arXiv:1705.07477v1 [cs.LG] 21 May 2017

Statistical inference using SGD

Tianyang Li Liu Liu Anastasios Kyrillidis Constantine Caramanis

May 23, 2017

Abstract

We present a novel method for frequentist statistical inference in M-estimation problems, based on stochastic gradient descent (SGD) *with a fixed step size*: we demonstrate that the average of such SGD sequences can be used for statistical inference, after proper scaling. An intuitive analysis using the Ornstein-Uhlenbeck process suggests that such averages are asymptotically normal. From a practical perspective, our SGD-based inference procedure is a first order method, and is well-suited for large scale problems. To show its merits, we apply it to both synthetic and real datasets, and demonstrate that its accuracy is comparable to classical statistical methods, while requiring potentially far less computation.

1 Introduction

In M-estimation, minimizing empirical risk functions (RFs) provides point estimates of the model parameters. Statistical inference then seeks to assess the quality of these estimates, for example, obtaining confidence intervals or solving hypothesis testing problems. A classical result in statistics states that the asymptotic distribution of the empirical RF's minimizer is normal, centered around the population RF's minimizer [23]. Thus, given the mean and covariance of this normal distribution, one can infer a range of values, along with probabilities, to quantify the probability that this interval includes the true minimizer.

The Bootstrap [8, 9] is a classical tool for obtaining estimates of the mean and covariance of this distribution. The Bootstrap operates by generating samples from this distribution. These are obtained by repeating the estimation procedure over different re-samplings of the entire data set. As the data dimensionality and size grow, the Bootstrap becomes increasingly –even prohibitively– expensive.

We follow a different path: we show that inference can also be accomplished by using stochastic gradient descent (SGD) *with a fixed step size over the data set*. Significantly, fixed step-size SGD is by and large the dominant method used for large scale data analysis. We prove, and also demonstrate empirically, that *the average of SGD sequences can be used for statistical inference.* Unlike the Bootstrap, our approach does not require creating many large-size subsamples from the data. Our method only uses first order information from gradient computations, and does not require any second order information. Both of these are important for large scale problems where re-sampling many times, or computing Hessians may be computationally prohibitive.

Outline and main contributions: This paper studies and analyzes a simple, *fixed step size*[1], SGD-based algorithm for inference in M-estimation problems. Our algorithm produces samples, whose covariance converges to the covariance of the M-estimate, without relying on bootstrap-based schemes, and also avoiding direct and costly computation of second order information. Much work has been done on asymptotic normality of SGD, as well as on Stochastic Gradient Langevin Dynamics (and variants) in the Bayesian setting. As we discuss in detail in Section 4, this is the first work to provide finite sample inference results for the quality of the estimates, using fixed step size, and without imposing overly restrictive assumptions on the convergence of fixed step size SGD.

The remainder of the paper is organized as follows. In Section 2 we define the inference problem for M-estimation, and recall basic results of asymptotic normality and how these are used. Section 3 is the main

[1] *Fixed step size* means we use the same step size every iteration, but the step size is smaller with more total number of iterations. *Constant step size* means the step size is constant no matter how many iterations taken.

body of the paper: we provide the algorithm for creating bootstrap-like samples, and also provide the main theorem of this work. As the details are involved, we provide an intuitive analysis of our algorithm and explanation of our main results, using an asymptotic Ornstein-Uhlenbeck process approximation for the SGD process [11, 17, 4, 12, 14], postponing the full proof to the appendix. We specialize our main theorem to the case of linear regression (see supplementary material), and also that of logistic regression. For logistic regression in particular, we require a somewhat different approach, as the logistic regression objective is not strongly convex. In Section 4, we present related work and elaborate how this work differs from existing research in the literature. Finally, in Section 5, we provide parts of our numerical experiments that illustrate the behavior of our algorithm, and corroborate our theoretical findings. We do this using synthetic data for linear and logistic regression, and also by considering the Higgs detection data set [3] and the LIBSVM Splice data set. A considerably expanded set of empirical results is deferred to the appendix.

Supporting our theoretical results, our empirical findings suggest that the SGD inference procedure produces results similar to bootstrap while using far fewer operations, thereby producing a more efficient inference procedure applicable in large scale settings where other approaches fail.

2 Statistical inference for M-estimators

Consider the problem of estimating a set of parameters $\theta^\star \in \mathbb{R}^p$ using n samples $\{X_i\}_{i=1}^n$, drawn from some distribution P on the sample space $\mathcal{X}$. In frequentist inference, we are interested in estimating the minimizer $\theta^\star$ of the population risk:

$$\theta^\star = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E}_P[f(\theta; X)] = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \int_x f(\theta; x)\, \mathrm{d}P(x), \tag{1}$$

where we assume that $f(\cdot; x) : \mathbb{R}^p \to \mathbb{R}$ is real-valued and convex; henceforth, we use $\mathbb{E} \equiv \mathbb{E}_P$, unless otherwise stated. In practice, the distribution P is unknown to us. We thus estimate $\theta^\star$ by solving an empirical risk minimization (ERM) problem, where we use the estimate $\widehat{\theta}$:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f(\theta; X_i). \tag{2}$$

Statistical inference consists of techniques for obtaining confidence intervals about the estimate $\widehat{\theta}$. These can be performed if there is an asymptotic limiting distribution associated with $\widehat{\theta}$ [24]. Indeed, under standard and well-understood regularity conditions, the solution to M-estimation problems satisfies asymptotic normality. That is, the distribution $\sqrt{n}(\widehat{\theta} - \theta^\star)$ converges weakly to a normal:

$$\sqrt{n}(\widehat{\theta} - \theta^\star) \longrightarrow \mathcal{N}(0, H^{\star -1} G^\star H^{\star -1}), \tag{3}$$

where $H^\star = \mathbb{E}[\nabla^2 f(\theta^\star; X)]$, and $G^\star = \mathbb{E}[\nabla f(\theta^\star; X) \nabla f(\theta^\star; X)^\top]$ (Theorem 5.21, [23]). We can therefore use this result, as long as we have a good estimate of the covariance matrix: $H^{\star -1} G^\star H^{\star -1}$. The central goal of this paper is obtaining accurate estimates for $H^{\star -1} G^\star H^{\star -1}$.

A naive way to estimate $H^{\star -1} G^\star H^{\star -1}$ is through the empirical estimator $\widehat{H}^{-1} \widehat{G} \widehat{H}^{-1}$ where:

$$\widehat{H} = \frac{1}{n} \sum_{i=1}^n \nabla^2 f(\widehat{\theta}; X_i) \qquad \text{and} \qquad \widehat{G} = \frac{1}{n} \sum_{i=1}^n \nabla f(\widehat{\theta}; X_i) \nabla f(\widehat{\theta}; X_i)^\top.$$

Beyond calculating $\widehat{H}$ and $\widehat{G}$,[2] this computation requires an inversion of $\widehat{H}$ and matrix-matrix multiplications in order to compute $\widehat{H}^{-1} \widehat{G} \widehat{H}^{-1}$ – a key computational bottleneck in high dimensions. Instead, our method uses SGD to directly estimate $\widehat{H}^{-1} \widehat{G} \widehat{H}^{-1}$.

[2] In the case of maximum likelihood estimation, we have $H^\star = G^\star$ which is called Fisher information, thus the covariance of interest is $H^{\star -1} = G^{\star -1}$. This can be estimated either using $\widehat{H}$ or $\widehat{G}$.

3 Statistical inference using SGD

In this section, we provide our main results, including the algorithm and its theoretical guarantees. We also describe its specialization to logistic regression (linear regression is deferred to the supplementary material).

Consider the optimization problem in (2). For instance, in maximum likelihood estimation (MLE), $f_i(\theta; X_i)$ is a negative log-likelihood function. For simplicity of notation, we use $f_i(\theta)$ and $f(\theta)$ in the rest of the paper.

The SGD algorithm with a fixed step size η, is given by the iteration

$$\theta_{t+1} = \theta_t - \eta g_s(\theta_t), \tag{4}$$

where $g_s(\cdot)$ is an unbiased estimator of the gradient, *i.e.*, $\mathbb{E}[g_s(\theta) \mid \theta] = \nabla f(\theta)$, where the expectation is w.r.t. the stochasticity in the $g_s(\cdot)$ calculation. A classical example of an unbiased estimator of the gradient is $g_s(\cdot) \equiv \nabla f_j(\cdot)$, where j is a uniformly random index over the samples X_j.

Our inference procedure uses the average of t SGD iterations. Denote such sequences as $\bar{\theta}_t$:

$$\bar{\theta}_t = \frac{1}{t}\sum_{i=1}^{t} \theta_i. \tag{5}$$

The algorithm proceeds as follows: Given a sequence of SGD iterates, we use the first SGD iterates $\theta_{-b}, \theta_{-b+1}, \ldots, \theta_0$ as a burn in period; we discard these iterates. Next, for each "segment" of $t + d$ iterates, we use the first t iterates to compute $\bar{\theta}_t^{(i)} = \frac{1}{t}\sum_{j=1}^{t} \theta_j^{(i)}$ and discard the last d iterates, where i indicates the i-th segment. This procedure is illustrated in Figure 1.

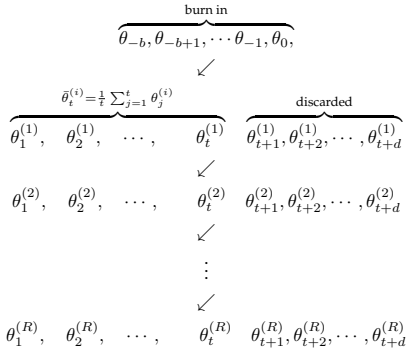


Figure 1: Our SGD inference procedure

Similar to ensemble learning [16], we use $i = 1, 2, \ldots, R$ estimators for statistical inference.

$$\theta^{(i)} = \widehat{\theta} + \frac{\sqrt{K_s}\sqrt{t}}{\sqrt{n}}(\bar{\theta}_t^{(i)} - \widehat{\theta}). \tag{6}$$

Here, K_s is a scaling factor that depends on how the stochastic gradient g_s is computed. We show examples of K_s for mini batch SGD in linear regression and logistic regression in the corresponding sections. In practice, we can use $\widehat{\theta} \approx \frac{1}{R}\sum_{i=1}^{R} \bar{\theta}_t^{(i)}$ [5].

Step size η selection and length t: Theorem 1 below is consistent only for SGD with fixed step size that depends on the number of samples taken. Our experiments, however, demonstrate that choosing a constant (large) η gives equally accurate results with significantly reduced running time. A better understanding of t's and η's influence requires (conjectured) stronger bounds for SGD with constant step size. Heuristically, calibration methods for parameter tuning in subsampling methods ([18], Ch. 9) could be used for hyper-parameter tuning in our SGD procedure. We leave the problem of finding maximal (provable) learning rates for future work.

Discarded length d: Based on the analysis of mean estimation, if we discard d SGD iterates in every segment, the correlation between consecutive $\theta^{(i)}$ and $\theta^{(i+1)}$ is on the order of $C_1 e^{-C_2 \eta d}$, where C_1 and C_2 are data dependent constants. This can be used as a rule of thumb to reduce correlation between samples from our SGD inference procedure.

Burn-in period b: The purpose of the burn-in period b, is to ensure that samples are generated when SGD iterates are sufficiently close to the optimum. This can be determined using heuristics for SGD convergence diagnostics. Another approach is to use other methods (*e.g.*, SVRG [10]) to find the optimum, and use a relatively small b for SGD to reach stationarity, similar to Markov Chain Monte Carlo burn in.

3.1 Theoretical guarantees

Next, we provide the main theorem of our paper. Essentially, this provides conditions under which our algorithm is guaranteed to succeed, and hence has inference capabilities.

Theorem 1. *For a differentiable convex function $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$, with gradient $\nabla f(\theta)$, let $\widehat{\theta} \in \mathbb{R}^p$ be its minimizer, according to* (2)*, and denote its Hessian at $\widehat{\theta}$ by $H := \nabla^2 f(\widehat{\theta})$. Assume that $\forall \theta \in \mathbb{R}^p$, f satisfies:*

(F_1) Weak strong convexity: $(\theta - \widehat{\theta})^\top \nabla f(\theta) \geq \alpha \|\theta - \widehat{\theta}\|_2^2$, for constant $\alpha > 0$,

(F_2) Lipschitz gradient continuity: $\|\nabla f(\theta)\|_2 \leq L\|\theta - \widehat{\theta}\|_2$, for constant $L > 0$,

(F_3) Bounded Taylor remainder: $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 \leq E\|\theta - \widehat{\theta}\|_2^2$, for constant $E > 0$,

(F_4) Bounded Hessian spectrum at $\widehat{\theta}$: $0 < \lambda_L \leq \lambda_i(H) \leq \lambda_U < \infty$, $\forall i$.

