eta}^{(i+1)\eta}\mathbb{E}\|\mathbf{h}(\xi_t)-\mathbf{h}(\xi_{i\eta})\|^2 dt\\
&\leq \frac{\ell^2\beta}{4}\sum_{i=0}^{k-1}\int_{i\eta}^{(i+1)\eta}\mathbb{E}\|\xi_t-\xi_{i\eta}\|^2 dt\\
&= \frac{\ell^2\beta}{4}\sum_{i=0}^{k-1}\int_{i\eta}^{(i+1)\eta}\mathbb{E}\|-(t-i\eta)\mathbf{h}(\xi_{i\eta})+\sqrt{2\beta^{-1}}(B_t-B_{i\eta})\|^2 dt\\
&\leq \frac{\ell^2\beta}{4}\sum_{i=0}^{k-1}\int_{i\eta}^{(i+1)\eta}\left[2(t-i\eta)^2\,\mathbb{E}\left\|\mathbf{h}(\xi_{i\eta})\right\|^2 + p\cdot 4\beta^{-1}(t-i\eta)\right]dt\\
&= \frac{\ell^2\beta}{4}\left[\frac{2}{3}\eta^3\sum_{i=0}^{k-1}\mathbb{E}\left\|\mathbf{h}(\xi_{i\eta})\right\|^2 + k\cdot 2p\beta^{-1}\eta^2\right]\\
&= \frac{\ell^2}{6}\beta\eta^3\sum_{i=0}^{k-1}\mathbb{E}\left\|\mathbf{h}(\xi_{i\eta})\right\|^2 + \frac{\ell^2 p}{2}k\eta^2
\end{aligned}$$

Now recall that $\mathbf{h}$ is M-bounded, therefore, we know,

$$D_{\mathrm{KL}}\left(\mu(\theta_t, 0\leq t\leq k\eta)||\mu(\xi_t, 0\leq t\leq k\eta)\right) \leq \left(\frac{\ell^2 M^2}{6}\beta\eta^3 + \frac{\ell^2 p}{2}\eta^2\right)\cdot k.$$

□

Proof of Lemma 4.1. First, let us focus on the line segment $\{cw_t+(1-c)w_{t+1}, 0\leq c\leq 1\}$, by the mean value theorem, we know there exist a $\tilde{c}\in[0,1]$ such that the following holds

$$L(w_{t+1}) = L(w_t) + \langle \mathbf{b}(w_t), w_{t+1}-w_t\rangle + \frac{1}{2}(w_{t+1}-w_t)^T\mathbf{H}\left(\tilde{c}w_t+(1-\tilde{c})w_{t+1}\right)(w_{t+1}-w_t).$$

Note $w_{t+1} = w_t - \eta\mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)$, let's abbreviate $\mathbf{H}_{\tilde{c}}$ for the Hessian matrix at the middle point,

$$\begin{aligned}
L(w_{t+1}) &= L(w_t) - \eta\langle \mathbf{b}(w_t), \mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)\rangle + \frac{\eta^2}{2}\left[\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\right]^T\mathbf{V}(w_t)^{-1/2}\mathbf{H}_{\tilde{c}}\mathbf{V}(w_t)^{-1/2}\left[\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\right],\\
&\leq L(w_t) - \eta\|\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\|^2 + \frac{\eta^2\gamma}{2}\|\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\|^2,\\
&= L(w_t) - \frac{1}{2\gamma}\|\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\|^2.
\end{aligned}$$

if we choose $\eta = \frac{1}{\gamma}$.

For any w, on line segment $cw + (1-c)w_t$, we can use mean value theorem again,

$$\begin{aligned}
&L(w) - L(w_t) \\
&= \langle \mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t), \mathbf{V}(w_t)^{1/2}(w - w_t)\rangle + \frac{1}{2}(w-w_t)^T \mathbf{H}\left(\tilde{c}w + (1-\tilde{c})w_t\right)(w - w_t) \\
&= \langle \mathbf{V}(w)^{-1/2}\mathbf{b}(w_t), \mathbf{V}(w)^{1/2}(w - w_t)\rangle \\
&\quad + \frac{1}{2}\left[\mathbf{V}(w_t)^{1/2}(w-w_t)\right]^T \mathbf{V}(w_t)^{-1/2}\mathbf{H}_{\tilde{c}}\mathbf{V}(w_t)^{-1/2}\left[\mathbf{V}(w_t)^{1/2}(w-w_t)\right] \\
&\geq \langle \mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t), \mathbf{V}(w_t)^{1/2}(w - w_t)\rangle + \frac{\alpha}{2}\|\mathbf{V}(w_t)^{1/2}(w-w_t)\|^2 \\
&\geq -\frac{1}{2\alpha}\|\mathbf{V}(w_t)^{-1/2}\mathbf{b}(w_t)\|^2.
\end{aligned}$$