Furthermore, let $g_s(\theta)$ be a stochastic gradient of f, satisfying:

(G_1) $\mathbb{E}\left[g_s(\theta) \mid \theta\right] = \nabla f(\theta)$,

(G_2) $\mathbb{E}\left[\|g_s(\theta)\|_2^2 \mid \theta\right] \leq A\|\theta - \widehat{\theta}\|_2^2 + B$,

(G_3) $\mathbb{E}\left[\|g_s(\theta)\|_2^4 \mid \theta\right] \leq C\|\theta - \widehat{\theta}\|_2^4 + D$,

(G_4) $\left\|\mathbb{E}\left[g_s(\theta)g_s(\theta)^\top \mid \theta\right] - G\right\|_2 \leq A_1\|\theta - \widehat{\theta}\|_2 + A_2\|\theta - \widehat{\theta}\|_2^2 + A_3\|\theta - \widehat{\theta}\|_2^3 + A_4\|\theta - \widehat{\theta}\|_2^4$,

for positive, data dependent constants A, B, C, D, A_i, for $i = 1, \ldots, 4$. Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$; then for sufficiently small step size $\eta > 0$, the average SGD sequence in (5) *satisfies:*

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\tfrac{1}{t\eta} + t\eta^2},$$

where $G = \mathbb{E}[g_s(\widehat{\theta})g_s(\widehat{\theta})^\top \mid \widehat{\theta}]$.

We provide the full proof in the appendix, and also we give precise (data-dependent) formulas for the above constants. For ease of exposition, we leave them as constants in the expressions above.

Discussion. For linear regression, assumptions (F_1), (F_2), (F_3), and (F_4) are satisfied when the empirical risk function is not degenerate. In mini batch SGD using sampling with replacement, assumptions (G_1), (G_2), (G_3), and (G_4) are satisfied. Linear regression's result is presented in Corollary 1.

For logistic regression, assumption (F_1) is not satisfied because the empirical risk function in this case is strictly but not strongly convex. Thus, we cannot apply Theorem 1 directly. Instead, we consider the use of SGD on the *square of the empirical risk function plus a constant; see eq.* (12) *below*. When the empirical risk function is not degenerate, (12) satisfies assumptions (F_1), (F_2), (F_3), and (F_4). We cannot directly use vanilla SGD to minimize (12), instead we describe a modified SGD procedure for minimizing (12) in Section 3.5, which satisfies assumptions (G_1), (G_2), (G_3), and (G_4). We believe that this result is of interest by its own. We present the result specialized for logistic regression in Corollary 2.

Note that Theorem 1 proves consistency for SGD with fixed step size, requiring $\eta \to 0$ when $t \to \infty$. However, we empirically observe in our experiments that a sufficiently large *constant* η gives better results. We conjecture that the average of consecutive iterates in SGD with *larger constant step size* converges to the optimum and we consider it for future work.

3.2 Intuitive interpretation via the Ornstein-Uhlenbeck process approximation

Here, we describe a continuous approximation of the discrete SGD process and relate it to the Ornstein-Uhlenbeck process [20], to give an intuitive explanation of our results—the complete proofs appear in the appendix. In particular, under regularity conditions, the stochastic process $\Delta_t = \theta_t - \widehat{\theta}$ asymptotically converges to an Ornstein-Uhlenbeck process $\Delta(t)$, [11, 17, 4, 12, 14] that satisfies:

$$\mathrm{d}\Delta(T) = -H\Delta(T)\,\mathrm{d}T + \sqrt{\eta}G^{\frac{1}{2}}\,\mathrm{d}B(T), \tag{7}$$

where $B(T)$ is a standard Brownian motion. Given (7), $\sqrt{t}(\bar{\theta}_t - \widehat{\theta})$ can be approximated as

$$\sqrt{t}(\bar{\theta}_t - \widehat{\theta}) = \tfrac{1}{\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta}) = \tfrac{1}{\eta\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta})\eta \approx \tfrac{1}{\eta\sqrt{t}}\int_0^{t\eta}\Delta(T)\,\mathrm{d}T, \tag{8}$$

where we use the approximation that $\eta \approx \mathrm{d}T$. By rearranging terms in (7) and multiplying both sides by H^{-1}, we can rewrite the stochastic differential equation (7) as $\Delta(T)\,\mathrm{d}T = -H^{-1}\,\mathrm{d}\Delta(T) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}\,\mathrm{d}B(T)$. Thus, we have

$$\int_0^{t\eta} \Delta(T)\,\mathrm{d}T = -H^{-1}(\Delta(t\eta) - \Delta(0)) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}B(t\eta). \tag{9}$$

After plugging (9) into (8) we have

$$\sqrt{t}\left(\bar{\theta}_t - \widehat{\theta}\right) \approx -\tfrac{1}{\eta\sqrt{t}}H^{-1}\left(\Delta(t\eta) - \Delta(0)\right) + \tfrac{1}{\sqrt{t\eta}}H^{-1}G^{\frac{1}{2}}B(t\eta).$$

When $\Delta(0) = 0$, the variance $\mathrm{Var}\left[-{}^{1}\!/_{\eta\sqrt{t}} \cdot H^{-1}\left(\Delta(t\eta) - \Delta(0)\right)\right] = O\left({}^{1}\!/_{t\eta}\right)$. Since ${}^{1}\!/_{\sqrt{t\eta}} \cdot H^{-1}G^{\frac{1}{2}}B(t\eta) \sim \mathcal{N}(0, H^{-1}GH^{-1})$, when $\eta \to 0$ and $\eta t \to \infty$, we conclude that

$$\sqrt{t}(\bar{\theta}_t - \widehat{\theta}) \sim \mathcal{N}(0, H^{-1}GH^{-1}).$$

3.3 Exact analysis of mean estimation

In this section, we give an exact analysis of our method in the least squares, mean estimation problem. For n i.i.d. samples $X_1, X_2, \ldots, X_n$, the mean is estimated by solving the following optimization problem

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2}\|X_i - \theta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In the case of mini-batch SGD, we sample $S = O(1)$ indexes uniformly randomly with replacement from $[n]$; denote that index set as I_t. For convenience, we write $Y_t = \frac{1}{S}\sum_{i \in I_t} X_i$, Then, in the t^{th} mini batch SGD step, the update step is

$$\theta_{t+1} = \theta_t - \eta(\theta_t - Y_t) = (1-\eta)\theta_t + \eta Y_t, \tag{10}$$

which is the same as the exponential moving average. And we have

$$\sqrt{t}\widehat{\theta}_t = -\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) + \frac{1}{\sqrt{t}}\sum_{i=1}^{n} Y_i. \tag{11}$$

Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$, then from Chebyshev's inequality $-\frac{1}{\eta\sqrt{t}}(\theta_{t+1} - \theta_1) \to 0$ almost surely when $t\eta \to \infty$. By the central limit theorem, $\frac{1}{\sqrt{t}}\sum_{i=1}^{n} Y_i$ converges weakly to $\mathcal{N}(\widehat{\theta}, \frac{1}{S}\hat{\Sigma})$ with $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\theta})(X_i - \widehat{\theta})^\top$. From (10), we have $\|\operatorname{Cov}(\theta_a, \theta_b)\|_2 = O(\eta(1-\eta)^{|a-b|})$ uniformly for all a, b, where the constant is data dependent. Thus, for our SGD inference procedure, we have $\|\operatorname{Cov}(\theta^{(i)}, \theta^{(j)})\|_2 = O(\eta(1-\eta)^{d+t|i-j|})$. Our SGD inference procedure does not generate samples that are independent conditioned on the data, whereas replicates are independent conditioned on the data in bootstrap, but this suggests that our SGD inference procedure can produce "almost independent" samples if we discard sufficient number of SGD iterates in each segment.

When estimating a mean using our SGD inference procedure where each mini batch is S elements sampled with replacement, we set $K_s = S$ in (6).

3.4 Linear Regression

In linear regression, the empirical risk function satisfies:

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2}(\theta^\top x_i - y_i)^2,$$

where y_i denotes the observations of the linear model and x_i are the regressors. To find an estimate to $\theta^\star$, one can use SGD with stochastic gradient give by:

$$g_s[\theta_t] = \frac{1}{S}\sum_{i\in I_t} \nabla f_i(\theta_t),$$

where I_t are S indices uniformly sampled from $[n]$ with replacement.

Next, we state a special case of Theorem 1. Because the Taylor remainder $\nabla f(\theta) - H(\theta - \widehat{\theta}) = 0$, linear regression has a stronger result than general M-estimation problems.

Corollary 1. *Assume that $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$, we have*

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \frac{1}{\sqrt{t\eta}},$$

where $H = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $G = \frac{1}{S}\frac{1}{n}\sum_{i=1}^n (x_i^\top \widehat{\theta} - y_i)^2 x_i x_i^\top$.

We assume that $S = O(1)$ is bounded, and quantities other than t and η are data dependent constants.

As with our main theorem, in the appendix we provide explicit data-dependent expressions for the constants in the result.

Because in linear regression the estimate's covariance is $\frac{1}{n}(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)^{-1})(\frac{1}{n}(x_i^\top\widehat{\theta} - y_i)(x_i^\top\widehat{\theta} - y_i)^\top)(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)^{-1})$, we set the scaling factor $K_s = S$ in (6) for statistical inference.

3.5 Logistic regression

In logistic regression, we have n samples $(X_1, y_1), (X_2, y_2), \dots (X_n, y_n)$ where $X_i \in \mathbb{R}^p$ consists of features and $y_i \in \{+1, -1\}$ is the label. We estimate θ of a linear classifier $\operatorname{sign}(\theta^T X)$ by:

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-y_i\theta^\top X_i)).$$

We cannot apply Theorem 1 directly because the empirical logistic risk is not strongly convex; 'it does not satisfy assumption (F_1). Instead, we consider the convex function

$$f(\theta) = \frac{1}{2}\left(c + \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)^2, \quad \text{where } c > 0 \text{ (e.g., } c = 1). \tag{12}$$

The gradient of $f(\theta)$ is a product of two terms

$$\nabla f(\theta) = \underbrace{\left(c + \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)}_{\Psi} \cdot \underbrace{\nabla\left(\frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)}_{\Upsilon}.$$

Therefore, we can compute a stochastic gradient, $g_s = \Psi_s\Upsilon_s$, *using two independent random variables satisfying* $\mathbb{E}[\Psi_s \mid \theta] = \Psi$ and $\mathbb{E}[\Upsilon_s \mid \theta] = \Upsilon$. For Υ_s, we have $\Upsilon_s = \frac{1}{S_\Upsilon}\sum_{i\in I_t^\Upsilon} \nabla\log(1 + \exp(-y_i\theta^\top X_i))$, where I_t^Υ are S_Υ indices sampled from $[n]$ uniformly at random with replacement. For Ψ_s, we have $\Psi_s = c + \frac{1}{S_\Psi}\sum_{i\in I_t^\Psi}\log(1 + \exp(-y_i\theta^\top X_i))$, where I_t^Ψ are S_Ψ indices uniformly sampled from $[n]$ with or without replacement. Given the above, we have $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some constant α by the generalized self concordance of logistic regression [1, 2], and therefore the assumptions are now satisfied.

For convenience, we write $k(\theta) = \frac{1}{n}\sum_{i=1}^n k_i(\theta)$ where $k_i(\theta) = \log(1 + \exp(-y_i\theta^\top X_i))$. Thus $f(\theta) = (k(\theta + c)^2$, $\mathbb{E}[\Psi_s \mid \theta] = k(\theta) + c$, and $\mathbb{E}[\Upsilon_s \mid \theta] = \nabla k(\theta)$.

Corollary 2. *Assume $\|\theta_1 - \widehat{\theta}\|_2^2 = O(\eta)$; also $S_\Psi = O(1)$, $S_\Upsilon = O(1)$ are bounded. Then, we have*

$$\left\| t\mathbb{E}\left[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top\right] - H^{-1}GH^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\tfrac{1}{t\eta} + t\eta^2},$$

where $H = \nabla^2 f(\widehat{\theta}) = (c + k(\widehat{\theta}))\nabla^2 k(\widehat{\theta})$. Here, $G = \frac{1}{S_\Upsilon}K_G(\widehat{\theta})\frac{1}{n}\sum_{i=1}^n \nabla k_i(\widehat{\theta})k_i(\widehat{\theta})^\top$ with $K_G(\theta) = \mathbb{E}[\Psi(\theta)^2]$ depending on how indexes are sampled to compute Ψ_s:

- *with replacement:* $K_G(\theta) = \frac{1}{S_\Psi}(\frac{1}{n}\sum_{i=1}^n (c + k_i(\theta))^2) + \frac{S_\Psi - 1}{S_\Psi}(c + k(\theta))^2$,
- *without replacement:* $K_G(\theta) = \frac{1 - \frac{S_\Psi - 1}{n-1}}{S_\Psi}(\frac{1}{n}\sum_{i=1}^n (c + k_i(\theta))^2) + \frac{S_\Psi - 1}{S_\Psi}\frac{n}{n-1}(c + k(\theta))^2$.

Quantities other than t and η are data dependent constants.

As with the results above, in the appendix we give data-dependent expressions for the constants. Simulations suggest that the term $t\eta^2$ in our bound is an artifact of our analysis. Because in logistic regression the estimate's covariance is $\frac{1}{n}\left(\nabla^2 k(\widehat{\theta})\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^n \nabla k_i(\widehat{\theta}) \nabla k_i(\widehat{\theta})^\top\right) \cdot \left(\nabla^2 k(\widehat{\theta})\right)^{-1}$, we set the scaling factor $K_s = \frac{(c+k(\widehat{\theta}))^2}{K_G(\widehat{\theta})}$ in (6) for statistical inference. Note that $K_s \approx 1$ for sufficiently large S_Ψ.