Therefore, choose w that attains the minimum of L, combine the above two bounds, we know

$$\begin{aligned}
L(w_{t+1}) - L(w_t) &\leq \frac{\alpha}{\gamma}(L(w) - L(w_t)), \\
L(w_{t+1}) - L(w) &\leq (1 - \frac{\alpha}{\gamma})(L(w_t) - L(w)).
\end{aligned}$$

□

Proof of Theorem 4.1. Mimic the proof in Lemma. 4.1, we have

$$\begin{aligned}
&\mathbb{E}\left\{L(\xi_{t+1})|\xi_t\right\} \\
&= \mathbb{E}\left\{L(\xi_t) + \langle \mathbf{b}(\xi_t), \xi_{t+1} - \xi_t\rangle + \frac{1}{2}(\xi_{t+1}-\xi_t)^T\mathbf{H}\left(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1}\right)(\xi_{t+1}-\xi_t)|\xi_t\right\}, \\
&= L(\xi_t) - \eta\langle \mathbf{b}(\xi_t), \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t)\rangle + \mathbb{E}\left\{\frac{\gamma}{2}\|\eta\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t) + \mathbf{V}(\xi_t)^{1/2}\sqrt{2\beta^{-1}\eta}\mathbf{g}_t\|^2|\xi_t\right\}, \\
&\leq L(\xi_t) - \eta\|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 + \frac{\eta^2\gamma}{2}\|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 + \beta^{-1}\eta\gamma\,\mathbb{E}\left\{\|\mathbf{V}(\xi_t)^{1/2}\mathbf{g}_t\|^2|\xi_t\right\}, \\
&= L(\xi_t) - \frac{1}{2\gamma}\|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 + \beta^{-1}\langle \mathbf{I}_p, \mathbf{V}(\xi_t)\rangle.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\mathbb{E}\,L(\xi_{t+1}) - \min_\xi L(\xi) &\leq (1-\frac{\alpha}{\gamma})(\mathbb{E}\,L(\xi_t) - \min_\xi L(\xi)) + \beta^{-1}\langle \mathbf{I}_p, \mathbb{E}\,\mathbf{V}(\xi_t)\rangle \\
\mathbb{E}\,L(\xi_{t+1}) - \min_\xi L(\xi) &\leq (1-\frac{\alpha}{\gamma})(\mathbb{E}\,L(\xi_t) - \min_\xi L(\xi)) + \beta^{-1}p\cdot\max_\xi\|\mathbf{V}(\xi)\| \\
\mathbb{E}\,L(\xi_k) - \min_\xi L(\xi) &\leq (1-\frac{\alpha}{\gamma})^k(\mathbb{E}\,L(\xi_0) - \min_\xi L(\xi)) + \frac{\beta^{-1}p\cdot\max_\xi\|\mathbf{V}(\xi)\|}{1-(1-\frac{\alpha}{\gamma})}.
\end{aligned}$$

We know $\xi_0 = \theta_0$. It is easily seen that the same argument holds with θ_t as the conditional second moment of the Gaussian approximation using ξ_{t+1} matches θ_{t+1}.

□

Proof of Theorem 4.2. Clearly, we know that $\mathrm{Cov}[\beta(\mathbf{x}, w)\mathbf{x}] \preceq \mathbb{E}[\beta(\mathbf{x}, w)^2\mathbf{x}\mathbf{x}^T]$. Recall that,

$$\mathbf{V}(w) = \left(\mathbb{E}[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T] + \mathrm{Cov}[\beta(\mathbf{x}, w)\mathbf{x}]\right)^{1/2}, \quad \mathbf{H}(w) = \mathbb{E}\left[c''(\mathbf{x}^T w)\mathbf{x}\mathbf{x}^T\right]$$

Therefore, the following matrix inequalities hold

$$\mathbb{E}[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T] \preceq \mathbf{V}(w)^2 \preceq \mathbb{E}[(\xi(\mathbf{x})^2 + \beta(\mathbf{x}, w)^2)\mathbf{x}\mathbf{x}^T]$$

where $A \preceq B$ denotes that $B - A$ being a positive semi-definite matrix.