4 Related work

Bayesian inference: First and second order iterative optimization algorithms –including stochastic gradient descent, gradient descent, and variants– naturally define a Markov chain. Based on this principle, there is a long line of works focused on creating variants that have a particular steady state distribution. Most related to this work is the case of stochastic gradient Langevin dynamics (SGLD) for Bayesian inference –namely, for sampling from the posterior distributions– using a variant of stochastic gradient descent [25, 6, 14, 15]. We note that, here as well, the vast majority of the results rely on using a decreasing step size. Very recently, [15] uses a heuristic approximation for Bayesian inference, and provides results for fixed step size.

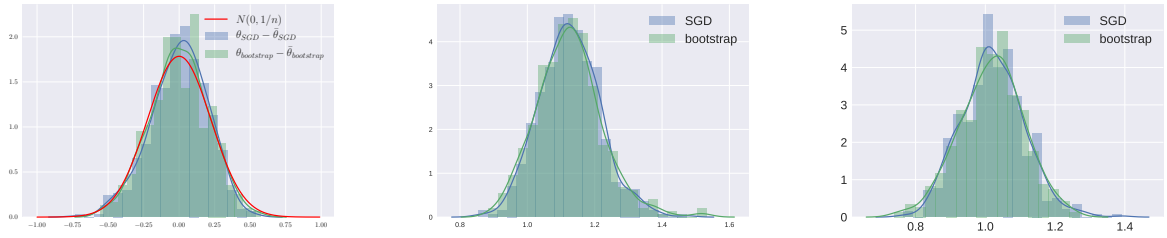
Our problem is different in important ways from the Bayesian inference problem. In such likelihood parameter estimation problems, the covariance of the estimator only depends on the gradient of the likelihood function. This is not the case, however, in general frequentist M-estimation problems (*e.g.*, linear regression), which is exactly the setting of this paper. In these cases, the covariance of the estimator depends both on the gradient and Hessian of the empirical risk function. For this reason, without second order information, SGLD methods are poorly suited for general M-estimation problems in frequentist inference. In contrast, our method exploits properties of averaged SGD, and computes the estimator's covariance without second order information. As we discuss below, a central challenge we face, therefore, is estimating second order information even though SGD's covariance need not converge if using fixed step size. This issue is avoided in the Bayesian setting since only first order information is needed, and (see more below) in the stochastic approximation setting by using decreasing step size.

Connection with Bootstrap methods: While methodologically different, the classical approach for statistical inference is to use the bootstrap [9, 21]. Bootstrap samples are generated by essentially replicating the entire data set by resampling, and then solving the optimization problem (by any means) on each generated set of the data. Our approach offers an alternative to this, using fixed step size SGD. We identify our algorithm and its analysis as an alternative to bootstrap methods. Our analysis is also specific to SGD, and thus sheds light on the statistical properties of this very widely used algorithm.

Connection with stochastic approximation methods: It has been long observed in stochastic approximation that under certain conditions, SGD displays asymptotic normality for both the setting of *decreasing step size*, e.g., [13, 19], and more recently, [22, 7]; and also for *fixed step size*, e.g., [4], Chapter 4. All of these results, however, provide their guarantees with the requirement that the stochastic approximation iterate converges to the optimum. For decreasing step size, this is not an overly burdensome assumption, since with mild assumptions it can be shown directly. As far as we know, however, it is not clear if this holds in the fixed step size regime. To side-step this issue, [4] provides results only when the (constant) step-size approaches 0 (see Section 4.4 and 4.6, and in particular Theorem 7 in [4]). Similarly, while [12] has asymptotic results on the average of consecutive stochastic approximation iterates with constant step size, it assumes convergence of iterates (assumption A1.7 in Ch. 10) – an assumption we are unable to justify in even simple settings.

Indeed, the challenge with SGD is that, when using constant step size, each iterate is distributed around the optimum with non-vanishing variance, and individual iterates do not converge to the optimum.

Beyond the critical difference in the assumptions, the majority of the "classical" subject matter seeks to prove asymptotic results about different flavors of SGD, but does not properly consider its use for inference. Key exceptions are the recent work in [22] and [7], which follow up on [19]. Both of these rely on decreasing step size, for reasons mentioned above. The work in [7] uses SGD with decreasing step size for estimating



(a) Normal. (b) Exponential. (c) Poisson.

Figure 2: Estimation in univariate models.

an M-estimate's covariance. Work in [22] studies implicit SGD with decreasing step size and proves results similar to [19], however it does not use SGD to compute confidence intervals.

Through SGD with constant step size, we can generate almost "symmetric" samples for statistical inference, and rescaling samples for statistical inference is nontrivial in SGD with decreasing step size. To the best of our knowledge, there are no prior results establishing asymptotic normality for SGD with fixed step size for general M-estimation problems (that do not rely on overly restrictive assumptions, as discussed).

5 Experiments

5.1 Synthetic data

The coverage probability is defined as $\frac{1}{p}\sum_{i=1}^{p}\mathbb{P}[\theta_i^\star \in \hat{C}_i]$ where $\theta^\star = \arg\min_\theta \mathbb{E}[f(\theta, X)] \in \mathbb{R}^p$, and $\hat{C}_i$ is the estimated confidence interval for the i^{th} coordinate. The average confidence interval width is defined as $\frac{1}{p}\sum_{i=1}^{p}(\hat{C}_i^u - \hat{C}_i^l)$ where $[\hat{C}_i^l, \hat{C}_i^u]$ is the estimated confidence interval for the i^{th} coordinate. In our experiments, coverage probability and average confidence interval width are estimated through simulation. We use the empirical quantile of our SGD inference procedure and bootstrap to compute the 95% confidence intervals for each coordinate of the parameter. Because theoretical justifications of our SGD inference procedure do not yet deal with pivotal quantities, here we have not included such comparisons. For results given as a pair (α, β), it usually indicates (coverage probability, confidence interval length).

5.1.1 Univariate models

In Figure 2, we compare our SGD inference procedure with (i) Bootstrap and (ii) normal approximation with inverse Fisher information in univariate models. We observe that our method and Bootstrap have similar statistical properties. Figure 7 in the appendix shows Q-Q plots of samples from our SGD inference procedure. *Normal distribution mean estimation:* Figure 2a compares 500 samples from SGD inference procedure and Bootstrap versus the distribution $\mathcal{N}(0, 1/n)$, using $n = 20$ i.i.d. samples from $\mathcal{N}(0, 1)$. We used mini batch SGD described in Sec. 3.3. For the parameters, we used $\eta = 0.8$, $t = 5$, $d = 10$, $b = 20$, and mini batch size of 2. Our SGD inference procedure gives (0.916 , 0.806), Bootstrap gives (0.926 , 0.841), and normal approximation gives (0.922 , 0.851). *Exponential distribution parameter estimation:* Figure 2b compares 500 samples from inference procedure and Bootstrap, using $n = 100$ samples from an exponential distribution with PDF $\lambda e^{-\lambda x}$ where $\lambda = 1$. We used SGD for MLE with mini batch sampled with replacement. For the parameters, we used $\eta = 0.1$, $t = 100$, $d = 5$, $b = 100$, and mini batch size of 5. Our SGD inference procedure gives (0.922, 0.364), Bootstrap gives (0.942 , 0.392), and normal approximation gives (0.922, 0.393). *Poisson distribution parameter estimation:* Figure 2c compares 500 samples from inference procedure and Bootstrap, using $n = 100$ samples from a Poisson distribution with PDF $\lambda^x e^{-\lambda x}$ where $\lambda = 1$. We used SGD for MLE with mini batch sampled with replacement. For the parameters, we used $\eta = 0.1$, $t = 100$, $d = 5$, $b = 100$, and mini batch size of 5. Our SGD inference procedure gives (0.942 , 0.364), Bootstrap gives (0.946 , 0.386), and normal approximation gives (0.960 , 0.393).

η	$t = 100$	$t = 500$	$t = 2500$
0.1	(0.957, 4.41)	(0.955, 4.51)	(0.960, 4.53)
0.02	(0.869, 3.30)	(0.923, 3.77)	(0.918, 3.87)
0.004	(0.634, 2.01)	(0.862, 3.20)	(0.916, 3.70)

(a) Bootstrap (0.941, 4.14), normal approximation (0.928, 3.87)

η	$t = 100$	$t = 500$	$t = 2500$
0.1	(0.949, 4.74)	(0.962, 4.91)	(0.963, 4.94)
0.02	(0.845, 3.37)	(0.916, 4.01)	(0.927, 4.17)
0.004	(0.616, 2.00)	(0.832, 3.30)	(0.897, 3.93)

(b) Bootstrap (0.938, 4.47), normal approximation (0.925, 4.18)

Table 1: Linear regression. *Left*: Experiment 1, *Right*: Experiment 2.

η	$t = 100$	$t = 500$	$t = 2500$
0.1	(0.872, 0.204)	(0.937, 0.249)	(0.9391, 0.258)
0.02	(0.610, 0.112)	(0.871, 0.196)	(0.926, 0.237)
0.004	(0.312, 0.051)	(0.596, 0.111)	(0.86, 0.194)

(a) Bootstrap (0.932, 0.253), normal approximation (0.957, 0.264)

η	$t = 100$	$t = 500$	$t = 2500$
0.1	(0.859, 0.206)	(0.931, 0.255)	(0.947, 0.266)
0.02	(0.600, 0.112)	(0.847, 0.197)	(0.931, 0.244)
0.004	(0.302, 0.051)	(0.583, 0.111)	(0.851, 0.195)

(b) Bootstrap (0.932, 0.245), normal approximation (0.954, 0.256)

Table 2: Logistic regression. *Left*: Experiment 1, *Right*: Experiment 2.

5.1.2 Multivariate models

In these experiments, we set $d = 100$, used mini-batch size of 4, and used 200 SGD samples. In all cases, we compared with Bootstrap using 200 replicates. We computed the coverage probabilities using 500 simulations. Also, we denote $1_p = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top \in \mathbb{R}^p$. Additional simulations comparing covariance matrix computed with different methods are given in Sec. B.1.2.

Linear regression: *Experiment 1:* Results for the case where $X \sim \mathcal{N}(0, I) \in \mathbb{R}^{10}$, $Y = {w^*}^T X + \epsilon$, $w^* = 1_p/\sqrt{p}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2)$ with $n = 100$ samples is given in Table 1a. Bootstrap gives (0.941, 4.14), and confidence intervals computed using the error covariance and normal approximation gives (0.928, 3.87). *Experiment 2:* Results for the case where $X \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.3^{|i-j|}$, $Y = {w^*}^T X + \epsilon$, $w^* = 1_p/\sqrt{p}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2)$ with $n = 100$ samples is given in Table 1b. Bootstrap gives (0.938, 4.47), and confidence intervals computed using the error covariance and normal approximation gives (0.925, 4.18).

Logistic regression: Here we show results for logistic regression trained using vanilla SGD with mini batch sampled with replacement. Results for modified SGD (Sec. 3.5) are given in Sec. B.1.2. *Experiment 1:* Results for the case where $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X \mid Y \sim \mathcal{N}(0.01 Y 1_p/\sqrt{p}, I) \in \mathbb{R}^{10}$ with $n = 1000$ samples is given in Table 2a. Bootstrap gives (0.932, 0.245), and confidence intervals computed using inverse Fisher matrix as the error covariance and normal approximation gives (0.954, 0.256). *Experiment 2:* Results for the case where $\mathbb{P}[Y = +1] = \mathbb{P}[Y = -1] = 1/2$, $X \mid Y \sim \mathcal{N}(0.01 Y 1_p/\sqrt{p}, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.2^{|i-j|}$ with $n = 1000$ samples is given in Table 2b. Bootstrap gives (0.932, 0.253), and confidence intervals computed using inverse Fisher matrix as the error covariance and normal approximation gives (0.957, 0.264).

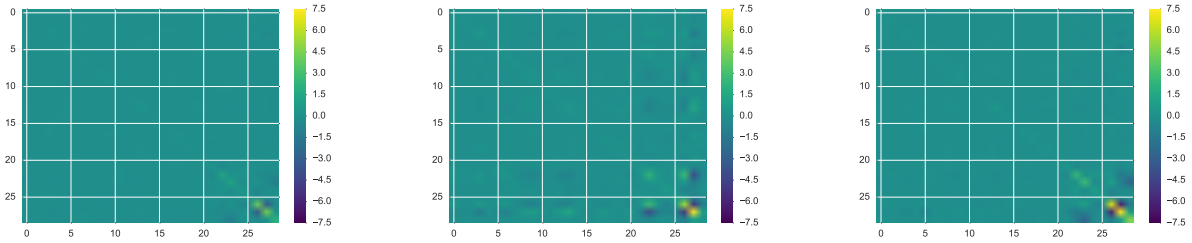
5.2 Real data

Here, we compare covariance matrix computed using our SGD inference procedure, bootstrap, and inverse Fisher information matrix on the Higgs data set [3] and the LIBSVM Splice data set, and we observe that they have similar statistical properties.