Under the condition that there exists $C > 1$ such that

$$C^{-1/3} < \frac{c''(\mathbf{x}^T v)}{\xi(x)^2 + \beta(x, w)^2} \leq \frac{c''(x^T v)}{\xi(x)^2} < C^{1/3},$$

then we have

$$\mathbf{H}(v) = \mathbb{E}\left[c''(\mathbf{x}^T v)\mathbf{x}\mathbf{x}^T\right] = \mathbb{E}\left[\frac{c''(\mathbf{x}^T v)}{\xi(\mathbf{x})^2}\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T\right] \prec C^{1/3}\mathbf{V}(w)^2,$$
$$\mathbf{H}(v) = \mathbb{E}\left[c''(\mathbf{x}^T v)\mathbf{x}\mathbf{x}^T\right] = \mathbb{E}\left[\frac{c''(\mathbf{x}^T v)}{\xi(\mathbf{x})^2 + \beta(\mathbf{x}, w)^2}\left(\xi(\mathbf{x})^2 + \beta(\mathbf{x}, w)^2\right)\mathbf{x}\mathbf{x}^T\right] \succ C^{-1/3}\mathbf{V}(w)^2.$$

Let's recall the following facts that if $A \prec B$, then $\lambda_{\max}(A) < \lambda_{\max}(B)$ because take v to be the top unit eigenvector of A,

$$\lambda_{\max}(A) = v^T A v < v^T B v \leq \lambda_{\max}(B).$$

Similarly, we have $\lambda_{\min}(A) < \lambda_{\min}(B)$. Also, if $A \prec B$, then for any symmetric matrix C, $CAC \prec CBC$.

Now because $\mathbf{H}(v) \prec C^{1/3}\mathbf{V}(w)^2$, take w, v that maximize the LHS of the following

$$\lambda_{\max}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right) < C^{1/3}\lambda_{\max}\left([\mathbf{V}(w)]^{-1/2}\mathbf{V}(w)^2[\mathbf{V}(w)]^{-1/2}\right) \leq C^{1/3}\max_w \lambda_{\max}(\mathbf{V}(w)).$$

Similarly, because $C^{-1/3}\mathbf{V}(w)^2 \prec \mathbf{H}(v)$,

$$\lambda_{\min}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right) > C^{-1/3}\lambda_{\min}\left([\mathbf{V}(w)]^{-1/2}\mathbf{V}(w)^2[\mathbf{V}(w)]^{-1/2}\right) \geq C^{-1/3}\min_w \lambda_{\min}(\mathbf{V}(w))$$

Recall the definition of κ_{MasGrad}, we know

$$\begin{aligned}
\kappa_{\text{MasGrad}} &= \frac{\max_{w,v}\lambda_{\max}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right)}{\min_{w,v}\lambda_{\min}\left([\mathbf{V}(w)]^{-1/2}\mathbf{H}(v)[\mathbf{V}(w)]^{-1/2}\right)}, \\
&\leq \frac{C^{1/3}\max_w \lambda_{\max}(\mathbf{V}(w))}{C^{-1/3}\min_w \lambda_{\min}(\mathbf{V}(w))} \\
&\leq C^{2/3}\sqrt{\frac{\max_w \lambda_{\max}(\mathbf{V}(w)^2)}{\min_w \lambda_{\min}(\mathbf{V}(w)^2)}} \leq C\sqrt{\frac{\max_v \lambda_{\max}(\mathbf{H}(v))}{\min_v \lambda_{\min}(\mathbf{H}(v))}} = C\sqrt{\kappa_{\text{GD}}}
\end{aligned}$$

where the last step also uses the fact that

$$C^{-1/3}\mathbf{V}(w)^2 \prec \mathbf{H}(v) \prec C^{1/3}\mathbf{V}(w)^2.$$

□

PROOF OF THEOREM 4.3. Now let's analyze Moment Adjusted Proximal Gradient Descent in Eq. (4.6). For any w, and any $z \in \partial h(w_{t+1})$

$$\begin{aligned}
L(w_{t+1}) &= g(w_{t+1}) + h(w_{t+1}) \\
&\leq \left[g(w_t) + \langle \nabla g(w_t), w_{t+1} - w_t \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2_{\mathbf{H}(\tilde{c})} \right] + [h(w) + \langle z, w_{t+1} - w \rangle] \\
&\leq \left[g(w) + \langle \nabla g(w_t), w_t - w \rangle - \frac{1}{2}\|w_t - w\|^2_{\mathbf{H}(c')} \right] + \left[\langle \nabla g(w_t), w_{t+1} - w_t \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2_{\mathbf{H}(\tilde{c})} \right] \\
&\quad + [h(w) + \langle z, w_{t+1} - w \rangle] \\
&= \left[g(w) + \langle \nabla g(w_t), w_{t+1} - w \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2}\|w_t - w\|^2_{\mathbf{H}(c')} \right] + [h(w) + \langle z, w_{t+1} - w \rangle] \\
&= L(w) + \langle \nabla g(w_t) + z, w_{t+1} - w \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2}\|w_t - w\|^2_{\mathbf{H}(c')}.
\end{aligned} \tag{8.4}$$