5.2.1 Higgs data set

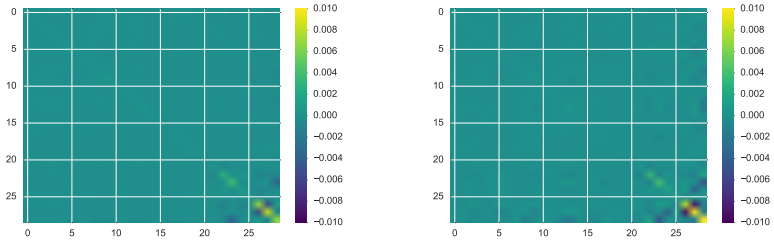
The Higgs data set [3] [3] contains 28 distinct features with 11,000,000 data samples. This is a classification problem between two types of physical processes: one produces Higgs bosons and the other is a background process that does not. We use a logistic regression model, trained using vanilla SGD, instead of the modified SGD described in Section 3.5.

[3] `https://archive.ics.uci.edu/ml/datasets/HIGGS`



(a) Inverse Fisher information (b) SGD inference covariance (c) Bootstrap estimated covariance

Figure 3: Higgs data set with $n = 200$



(a) Inverse Fisher information (b) SGD inference covariance

Figure 4: Higgs data set with $n = 50000$

To understand different settings of sample size, we subsampled the data set with different sample size levels: $n = 200$ and $n = 50000$. We investigate the empirical performance of SGD inference on this subsampled data set. In all experiments below, the batch size of the mini batch SGD is 10.

In the case $n = 200$, the asymptotic normality for the MLE is not a good enough approximation. Hence, in this small-sample inference, we compare the SGD inference covariance matrix with the one obtained by inverse Fisher information matrix and bootstrap in Figure 3.

For our SGD inference procedure, we use $t = 100$ samples to average, and discard $d = 50$ samples. We use $R = 20$ averages from 20 segments (as in Figure 1). For bootstrap, we use 2000 replicates, which is much larger than the sample size $n = 200$.

Figure 3 shows that the covariance matrix obtained by SGD inference is comparable to the estimation given by bootstrap and inverse Fisher information.

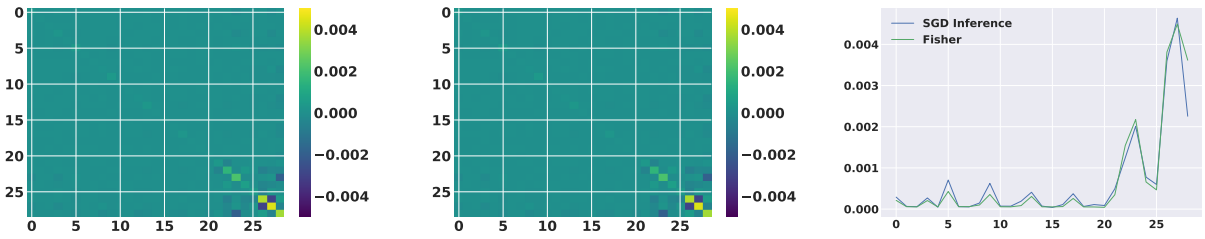
In the case $n = 50000$, we use $t = 5000$ samples to average, and discard $d = 500$ samples. We use $R = 20$ averages from 20 segments (as in Figure 1). For this large data set, we present the estimated covariance of SGD inference procedure and inverse Fisher information (the asymptotic covariance) in Figure 4 because bootstrap is computationally prohibitive. Similar to the small sample result in Figure 3, the covariance of our SGD inference procedure is comparable to the inverse Fisher information.

In Figure 5, we compare the covariance matrix computed using our SGD inference procedure and inverse Fisher information with $n = 90000$ samples . We used 25 samples from our SGD inference procedure with $t = 5000$, $d = 1000$, $\eta = 0.2$, and mini batch size of 10.

5.2.2 Splice data set

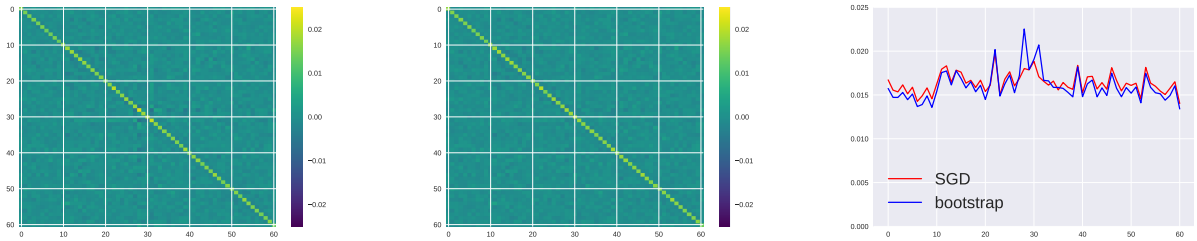
The Splice data set [4] contains 60 distinct features with 1000 data samples. This is a classification problem between two classes of splice junctions in a DNA sequence. Similar to the Higgs data set, we use a logistic regression model, trained using vanilla SGD.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html



(a) Inverse Fisher information (b) SGD inference covariance (c) Diagonal terms

Figure 5: Higgs data set with $n = 90000$



(a) Bootstrap (b) SGD inference covariance (c) Diagonal terms

Figure 6: Splice data set

In Figure 5, we compare the covariance matrix computed using our SGD inference procedure and bootstrap $n = 1000$ samples. We used 10000 samples from both bootstrap and our SGD inference procedure with $t = 500$, $d = 100$, $\eta = 0.2$, and mini batch size of 6.

5.3 Discussion

In our experiments, we observed that using a larger step size η produces accurate results with significantly accelerated convergence time. This might imply that the η term in Theorem 1's bound is an artifact of our analysis. Indeed, although Theorem 1 only applies to SGD with fixed step size, where $\eta t \to \infty$ and $\eta^2 t \to 0$ imply that the step size should be smaller when the number of consecutive iterates used for the average is larger, our experiments suggest that we can use a (data dependent) constant step size η and only require $\eta t \to \infty$.

In the experiments, our SGD inference procedure uses $(t + d) \cdot S \cdot p$ operations to produce a sample, and Newton method uses $n \cdot$(matrix inversion complexity $= \Omega(p^2)) \cdot$(number of Newton iterations t) operations to produce a sample. The experiments therefore suggest that our SGD inference procedure produces results similar to Bootstrap while using far fewer operations.

References

[1] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

[3] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5, 2014.

[4] A. Benveniste, P. Priouret, and M. Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[5] S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, Nov. 2015.

[6] S. Bubeck, R. Eldan, and J. Lehec. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251, 2015.

[7] X. Chen, J. Lee, X. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

[8] B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[9] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[10] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[11] H. Kushner and H. Huang. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981.

[12] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer New York, 2003.

[13] L. Ljung, G. C. Pflug, and H. Walk. *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser, 2012.

[14] S. Mandt, M. Hoffman, and D. Blei. A Variational Analysis of Stochastic Gradient Algorithms. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 354–363, 2016.

[15] S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv preprint arXiv:1704.04289*, 2017.

[16] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[17] G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

[18] D. Politis, J. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer New York, 2012.

[19] B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[20] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[21] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[22] P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *arXiv preprint arXiv:1408.2923*, 2014.

[23] A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[24] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[25] M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

A Proofs

A.1 Proof of Theorem 1

We first assume that $\theta_1 = \widehat{\theta}$ for more precise constants in our bounds, the same analysis applies when $\|\theta_1\|_2^2$

For ease of notation, we denote

$$\Delta_t = \theta_t - \widehat{\theta}, \tag{13}$$

and, without loss of generality, we assume that $\widehat{\theta} = 0$. The stochastic gradient descent recursion satisfies:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \cdot g_s(\theta_t) \\ &= \theta_t - \eta \cdot (g_s(\theta_t) - \nabla f(\theta_t) + \nabla f(\theta_t)) \\ &= \theta_t - \eta \cdot \nabla f(\theta_t) - \eta \cdot e_t,\end{aligned}$$

where $e_t = g_s(\theta_t) - \nabla f(\theta_t)$. Note that $e_1, e_2, \ldots$ is a martingale difference sequence. We use

$$g_i = \nabla f_i(\widehat{\theta}) \qquad \text{and} \qquad H_i = \nabla^2 f_i(\widehat{\theta}) \tag{14}$$

to denote the gradient component at index i, and the Hessian component at index i, at optimum $\widehat{\theta}$, respectively. Note that $\sum g_i = 0$ and $\frac{1}{n}\sum H_i = H$.

For each f_i, its Taylor expansion around $\widehat{\theta}$ is

$$f_i(\theta) = f_i(\widehat{\theta}) + g_i^\top(\theta - \widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^\top H_i(\theta - \widehat{\theta}) + R_i(\theta, \widehat{\theta}), \tag{15}$$

where $R_i(\theta, \widehat{\theta})$ is the remainder term. For convenience, we write $R = \frac{1}{n}\sum R_i$.

For the proof, we require the following lemmata. The following lemma states that $\mathbb{E}[\|\Delta_t\|_2^2] = O(\eta)$ as $t \to \infty$ and $\eta \to 0$.

Lemma 1. *For data dependent, positive constants α, A, B according to assumptions (F_1) and (G_2) in Theorem 1, and given assumption (G_1), we have*

$$\mathbb{E}\left[\|\Delta_t\|_2^2\right] \leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}, \tag{16}$$

under the assumption $\eta < \frac{2\alpha}{A}$.

Proof. As already stated, we assume without loss of generality that $\widehat{\theta} = 0$. This further implies that: $g_s(\theta_t) = g_s(\theta_t - \widehat{\theta}) = g_s(\Delta_t)$, and

$$\Delta_{t+1} = \Delta_t - \eta \cdot g_s(\Delta_t).$$

Given the above and assuming expectation $\mathbb{E}[\cdot]$ w.r.t. the selection of a sample from $\{X_i\}_{i=1}^n$, we have:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^2 \mid \Delta_t\right] &= \mathbb{E}\left[\|\Delta_t - \eta g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] \\ &= \mathbb{E}\left[\|\Delta_t\|_2^2 \mid \Delta_t\right] + \eta^2 \cdot \mathbb{E}\left[\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] - 2\eta \cdot \mathbb{E}\left[g_s(\Delta_t)^\top \Delta_t \mid \Delta_t\right] \\ &= \|\Delta_t\|_2^2 + \eta^2 \cdot \mathbb{E}\left[\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\right] - 2\eta \cdot \nabla f(\Delta_t)^\top \Delta_t \\ &\overset{(i)}{\leq} \|\Delta_t\|_2^2 + \eta^2 \cdot \left(A \cdot \|\Delta_t\|_2^2 + B\right) - 2\eta \cdot \alpha \|\Delta_t\|_2^2 \\ &= (1 - 2\alpha\eta + A\eta^2)\|\Delta_t\|_2^2 + \eta^2 B.\end{aligned} \tag{17}$$

where (i) is due to assumptions (F_1) and (G_2) of Theorem 1. Taking expectations for every step $t = 1, \cdot$ over the whole history, we obtain the recursion:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^2\right] &\leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \eta^2 B \cdot \sum_{i=0}^{t-1}(1 - 2\alpha\eta + A\eta^2)^i \\ &= (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \eta^2 B \cdot \tfrac{1-(1-2\alpha\eta+A\eta^2)^t}{2\alpha\eta - A\eta^2} \\ &\leq (1 - 2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \tfrac{\eta B}{2\alpha - A\eta}.\end{aligned}$$

□

The following lemma states that $\mathbb{E}[\|\Delta_t\|_2^4] = O(\eta^2)$ as $t \to \infty$ and $\eta \to 0$.

Lemma 2. *For data dependent, positive constants α, A, B, C, D according to assumptions (F_1), (G_1), (G_2) in Theorem 1, we have:*

$$\begin{aligned}\mathbb{E}[\|\Delta_t\|_2^4] \leq &(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &+ \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - B(3 + \eta) - C(2\eta^2 + \eta^3)}.\end{aligned} \tag{18}$$

Proof. Given Δ_t, we have the following sets of (in)equalities:

$$\begin{aligned}&\mathbb{E}\left[\|\Delta_{t+1}\|_2^4 \mid \Delta_t\right] \\ =&\mathbb{E}\left[\|\Delta_t - \eta g_s(\Delta_t)\|_2^4 \mid \Delta_t\right] \\ =&\mathbb{E}\left[(\|\Delta_t\|_2^2 - 2\eta \cdot g_s(\Delta_t)^\top \Delta_t + \eta^2\|g_s(\Delta_t)\|_2^2)^2 \mid \Delta_t\right] \\ =&\mathbb{E}\big[\|\Delta_t\|_2^4 + 4\eta^2(g_s(\Delta_t)^\top \Delta_t)^2 + \eta^4\|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t\|\Delta_t\|_2^2 \\ &\qquad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2\|\Delta_t\|_2^2 - 4\eta^3 \cdot g_s(\Delta_t)^\top \Delta_t\|g_s(\Delta_t)\|_2^2 \mid \Delta_t\big] \\ \overset{(i)}{\leq}&\mathbb{E}\big[\|\Delta_t\|_2^4 + 4\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + \eta^4\|g_s(\Delta_t)\|_2^4 - 4\eta \cdot g_s(\Delta_t)^\top \Delta_t\|\Delta_t\|_2^2 \\ &\qquad + 2\eta^2 \cdot \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2 + 2\eta^3 \cdot (\|g_s(\Delta_t)\|_2^2 + \|\Delta_t\|_2^2) \cdot \|g_s(\Delta_t)\|_2^2 \mid \Delta_t\big] \\ \overset{(ii)}{\leq}&\mathbb{E}\left[\|\Delta_t\|_2^4 + (2\eta^3 + \eta^4)\|g_s(\Delta_t)\|_2^4 + (6\eta^2 + 2\eta^3)\|g_s(\Delta_t)\|_2^2\|\Delta_t\|_2^2 \mid \Delta_t\right] - 4\alpha\eta\|\Delta_t\|_2^4 \\ \overset{(iii)}{\leq}&(1 - 4\alpha\eta)\|\Delta_t\|_2^4 + (6\eta^2 + 2\eta^3)(A\|\Delta_t\|_2^2 + B)\|\Delta_t\|_2^2 + (2\eta^3 + \eta^4)(C\|\Delta_t\|_2^4 + D) \\ =&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))\|\Delta_t\|_2^4 + B(6\eta^2 + 2\eta^3)\|\Delta_t\|_2^2 + D(2\eta^3 + \eta^4) \\ \overset{(iv)}{\leq}&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B(3\eta + \eta^2)(\eta^2 + \|\Delta_t\|_2^4) + D(2\eta^3 + \eta^4) \\ =&(1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + B(3\eta + \eta^2) + C(2\eta^3 + \eta^4)) \cdot \|\Delta_t\|_2^4 + B\eta^2(3\eta + \eta^2) + D(2\eta^3 + \eta^4),\end{aligned} \tag{19}$$

where (i) is due to $(g_s(\Delta_t)^\top \Delta_t)^2 \leq \|g_s(\Delta_t)\|_2^2 \cdot \|\Delta_t\|_2^2$ and $-2g_s(\Delta_t)^\top \Delta_t \leq \|g_s(\Delta_t)\|_2^2 + \|\Delta_t\|_2^2$, (ii) is due to assumptions (G_1) and (F_1) in Theorem 1, (iii) is due to assumptions (G_2) and (G_3) in Theorem 1, and (iv) is due to $2\eta\|\Delta_t\|_2^2 \leq \eta^2 + \|\Delta_t\|_2^4$. Similar to the proof of the previous lemma, applying the above rule recursively and w.r.t. the whole history of estimates, we

obtain:

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2^4\right] &\leq (1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &\quad+\left(B\eta^2(3\eta+\eta^2)+D(2\eta^3+\eta^4)\right)\cdot\sum_{i=0}^{t-1}\left(1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4)\right)^i \\ &\leq (1-4\alpha\eta+A(6\eta^2+2\eta^3)+B(3\eta+\eta^2)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4 \\ &\quad+\frac{B\eta^2(3\eta+\eta^2)+D(2\eta^3+\eta^4)}{4\alpha\eta-A(6\eta^2+2\eta^3)-B(3\eta+\eta^2)-C(2\eta^3+\eta^4)},\end{aligned}$$

which is the target inequality, after simple transformations. □

For SGD, we have

$$\begin{aligned}\Delta_t &= (I-\eta H)\Delta_{t-1}-\eta(\nabla R(\Delta_{t-1})+e_{t-1}) \\ &= (I-\eta H)^{t-1}\Delta_1-\eta\sum_{i=1}^{t-1}(I-\eta H)^{t-1-i}(e_i+\nabla R(\Delta_i)).\end{aligned} \tag{20}$$

For $t\geq 2$,

$$\begin{aligned}t(\bar{\theta}-\widehat{\theta}) &= \sum_{i=1}^{\top}\Delta_i \\ &= (I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1-\eta\sum_{j=1}^{t-1}\sum_{i=1}^{j}(I-\eta H)^{j-1-i}(e_i+\nabla R(\Delta_i)).\end{aligned} \tag{21}$$

For the latter term,

$$\begin{aligned}&\eta\sum_{j=1}^{t-1}\sum_{i=1}^{j}(I-\eta H)^{j-i}(e_i+\nabla R(\Delta_i)) \\ =&\eta\sum_{i=1}^{t-1}(\sum_{j=0}^{t-i-1}(I-\eta H)^j)(e_i+\nabla R(\Delta_i)) \\ =&\sum_{i=1}^{t-1}(I-(I-\eta H)^{t-i})H^{-1}(e_i+\nabla R(\Delta_i)) \\ =&H^{-1}\sum_{i=1}^{t-1}e_i+H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)-H^{-1}\sum_{i=1}^{t-1}(I-\eta H)^{t-i}(e_i+\nabla R(\Delta_i)) \\ \overset{(i)}{=}&H^{-1}\sum_{i=1}^{t-1}e_i+H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)+H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1),\end{aligned} \tag{22}$$

where step (i) follows from the fact $\sum_{i=1}^{t-1}(I-\eta H)^{t-i}(e_i+\nabla R(\Delta_i))=(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1)$.

Thus, we have

$$\begin{aligned}\sqrt{t}\bar{\Delta}_t =&\frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1\\ &-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i\\ &-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)\\ &-\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1).\end{aligned} \tag{23}$$

In the statement of the theorem we have $\Delta_1 = 0$ (however similar bounds will hold if $\|\Delta_1\|_2^2 = O(\eta)$), thus for above terms we have

$$\frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1 = 0, \tag{24}$$

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1)\|_2^2]\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\mathbb{E}[\frac{\|\Delta_t\|_2^2}{\eta^2 t}]\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\frac{1}{\eta^2 t}((1-2\alpha\eta+A\eta^2)^{t-1}\|\Delta_1\|_2^2+\frac{B\eta}{2\alpha-A\eta})\\ \leq&\frac{1-\eta\lambda_U}{\lambda_L}\frac{B}{t\eta(2\alpha-A\eta)}\\ =&O(\frac{1}{t\eta}).\end{aligned} \tag{25}$$

$$
\begin{aligned}
&\mathbb{E}[\|\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i)\|_2^2] \\
\leq&\mathbb{E}[\frac{1}{\lambda_L}\frac{1}{t}(\sum_{i=1}^{t-1}\|\nabla R(\Delta_i)\|_2)^2] \\
\leq&\mathbb{E}[\frac{E^2}{\lambda_L t}(\sum_{i=1}^{t-1}\|\Delta_i\|_2^2)^2] \\
\leq&\frac{E^2}{\lambda_L t}(t-1)\mathbb{E}[\sum_{i=1}^{t-1}\|\Delta_i\|_2^4] \\
\leq&\frac{E^2}{\lambda_L}\frac{t}{t-1}\sum_{i=1}^{t-1}((1-4\alpha\eta+A(6\eta^2+2\eta^3)+C(2\eta^3+\eta^4))^{t-1}\|\Delta_1\|_2^4+\frac{B(3\eta^2+\eta^3)+D(2\eta^2+\eta^3)}{4\alpha-A(6\eta+2\eta^2)-C(2\eta^2+\eta^3)}) \\
=&\frac{E^2}{\lambda_L}t\frac{B(3\eta^2+\eta^3)+D(2\eta^2+\eta^3)}{4\alpha-A(6\eta+2\eta^2)-C(2\eta^2+\eta^3)} \\
=&O(t\eta^2).
\end{aligned} \tag{26}
$$

For the term $-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i$, we have

$$
\begin{aligned}
&\mathbb{E}[\|-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i\|_2^2] \\
\overset{(i)}{=}&\frac{1}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|H^{-1}e_i\|_2^2] \\
\leq&\frac{\lambda_U}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|e_i\|_2^2] \\
=&\frac{\lambda_U}{t}\sum_{i=1}^{t-1}\mathbb{E}[\|g_s(\Delta_i)-\nabla f(\Delta_i)\|_2^2] \\
\leq&2\frac{\lambda_U}{t}(\sum_{i=1}^{t-1}\mathbb{E}[\|g_s(\Delta_i)\|_2^2]+\sum_{i=1}^{t-1}\mathbb{E}[\|\nabla f(\Delta_i)\|_2^2]) \\
\leq&2\frac{\lambda_U}{t}((t-1)B+(A+L^2)\sum_{i=1}^{t-1}\|\Delta_i\|_2^2) \\
\leq&2\frac{\lambda_U}{t}((t-1)B+(A+L^2)\sum_{i=1}^{t-1}((1-2\alpha\eta+A\eta^2)^{t-1}\|\Delta_1\|_2^2+\frac{B\eta}{2\alpha-A\eta})) \\
=&2\lambda_U\frac{t-1}{t}(B+(A+L^2)\frac{B\eta}{2\alpha-A\eta}) \\
=&O(1),
\end{aligned} \tag{27}
$$

where step (i) follows from $i \neq j$ leading to $\mathbb{E}[(H^{-1}e_i)^\top H^{-1}e_j] = 0$. We also have

$$\begin{aligned}&\mathbb{E}[(-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i)(-\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i)^\top]\\=&\frac{1}{t}H^{-1}(\sum_{i=1}^{t-1}\mathbb{E}[e_ie_i^\top])H^{-1}.\end{aligned} \tag{28}$$

For each term $\mathbb{E}[e_ie_i^\top]$, we have

$$\begin{aligned}&\|\mathbb{E}[e_ie_i^\top] - G\|_2\\=&\|\mathbb{E}[g_s(\Delta_i)g_s(\Delta_i)^\top] - \mathbb{E}[(\nabla f(\Delta_i))(\nabla f(\Delta_i))^\top] - G\|_2\\\leq&\mathbb{E}[\|\nabla f(\Delta_i)\|_2^2] + \mathbb{E}[A_1\|\Delta_i\|_2 + A_2\|\Delta_i\|_2^2 + A_3\|\Delta_i\|_2^3 + A_4\|\Delta_i\|_2^4]\\\leq&L^2\mathbb{E}[\|\Delta_i\|_2^2] + A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + A_2\mathbb{E}[\|\Delta_i\|_2^2] + \frac{A_3}{2}\mathbb{E}[\|\Delta_i\|_2^2 + \|\Delta_i\|_2^4] + A_4\mathbb{E}[\|\Delta_i\|_2^4]\\=&A_1\sqrt{\mathbb{E}[\|\Delta_i\|_2^2]} + (L^2 + A_2 + \frac{A_3}{2})\mathbb{E}[\|\Delta_i\|_2^2] + (\frac{A_3}{2} + A_4)\mathbb{E}[\|\Delta_i\|_2^4]\\\leq&A_1\sqrt{(1-2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})((1-2\alpha\eta + A\eta^2)^{t-1}\|\Delta_1\|_2^2 + \frac{B\eta}{2\alpha - A\eta})\\&+ (\frac{A_3}{2} + A_4)((1 - 4\alpha\eta + A(6\eta^2 + 2\eta^3) + C(2\eta^3 + \eta^4))^{t-1}\|\Delta_1\|_2^4 + \frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)})\\=&A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})\frac{B\eta}{2\alpha - A\eta} + (\frac{A_3}{2} + A_4)\frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)}.\end{aligned} \tag{29}$$

Thus, we have

$$\begin{aligned}&\|\frac{1}{t}H^{-1}(\sum_{i=1}^{t-1}\mathbb{E}[e_ie_i^\top])H^{-1} - H^{-1}GH^{-1}\|_2\\\leq&\frac{1}{t}\|H^{-1}GH^{-1}\|_2\\&+ \frac{t-1}{t}\frac{1}{\lambda_L^2}(A_1\sqrt{\frac{B\eta}{2\alpha - A\eta}} + (L^2 + A_2 + \frac{A_3}{2})\frac{B\eta}{2\alpha - A\eta} + (\frac{A_3}{2} + A_4)\frac{B(3\eta^2 + \eta^3) + D(2\eta^2 + \eta^3)}{4\alpha - A(6\eta + 2\eta^2) - C(2\eta^2 + \eta^3)})\\=&O(\sqrt{\eta}).\end{aligned} \tag{30}$$

For convenience, denote

$$\begin{aligned}
\square_0 &= \frac{1}{\sqrt{t}}(I-(I-\eta H)^t)\frac{H^{-1}}{\eta}\Delta_1,\\
\square_1 &= -\frac{1}{\sqrt{t}}H^{-1}(I-\eta H)\frac{1}{\eta}(\Delta_t-(I-\eta H)^{t-1}\Delta_1),\\
\square_2 &= -\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}\nabla R(\Delta_i),\\
\square_3 &= -\frac{1}{\sqrt{t}}H^{-1}\sum_{i=1}^{t-1}e_i,
\end{aligned} \tag{31}$$

and we have $\mathbb{E}[t\bar{\Delta}_t\bar{\Delta}_t]=\mathbb{E}[(\square_0+\square_1+\square_2+\square_3)(\square_0+\square_1+\square_2+\square_3)^\top]$.