Due to the optimality of the proximal updates in Eq. (4.6), we know

$$0 \in \frac{1}{\eta}\mathbf{V}(w_{t+1} - w_t + \eta \mathbf{V}^{-1}\nabla g(w_t)) + \partial h(w_{t+1}),$$

there exists $z \in \partial h(w_{t+1})$ such that

$$\nabla g(w_t) + z = \frac{1}{\eta}\mathbf{V}(w_t - w_{t+1}).$$

Continue with Eq. (8.4), and recall the definition of α, γ, one has

$$\begin{aligned}
L(w_{t+1}) &\leq L(w) + \langle \frac{1}{\eta}\mathbf{V}(w_t - w_{t+1}), w_{t+1} - w \rangle + \frac{1}{2}\|w_{t+1} - w_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2}\|w_t - w\|^2_{\mathbf{H}(c')} \\
&\leq L(w) + \langle \frac{1}{\eta}\mathbf{V}(w_t - w_{t+1}), w_{t+1} - w \rangle + \frac{\gamma}{2}\|w_{t+1} - w_t\|^2_{\mathbf{V}} - \frac{\alpha}{2}\|w_t - w\|^2_{\mathbf{V}}.
\end{aligned}$$

Plug in $w = w_t$, we know if $\eta = \frac{1}{\gamma}$

$$L(w_{t+1}) \leq L(w_t) - (\frac{1}{\eta} - \frac{\gamma}{2})\|w_{t+1} - w_t\|^2_{\mathbf{V}} = L(w_t) - \frac{\gamma}{2}\|w_{t+1} - w_t\|^2_{\mathbf{V}} \leq L(w_t).$$

Plug in $w_* = \arg\min L(w)$, one has

$$\begin{aligned}
L(w_{t+1}) - L(w_*) &\leq \gamma \langle w_t - w_{t+1}, w_{t+1} - w_* \rangle_{\mathbf{V}} + \frac{\gamma}{2}\|w_{t+1} - w_t\|^2_{\mathbf{V}} - \frac{\alpha}{2}\|w_t - w_*\|^2_{\mathbf{V}} \\
&\leq \frac{\gamma - \alpha}{2}\|w_t - w_*\|^2_{\mathbf{V}} - \frac{\gamma}{2}\|w_{t+1} - w_*\|^2_{\mathbf{V}}, \\
\frac{2}{\gamma - \alpha}[L(w_{t+1}) - L(w_*)] &\leq \|w_t - w_*\|^2_{\mathbf{V}} - \frac{\gamma}{\gamma - \alpha}\|w_{t+1} - w_*\|^2_{\mathbf{V}}.
\end{aligned}$$

Sum the above equations for $t = 0, \ldots T-1$, one has

$$\frac{2}{\alpha}\left[\left(\frac{\gamma}{\gamma - \alpha} \right)^T - 1 \right] (L(w_T) - L(w_*)) \leq \frac{2}{\gamma - \alpha} \sum_{t=0}^{T-1} \left(\frac{\gamma}{\gamma - \alpha} \right)^t (L(w_{t+1}) - L(w_*)) \leq \|w_0 - w_*\|^2_{\mathbf{V}}.$$

Therefore we know if

$$T \geq \frac{\gamma}{\alpha} \log \left(\frac{\alpha}{2\epsilon} \|w_0 - w_*\|_{\mathbf{V}}^2 + 1 \right),$$

we have

$$L(w_T) - L(w_*) \leq \epsilon.$$

□

Proof of Theorem 5.1. Denote $C \triangleq \max_\theta \|\mathbf{V}(\theta)\|$. Let's start with the the mean value theorem on the line segment between ξ_{t+1} and ξ_t,