Combining above results, we can bound

$$\begin{aligned}
&\|t\mathbb{E}[(\bar{\theta}_t-\widehat{\theta})(\bar{\theta}_t-\widehat{\theta})^\top]-H^{-1}GH^{-1}\|_2\\
=&\|\mathbb{E}[(\square_0+\square_1+\square_2+\square_3)(\square_0+\square_1+\square_2+\square_3)^\top]-H^{-1}GH^{-1}\|_2\\
=&\|\mathbb{E}[\square_3\square_3^\top]-H^{-1}GH^{-1}+\mathbb{E}[\square_3(\square_0+\square_1+\square_2)^\top+(\square_0+\square_1+\square_2)\square_3^\top+(\square_0+\square_1+\square_2)(\square_0+\square_1+\square_2)^\top]\|_2\\
\lesssim&\|\mathbb{E}[\square_3\square_3^\top]-H^{-1}GH^{-1}\|_2+\sqrt{\mathbb{E}[\|\square_3\|_2^2](\mathbb{E}[\|\square_0\|_2^2]+\mathbb{E}[\|\square_1\|_2^2]+\mathbb{E}[\|\square_2\|_2^2])}+\mathbb{E}[\|\square_0\|_2^2]+\mathbb{E}[\|\square_1\|_2^2]+\mathbb{E}[\|\square_2\|_2^2]\\
\lesssim&\sqrt{\eta}+\sqrt{\frac{1}{t\eta}+t\eta^2}.
\end{aligned} \tag{32}$$

Here we have used the fact that for two p-dimensional random vectors a and b, the expectation of the matrix $ab^\top$ satisfies

$$\|\mathbb{E}[ab^\top]\|_2\le\sqrt{\mathbb{E}[\|a\|_2^2]\mathbb{E}[\|b\|_2^2]}\le\frac{1}{2}\mathbb{E}[\|a\|_2^2]+\mathbb{E}[\|b\|_2^2]. \tag{33}$$

Indeed, for any fixed unit vector u we have $\|\mathbb{E}[ab^\top]u\|_2=\|\mathbb{E}[a(b^\top u)]\|_2\le\mathbb{E}[\|a\|_2|b^\top u|]\le\mathbb{E}[\|a\|_2\|b\|_2]\le\sqrt{\mathbb{E}[\|a\|_2^2]\mathbb{E}[\|b\|_2^2]}$. Here we used the fact $\|\mathbb{E}[x]\|_2\le\mathbb{E}[\|x\|_2]$ because $\|x\|_2$ is convex.

■

A.2 Proof of Corollary 1

Proof of Corollary 1. Here we use the same notations as the proof of Theorem 1.

Because linear regression satisfies $\nabla f(\theta)-H(\theta-\widehat{\theta})=0$, we do not have to consider the Taylor remainder term in our analysis. And we do not need 4-th order bound for SGD.

Because the quadratic function is strongly convex, we have $\Delta^\top\nabla f(\Delta+\widehat{\theta})\ge\lambda_L\|\Delta\|_2^2$.

By sampling with replacement, we have

$$
\begin{aligned}
&\mathbb{E}[\|g_s(\theta_t)\|_2^2 \mid \theta_t] \\
=&\|\nabla f(\theta_t)\|_2^2 + \mathbb{E}[\|e_t\|_2^2 \mid \theta_t] \\
=&\|\nabla f(\theta_t)\|_2^2 + \frac{1}{S}(\frac{1}{n}\sum \|\nabla f_i(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2^2) \\
\leq& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + \frac{1}{S}\frac{1}{n}\sum \|x_i(x_i^\top \theta_t - y_i)\|_2^2 \\
=& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + \frac{1}{S}\frac{1}{n}\sum \|x_i x_i^\top \Delta_t + x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2 \\
\leq& L^2(1-\frac{1}{S})\|\Delta_t\|_2^2 + 2\frac{1}{S}\frac{1}{n}\sum (\|x_i x_i^\top \Delta_t\|_2^2 + \|x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2) \\
\leq& (L^2(1-\frac{1}{S}) + 2\frac{1}{S}\frac{1}{n}\sum \|x_i\|_2^4)\|\Delta_t\|_2^2 + 2\frac{1}{S}\frac{1}{n}\sum \|x_i x_i^\top \widehat{\theta} - y_i x_i\|_2^2.
\end{aligned} \tag{34}
$$

We also have

$$
\begin{aligned}
&\|\mathbb{E}[g_s(\theta) g_s(\theta)^\top \mid \theta] - G\|_2 \\
=&\|\frac{1}{S}\frac{1}{n}\sum \nabla f_i(\theta) f_i(\theta)^\top - \nabla f(\theta)\nabla f(\theta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum \nabla f_i(\theta) f_i(\theta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum (g_i + H_i\Delta)(g_i + H_i\Delta)^\top - G\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}\|\frac{1}{n}\sum H_i\Delta g_i^\top + g_i\Delta^\top H_i + H_i\Delta\Delta^\top H_i\|_2 \\
\leq& \|\nabla f(\theta)\|_2^2 + \frac{1}{S}(\frac{2}{n}\|H_i\|_2\|g_i\|_2)\|\Delta\|_2 + \frac{1}{S}(\frac{1}{n}\sum \|H_i\|_2^2)\|\Delta\|_2^2 \\
\leq& \frac{1}{S}(\frac{2}{n}\|H_i\|_2\|g_i\|_2)\|\Delta\|_2 + (L^2 + \frac{1}{S}\frac{1}{n}\sum \|H_i\|_2^2)\|\Delta\|_2^2,
\end{aligned} \tag{35}
$$

where $g_i = x_i(x_i^\top \widehat{\theta} - y_i)$ and $H_i = x_i x_i^\top$.

Following Theorem 1's proof, we have

$$
\|t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1}GH^{-1}\|_2 \lesssim \sqrt{\eta} + \frac{1}{\sqrt{t\eta}}. \tag{36}
$$

□

A.3 Proof of Corollary 2

Proof of Corollary 2. Here we use the same notations as the proof of Theorem 1.

Because $\nabla^2 f(\theta) = \nabla k(\theta)\nabla k(\theta)^\top + (k(\theta)+c)\nabla^2 k(\theta)$, $f(\theta)$ is convex.

The following lemma shows that $\nabla f(\theta) = (k(\theta)+c)\nabla k(\theta)$ is Lipschitz.

Lemma 3.

$$
\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2 \tag{37}
$$

for some data dependent constant L.

Proof. First, because

$$\nabla k(\theta) = \frac{1}{n}\sum -\frac{-y_i x_i}{1+\exp(y_i\theta^\top x_i)}, \tag{38}$$

we have

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{n}\sum \|x_i\|_2. \tag{39}$$

Also, we have

$$\begin{aligned}\|\nabla^2 k(\theta)\|_2 =& \|\frac{1}{n}\sum \frac{\exp(y_i\theta^\top x_i)}{(1+\exp(y_i\theta^\top x_i))^2} x_i x_i^\top\|_2 \\ \leq& \frac{1}{4}\frac{1}{n}\sum \|x_i\|_2^2,\end{aligned} \tag{40}$$

which implies

$$\|\nabla k(\theta)\|_2 \leq \frac{1}{4}\frac{1}{n}\sum \|x_i\|_2^2\|\Delta\|_2. \tag{41}$$

And, we have

$$\begin{aligned}k(\theta) =& \frac{1}{n}\sum \log(1+\exp(-y_i\Delta^\top x_i - y_i\widehat{\theta}^\top x_i)) \\ \leq& \frac{1}{n}\sum \log(1+\exp(\|x_i\|_2\|\Delta\|_2 - y_i\widehat{\theta}^\top x_i)) \\ \overset{(i)}{\leq}& \frac{1}{n}\sum (\log(1+\exp(-y_i\widehat{\theta}^\top x_i)) + \|x_i\|_2\|\Delta\|_2)\end{aligned} \tag{42}$$

where step (i) follows from $\log(1+\exp(a+b)) \leq \log(1+e^b)+|a|$. Thus, we have

$$\begin{aligned}&\|\nabla f(\theta)\|_2 \\ =&\|(k(\theta)+c)\nabla k(\theta)\|_2 \\ \leq& k(\theta)\|\nabla k(\theta)\|_2 + c\|\nabla k(\theta)\|_2 \\ \leq& (c+\frac{1}{n}\sum \log(1+\exp(-y_i\widehat{\theta}^\top x_i)))\|\nabla k(\theta)\|_2 + (\frac{1}{n}\sum \|x_i\|_2)^2\|\Delta\|_2,\end{aligned} \tag{43}$$

and we can conclude that $\|\nabla f(\theta)\|_2 \leq L\|\Delta\|_2$ for some data dependent constant L. □

Next, we show that $f(\theta)$ has a bounded Taylor remainder.

Lemma 4.

$$\|\nabla f(\theta) - H(\theta-\widehat{\theta})\|_2 \leq E\|\theta-\widehat{\theta}\|_2^2, \tag{44}$$

for some data dependent constant E.

Proof. Because $\nabla f(\theta) = (k(\theta) + c)\nabla k(\theta)$, we know that $\|\nabla f(\theta)\|_2 = O(\|\Delta\|_2)$ when $\|\Delta\|_2 = \Omega(1)$ where the constants are data dependent.

Because $f(\theta)$ is infinitely differentiable, by the Taylor expansion we know that $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 = O(\|\theta - \widehat{\theta}\|_2^2)$ when $\|\Delta\|_2 = O(1)$ where the constants are data dependent.

Combining the above, we can conclude $\|\nabla f(\theta) - H(\theta - \widehat{\theta})\|_2 \leq E\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant E.

□

In the following lemma, we will show that $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant α.

Lemma 5.

$$\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2, \tag{45}$$

for some data dependent constant α.

Proof.

$$\nabla f(\theta)^\top\Delta = (k(\theta) + c)\nabla k(\theta)^\top\Delta. \tag{46}$$

First, notice that locally (when $\|\Delta\|_2 = O(\frac{\lambda_L}{E})$) we have

$$\nabla k(\theta)^\top\Delta \gtrsim \Delta^\top H\Delta \gtrsim \lambda_L\|\Delta\|_2^2, \tag{47}$$

because of the optimality condition. This lower bounds $\nabla f(\theta)^\top(\theta - \widehat{\theta})$ when $\|\Delta\|_2 = O(\frac{\lambda_L}{E})$. Next we will lower bound it when $\|\Delta\|_2 = \Omega(\frac{\lambda_L}{E})$.

Consider the function for $t \in [0, \infty)$, we have

$$\begin{aligned} g(t) &= \nabla f(\widehat{\theta} + ut)^\top ut \\ &= (k(\widehat{\theta} + ut) + c)\nabla k(\widehat{\theta} + ut)^\top ut \\ &= k(\widehat{\theta} + ut)\nabla k(\widehat{\theta} + ut)^\top ut + c\nabla k(\widehat{\theta} + ut)^\top ut, \end{aligned} \tag{48}$$

where $u = \frac{\Delta}{\|\Delta\|_2}$.

Because $k(\theta)$ is convex, $\nabla k(\widehat{\theta} + ut)^\top u$ is an increasing function in t, thus we have $\nabla k(\widehat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. And we can deduce $\nabla k(\widehat{\theta} + ut)^\top ut = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Similarly, because $k(\theta)$ is convex, $k(\widehat{\theta} + ut)$ is an increasing function in t. Its derivative $\nabla k(\widehat{\theta} + ut)^\top u = \Omega(\frac{\lambda_L^2}{E})$ when $t = \Omega(\frac{\lambda_L}{E})$. So we have $k(\widehat{\theta} + ut) = \Omega(\frac{\lambda_L^2}{E}t)$ when $t = \Omega(\frac{\lambda_L}{E})$.

Thus, we have

$$k(\widehat{\theta} + ut)\nabla k(\widehat{\theta} + ut)^\top ut = \Omega(\frac{\lambda_L^4}{E^2}t^2), \tag{49}$$

when $t = \Omega(\frac{E}{\lambda_L})$.

And we can conclude that $\nabla f(\theta)^\top(\theta - \widehat{\theta}) \geq \alpha\|\theta - \widehat{\theta}\|_2^2$ for some data dependent constant $\alpha = \Omega(\min\{\lambda_L, \frac{\lambda_L^4}{E^2}\})$. □

Next, we will prove properties about $g_s = \Psi_s \Upsilon_s$.

$$\begin{aligned}
&\mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] = \frac{1}{S_\Upsilon}(\frac{1}{n}\sum \|\nabla k_i(\theta)\|_2^2 - \|\nabla k(\theta)\|_2^2) + \|\nabla k(\theta)\|_2^2 \\
\lesssim &\frac{1}{n}\|x_i\|_2^2
\end{aligned} \tag{50}$$

$$\begin{aligned}
&\mathbb{E}[\Psi_s^2] \\
\overset{(i)}{\leq} &\frac{1}{n}\sum (c + k_i(\theta))^2 \\
= &\frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)))^2 \\
\overset{(ii)}{\lesssim} &\frac{1}{n}\sum \|x_i\|^2 \|\Delta\|_2^2 + \frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)))^2,
\end{aligned} \tag{51}$$

where (i) follows from $\mathbb{E}[(\frac{\sum_{j=1}^S X_j}{S})^2] \leq \mathbb{E}[\frac{\sum_{j=1}^S X_j^2}{S}]$ and (ii) follows from $\log(1 + \exp(a+b)) \leq \log(1+e^b) + |a|$.

Thus we have

$$\begin{aligned}
&\mathbb{E}[\|g_s\|_2^2(\theta) \mid \theta] \\
= &\mathbb{E}[\Psi^2 \mid \theta]\mathbb{E}[\|\Upsilon\|_2^2 \mid \theta] \\
\lesssim &A\|\Delta\|_2^2 + B
\end{aligned} \tag{52}$$

for some data dependent constants A and B.

$$\begin{aligned}
&\mathbb{E}[\|\Upsilon\|_2^4 \mid \theta] \\
= &\mathbb{E}[\|\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \nabla \log(1 + \exp(-y_i \theta^\top x_i))\|_2^4] \\
\leq &\mathbb{E}[(\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \|\nabla \log(1 + \exp(-y_i \theta^\top x_i))\|_2)^4] \\
\leq &\mathbb{E}[(\frac{1}{S_\Upsilon}\sum_{i \in I_t^\Upsilon} \|x_i\|_2)^4] \\
\leq &\frac{1}{n}\sum \|x_i\|_2^4.
\end{aligned} \tag{53}$$

$$\begin{aligned}
&\mathbb{E}[\Psi_s^4] \\
\overset{(i)}{\leq} &\frac{1}{n}\sum (c + k_i(\theta))^4 \\
= &\frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i - y_i \Delta x_i)))^4 \\
\overset{(ii)}{\lesssim} &\frac{1}{n}\sum \|x_i\|^4 \|\Delta\|_2^4 + \frac{1}{n}\sum (c + \log(1 + \exp(-y_i \widehat{\theta}^\top x_i)))^4,
\end{aligned} \tag{54}$$

where (i) follows from $\mathbb{E}[(\frac{\sum_{j=1}^{S} X_j}{S})^4] \leq \mathbb{E}[\frac{\sum_{j=1}^{S} X_j^4}{S}]$ and (ii) follows from $\log(1+\exp(a+b)) \leq \log(1+e^b)+|a|$.

Thus we have

$$\begin{aligned}&\mathbb{E}[\|g_s\|_2^4(\theta) \mid \theta]\\ =&\mathbb{E}[\Psi^4 \mid \theta]\mathbb{E}[\|\Upsilon\|_2^4 \mid \theta]\\ \lesssim& C\|\Delta\|_2^4 + D,\end{aligned} \tag{55}$$

for some data dependent constants C and D.

$$\begin{aligned}&\|\mathbb{E}[\nabla g_s(\theta)\nabla g_s(\theta)^\top] - G\|_2\\ \leq&\|K_G(\theta)\frac{1}{n}\sum \nabla k_i(\theta)\nabla k_i(\theta)^\top - K_G(\widehat{\theta})\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2\\ \leq&\|K_G(\theta)\frac{1}{n}\sum \nabla k_i(\theta)\nabla k_i(\theta)^\top - K_G(\theta)\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top + K_G(\theta)\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top - K_G(\widehat{\theta})\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2\\ \leq&K_G(\theta)\frac{1}{n}\|\sum(\nabla k_i(\theta)\nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top)\|_2 + |K_G(\theta) - K_G(\widehat{\theta})|\|\frac{1}{n}\sum \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top\|_2.\end{aligned} \tag{56}$$

Because

$$K_G(\theta) = O(1+\|\Delta\|_2+\|\Delta\|_2^2), \tag{57}$$

$$\frac{1}{n}\|\sum(\nabla k_i(\theta)\nabla k_i(\theta)^\top - \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top)\|_2 = O(\|\Delta\|_2+\|\Delta\|_2^2), \tag{58}$$

$$|K_G(\theta) - K_G(\widehat{\theta})| = O(\|\Delta\|_2+\|\Delta\|_2^2), \tag{59}$$

where we have data dependent constants.

Then, we have

$$\|\mathbb{E}[g_s(\theta)g_s(\theta)^\top \mid \theta] - G\|_2 \leq A_1\|\theta-\widehat{\theta}\|_2 + A_2\|\theta-\widehat{\theta}\|_2^2 + A_3\|\theta-\widehat{\theta}\|_2^3 + A_4\|\theta-\widehat{\theta}\|_2^4, \tag{60}$$

for some data dependent constants A_1, A_2, A_3, and A_4.

Combining above results and using Theorem 1, we have

$$\begin{aligned}&\|t\mathbb{E}[(\bar{\theta}_t-\widehat{\theta})(\bar{\theta}_t-\widehat{\theta})^\top] - H^{-1}GH^{-1}\|_2\\ \lesssim&\sqrt{\eta} + \sqrt{\frac{1}{t\eta}+t\eta^2}.\end{aligned} \tag{61}$$

□

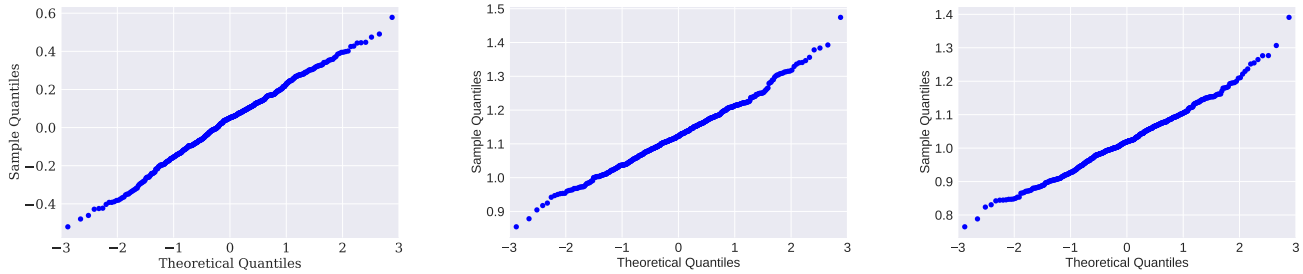
B Experiments

Here we present additional experiments on our SGD inference procedure.

B.1 Synthetic data

B.1.1 Univariate models

Figure 7 shows Q-Q plots for samples shown in Figure 2.



(a) Normal. (b) Exponential. (c) Poisson.

Figure 7: Estimation in univariate models: Q-Q plots for samples shown in Figure 2

B.1.2 Multivariate models

Here we show Q-Q plots per coordinate for samples from our SGD inference procedure.

Q-Q plots per coordinate for samples in linear regression experiment 1 is shown in Figure 8. Q-Q plots per coordinate for samples in linear regression experiment 2 is shown in Figure 9.

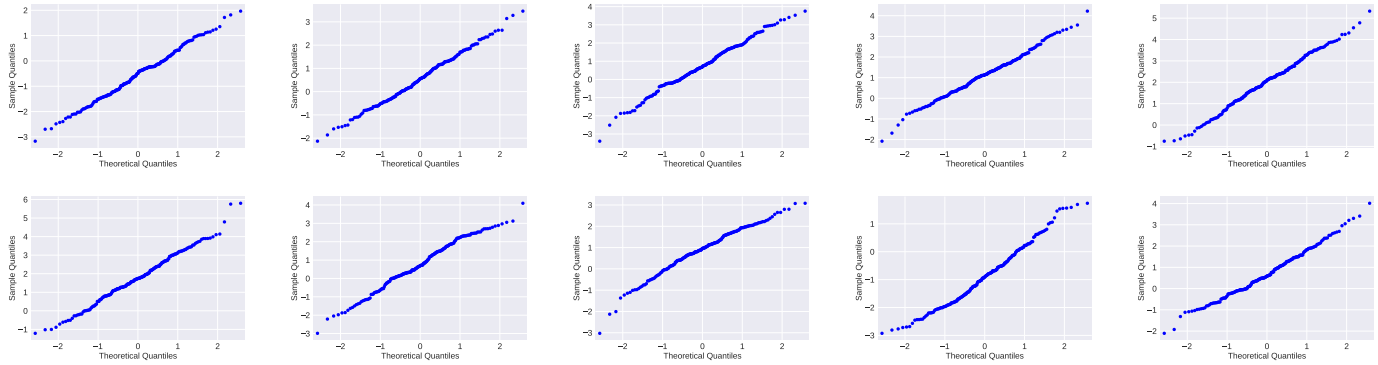


Figure 8: Linear regression experiment 1: Q-Q plots per coordinate

Q-Q plots per coordinate for samples in logistic regression experiment 1 is shown in Figure 10. Q-Q plots per coordinate for samples in logistic regression experiment 2 is shown in Figure 11.

Additional experiments

2-Dimensional Linear Regression. Consider:

$$y = x_1 + x_2 + \epsilon, \qquad \text{where } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right) \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2).$$

Each sample consists of $Y = y$ and $X = [x_1, \ x_2]^\top$. We use linear regression to estimate w_1, w_2 in $y = w_1 x_1 + w_2 x_2$. In this case, the minimizer of the population least square risk is $w_1^\star = 1, w_2^\star = 1$.

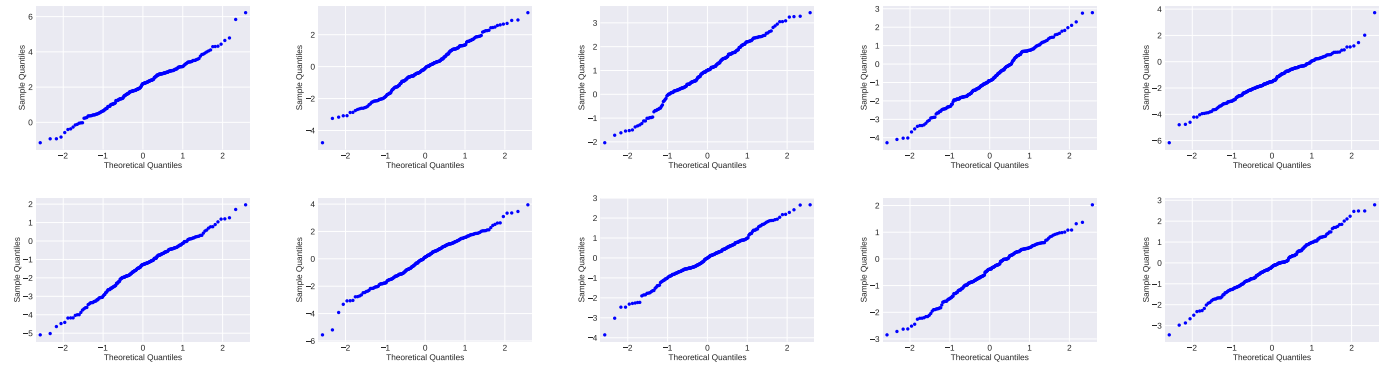


Figure 9: Linear regression experiment 2: Q-Q plots per coordinate

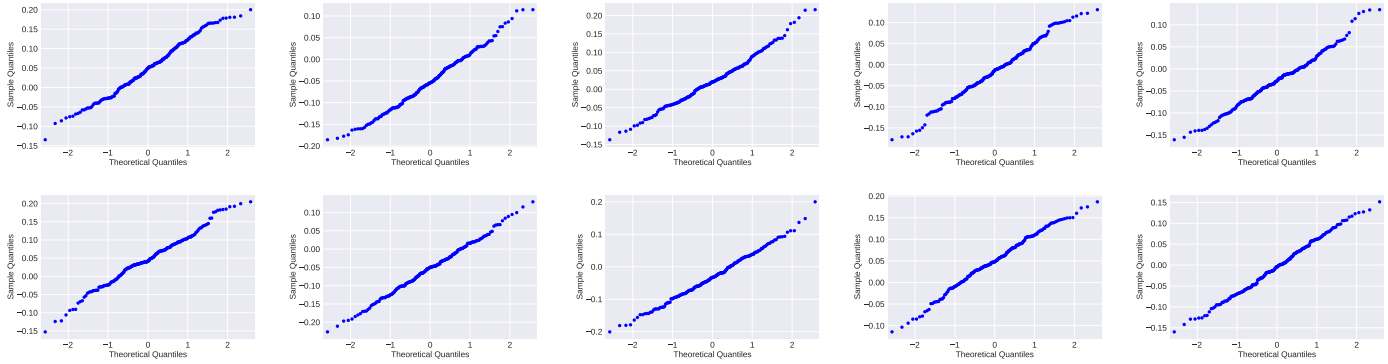


Figure 10: Logistic regression experiment 1: Q-Q plots per coordinate

For 300 i.i.d. samples, we plotted 100 samples from SGD inference in Figure 12. We compare our SGD inference procedure against bootstrap in Figure 12a. Figure 12b and Figure 12c show samples from our SGD inference procedure with different parameters.

10-Dimensional Linear Regression.

Here we consider the following model

$$y = x^\top w^\star + \epsilon,$$

where $w^\star = \frac{1}{\sqrt{10}}[1, \ 1, \ \cdots, \ 1]^\top \in \mathbb{R}^{10}$, $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.8^{|i-j|}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 20^2)$, and use $n = 1000$ samples. We estimate the parameter using

$$\widehat{w} = \operatorname*{argmin}_w \frac{1}{n} \sum_{i=1}^{n} \tfrac{1}{2}(x_i^\top w - y_i)^2.$$

Figure 13 shows the diagonal terms of of the covariance matrix computed using the sandwich estimator and our SGD inference procedure with different parameters. 100000 samples from our SGD inference procedure are used to reduce the effect of randomness.