$$\begin{aligned}
&\mathbb{E}\left\{L(\xi_{t+1})|\xi_t\right\} \\
&= \mathbb{E}\left\{ L(\xi_t) + \langle \mathbf{b}(\xi_t), \xi_{t+1} - \xi_t \rangle + \frac{1}{2}(\xi_{t+1} - \xi_t)^T \mathbf{H}\left(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1}\right)(\xi_{t+1} - \xi_t) | \xi_t \right\} \\
&= L(\xi_t) - \eta \langle \mathbf{b}(\xi_t), \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) \rangle + \mathbb{E}\left\{ \frac{\gamma}{2} \|\eta \mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{1/2}\mathbf{g}_t\|^2 | \xi_t \right\} \\
&= L(\xi_t) - \left(\eta - \frac{\eta^2\gamma}{2}\right) \|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 + \beta^{-1}\eta\gamma\, \mathbb{E}\, \|\mathbf{V}(\xi_t)^{1/2}\mathbf{g}_t\|^2 \\
&\leq L(\xi_t) - \left(\eta - \frac{\eta^2\gamma}{2}\right) \|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 + C^{1/2} \cdot p\beta^{-1}\eta\gamma
\end{aligned}$$

Therefore, summing over $t \in [T]$, we have

$$\begin{aligned}
L(\xi_0) - \min L(\theta) + C^{1/2} \cdot p\beta^{-1}\eta\gamma T &\geq \sum_{t=0}^{T-1} \left(\eta - \frac{\eta^2\gamma}{2}\right) \mathbb{E}\|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2, \\
\mathbb{E} \min_{t \leq T} \|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 &\leq \frac{L(\theta_0) - \min L(\theta) + C^{1/2} \cdot p\beta^{-1}\eta\gamma T}{T\left(\eta - \frac{\eta^2\gamma}{2}\right)}.
\end{aligned}$$

Therefore we the choice $\eta = \frac{1}{\gamma}$, we have

$$\mathbb{E} \min_{t \leq T} \|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 \leq \frac{2\gamma(L(\theta_0) - \min L(\theta))}{T} + C^{1/2} \cdot \frac{p}{n}.$$

To obtain an ϵ-stationary point in the sense that $\mathbb{E} \min_{t \leq T} \|\mathbf{b}(w_t)\| \leq \epsilon$, we need to

$$\frac{1}{C^{1/2}} \mathbb{E} \min_{t \leq T} \|\mathbf{b}(\xi_t)\|^2 \leq \mathbb{E} \min_{t \leq T} \|\mathbf{V}(\xi_t)^{-1/2}\mathbf{b}(\xi_t)\|^2 \leq \frac{2\gamma(L(\theta_0) - \min L(\theta))}{T} + C^{1/2} \cdot \frac{p}{n} \leq \frac{\epsilon^2}{C^{1/2}}.$$

Hence, one can choose

$$\begin{aligned}
T &= \frac{C^{1/2}\left[2\gamma(L(w_0) - \min L(w)) + C^{1/2} \cdot p\delta^2\right]}{\epsilon^2}, \\
n &= \frac{T}{\delta^2},
\end{aligned}$$

to ensure

$$\left(\mathbb{E} \min_{t \leq T} \|\mathbf{b}(w_t)\|\right)^2 \leq \mathbb{E} \min_{t \leq T} \|\mathbf{b}(w_t)\|^2 \leq \epsilon^2.$$

And due to Thm. 3.1, we know at the same time

$$D_{\mathrm{TV}}\left(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])\right) \leq \mathcal{O}(\sqrt{\frac{T}{n}}) = C\sqrt{\frac{T}{n}} = \mathcal{O}_\delta(\delta).$$

The total number of samples needed is $N = nT = \mathcal{O}(\epsilon^{-4}\delta^{-2})$. Again, it is easy to see that the same argument holds with θ_t as the conditional second moment of ξ_{t+1} matches that of θ_{t+1}. □

REFERENCES

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

Andrew R Barron. Entropy and the central limit theorem. *The Annals of probability*, pages 336–342, 1986.

Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Rate of convergence and edgeworth-type expansion in the entropic central limit theorem. *Ann. Probab.*, 41(4):2479–2512, 07 2013. . URL https://doi.org/10.1214/12-AOP780.

Sergey G Bobkov, Gennadiy P Chistyakov, and Friedrich Götze. Berry–esseen bounds in the entropic central limit theorem. *Probability Theory and Related Fields*, 159(3-4):435–478, 2014.

Antoine Bordes, Léon Bottou, and Patrick Gallinari. Sgd-qn: Careful quasi-newton stochastic gradient descent. *Journal of Machine Learning Research*, 10(Jul):1737–1754, 2009.

VS Borkar and SK Mitter. A strong approximation theorem for stochastic recursive algorithms. *Journal of optimization theory and applications*, 100(3):499–513, 1999.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564*, 2015.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017a.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017b.

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

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Boris T Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.