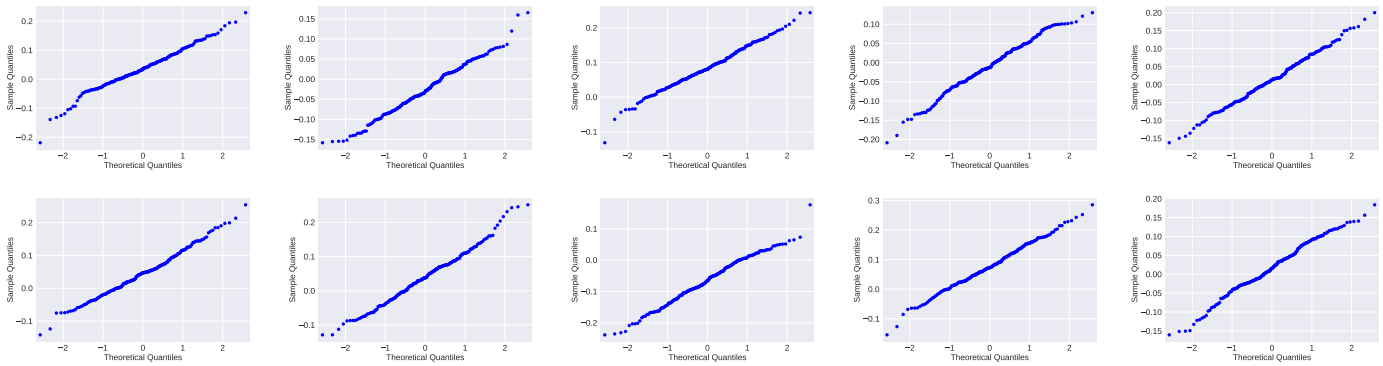


Figure 11: Logistic regression experiment 2: Q-Q plots per coordinate

2-Dimensional Logistic Regression.
Here we consider the following model

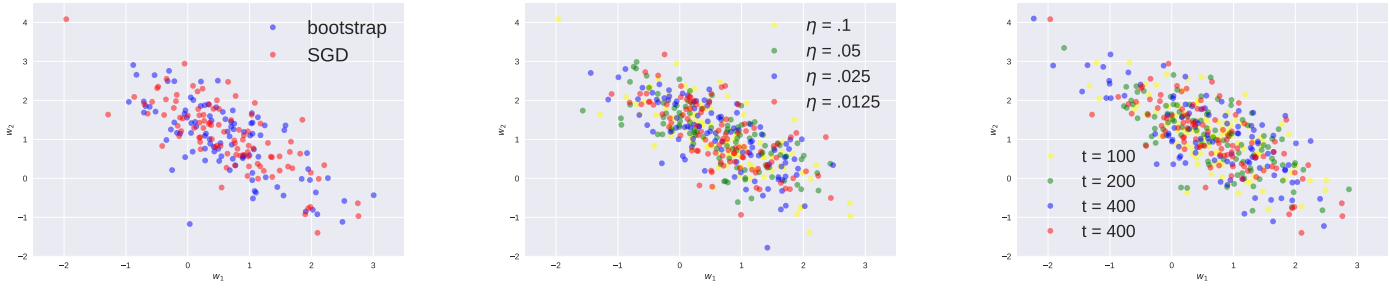
$$\mathbb{P}[Y=+1]=\mathbb{P}[Y=-1]=\frac{1}{2}, \qquad X \mid Y \sim \mathcal{N}\left(\mu=1.1+0.1Y,\ \sigma^2=1\right). \tag{62}$$

We use logistic regression to estimate w, b in the classifier $\operatorname{sign}(wx+b)$ where the minimizer of the population logistic risk is $w^\star = 0.2, b^\star = -0.22$.

For 100 i.i.d. samples, we plot 1000 samples from SGD in Figure 14. In our simulations, we notice that our modified SGD for logistic regression behaves similar to vanilla logistic regression. T his suggests that an assumption weaker than $(\theta-\widehat{\theta})^\top \nabla f(\theta) \geq \alpha\|\theta-\widehat{\theta}\|_2^2$ (assumption (F_1) in Theorem 1) is sufficient for SGD analysis. Figure 14b and Figure 14d suggest that the $t\eta^2$ term in Corollary 2 is an artifact of our analysis, and can be improved.

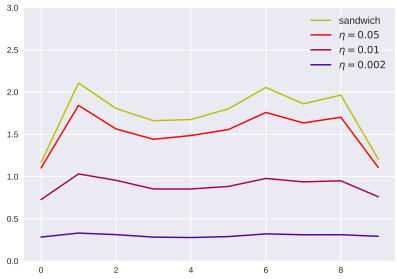
11-Dimensional Logistic Regression.
Here we consider the following model

$$\mathbb{P}[Y=+1]=\mathbb{P}[Y=-1]=\frac{1}{2}, \qquad X \mid Y \sim \mathcal{N}\left(0.01Y\mu, \Sigma\right),$$

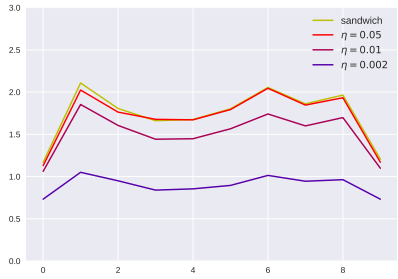


(a) SGD inference vs. bootstrap (b) $t = 800$ (c) $\eta = 0.1$

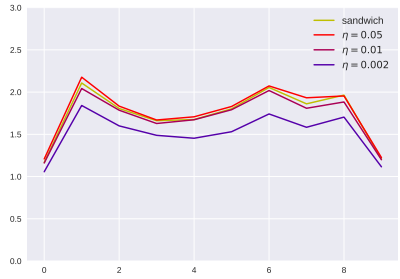
Figure 12: 2-dimensional linear regression



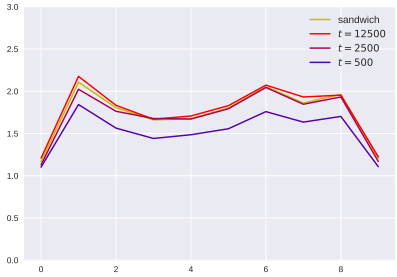
(a) $t = 500$



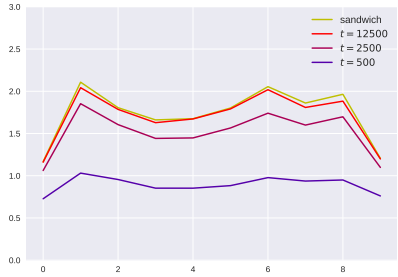
(b) $t = 2500$



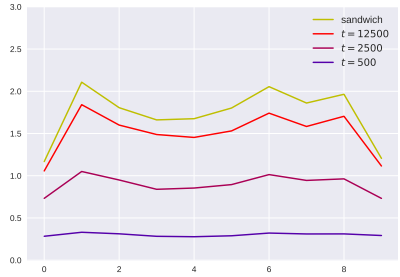
(c) $t = 12500$



(d) $\eta = 0.05$

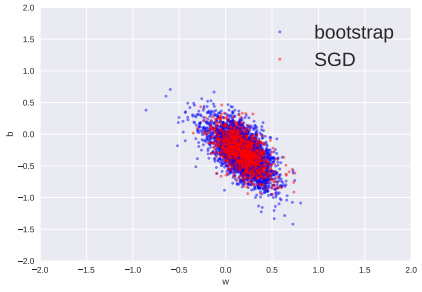


(e) $\eta = 0.01$

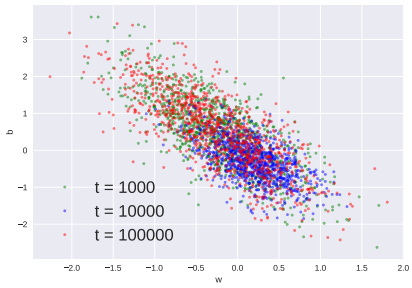


(f) $\eta = 0.002$

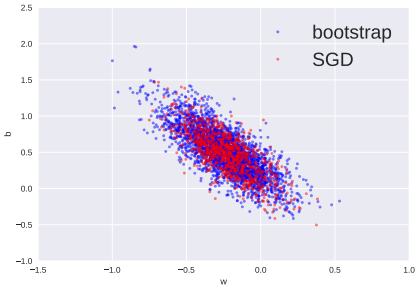
Figure 13: 11-dimensional linear regression: covariance matrix diagonal terms of SGD inference and sandwich estimator



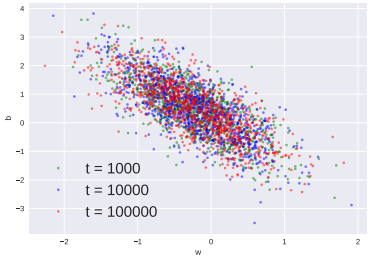
(a) Modified SGD with $t = 1000$ and $\eta = 0.1$



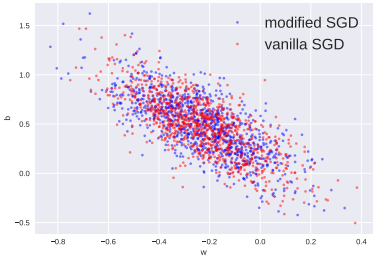
(b) Modified SGD with $\eta = 0.1$



(c) Vanilla SGD with $t = 1000$ and $\eta = 0.1$



(d) Vanilla SGD with $\eta = 0.1$



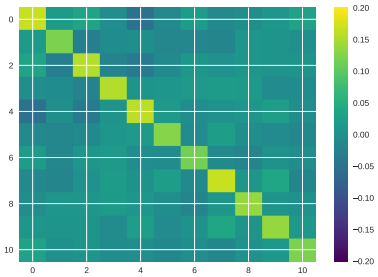
(e) $t = 1000$ and $\eta = 0.1$

Figure 14: 2-dimensional logistic regression

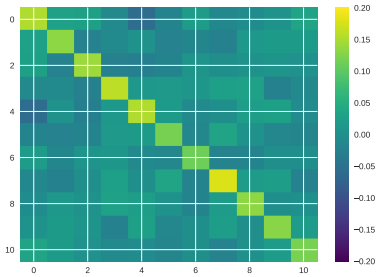
where $\Sigma_{ii} = 1$ and when $i \neq j$ $\Sigma_{ij} = \rho^{|i-j|}$ for some $\rho \in [0, 1)$, and $\mu = \frac{1}{\sqrt{10}}[1, \ 1, \ \cdots, \ 1]^\top \in \mathbb{R}^{10}$. We estimate a classifier $\text{sign}(w^\top x + b)$ using

$$\widehat{w}, \widehat{b} = \underset{w,b}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-Y_i(w^\top X_i + b))\right). \tag{63}$$

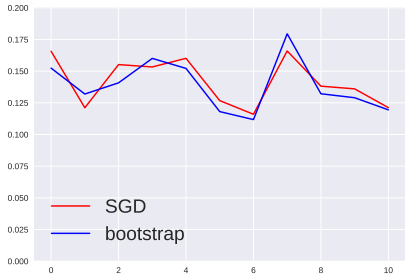
Figure 15 shows results for $\rho = 0$ with $n = 80$ samples. We use $t = 100$, $d = 70$, $\eta = 0.8$, and mini batch of size 4 in vanilla SGD. Bootstrap and our SGD inference procedure each generated 2000 samples. In bootstrap, we used Newton method to perform optimization over each replicate, and 6-7 iterations were used. In figure 16, we follow the same procedure for $\rho = 0.6$ with $n = 80$ samples. Here, we use $t = 200$, $d = 70$, $\eta = 0.85$; the rest of the setting is the same.



(a) SGD covariance

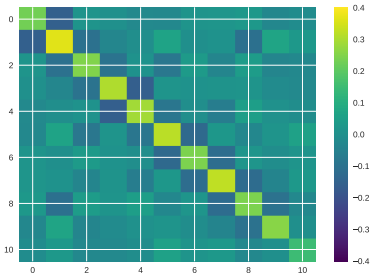


(b) Bootstrap estimated covariance

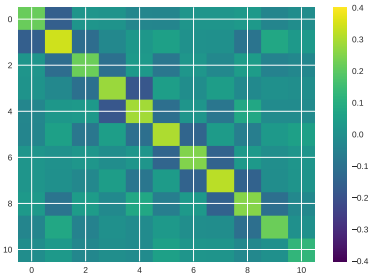


(c) Diagonal terms

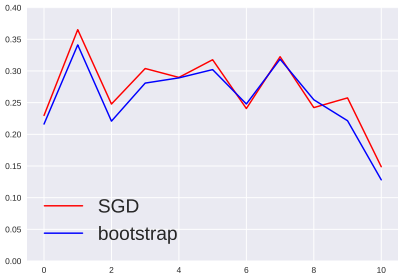
Figure 15: 11-dimensional logistic regression: $\rho = 0$



(a) SGD covariance



(b) Bootstrap estimated covariance



(c) Diagonal terms

Figure 16: 11-dimensional logistic regression: $\rho = 0.6$