The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# Statistical Inference Using SGD

**Tianyang Li**
The University of Texas at Austin
lty@cs.utexas.edu

**Liu Liu**
The University of Texas at Austin
liuliu@utexas.edu

**Anastasios Kyrillidis**
IBM T.J. Watson Research Center,
Yorktown Heights
anastasios.kyrillidis@ibm.com

**Constantine Caramanis**
The University of Texas at Austin
constantine@utexas.edu

**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. To show the merits of our scheme, we apply it to both synthetic and real data sets, and demonstrate that its accuracy is comparable to classical statistical methods, while requiring potentially far less computation.

## 1 Introduction

In $M$-estimation, the minimization of empirical risk functions (RFs) provides point estimates of the model parameters. Statistical inference then seeks to assess the quality of these estimates; *e.g.*, by obtaining confidence intervals or solving hypothesis testing problems. Within this context, 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 (van der Vaart 2000). Thus, given the mean and covariance of this normal distribution, we can infer a range of values, along with probabilities, that allows us to quantify the probability that this interval includes the true minimizer.

The Bootstrap (Efron 1982; Efron and Tibshirani 1994) is a classical tool for obtaining estimates of the mean and covariance of this distribution. The Bootstrap operates by generating samples from this distribution (usually, by re-sampling with or without replacement from the entire data set) and repeating the estimation procedure over these different re-samplings. As parameter dimensionality and data size grow, the Bootstrap becomes increasingly – even prohibitively– expensive.

In this context, we follow a different path: we show that inference can also be accomplished by directly using stochastic gradient descent (SGD), both for point estimates and inference, *with a fixed step size over the data set*. It is well-established that 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, obtained by minimizing RFs, can also be used for statistical inference.* Unlike the Bootstrap, our approach does not require creating many large-size subsamples from the data, neither re-running SGD from scratch for each of these subsamples. 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 the asymptotic normality of SGD, as well as on the 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, 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 the next section, 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 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 (Kushner and Huang 1981; Pflug 1986;



[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. In contrast, *constant step size* means the step size is constant no matter how many iterations taken.

Benveniste, Priouret, and Métivier 1990; Kushner and Yin 2003; Mandt, Hoffman, and Blei 2016), and we postpone the full proof until 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 the experimental section, 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 (Baldi, Sadowski, and Whiteson 2014) and the LIBSVM Splice data sets. 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, we produce 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:

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

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

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

Statistical inference consists of techniques for obtaining information beyond point estimates $\widehat{\theta}$, such as confidence intervals. These can be performed if there is an asymptotic limiting distribution associated with $\widehat{\theta}$ (Wasserman 2013). 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 distribution:

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

where

$$H^\star = \mathbb{E}[\nabla^2 f(\theta^\star; X)],$$

and

$$G^\star = \mathbb{E}[\nabla f(\theta^\star; X) \cdot \nabla f(\theta^\star; X)^\top];$$

see also Theorem 5.21 in (van der Vaart 2000). 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:

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

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^n \nabla f(\widehat{\theta}; X_i)\nabla f(\widehat{\theta}; X_i)^\top. \quad (4)$$

Beyond calculating[2] $\widehat{H}$ and $\widehat{G}$, 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 $\widehat{H}^{-1}\widehat{G}\widehat{H}^{-1}$.

## 3 Statistical inference using SGD

Consider the optimization problem in (2). For instance, in maximum likelihood estimation (MLE), $f(\theta; X_i)$ is a negative log-likelihood function. For simplicity of notation, we use $f_i(\theta)$ and $f(\theta)$ for $f(\theta; X_i)$ and $\frac{1}{n}\sum_{i=1}^n f(\theta; X_i)$, respectively, for the rest of the paper.

The SGD algorithm with a fixed step size $\eta$, is given by the iteration

$$\theta_{t+1} = \theta_t - \eta g_s(\theta_t), \quad (5)$$

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 $g_s(\cdot) \equiv \nabla f_i(\cdot)$, where $i$ is a uniformly random index over the samples $X_i$.

*Our inference procedure uses the average of $t$ consecutive SGD iterations*. In particular, 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. As the final empirical minimum $\widehat{\theta}$, we use in practice $\widehat{\theta} \approx \frac{1}{R}\sum_{i=1}^R \bar{\theta}_t^{(i)}$ (Bubeck 2015).

Some practical aspects of our scheme are discussed below.

*Step size $\eta$ 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) $\eta$ gives equally accurate results with significantly reduced running time. We conjecture that a better understanding of $t$'s and $\eta$'s influence requires stronger bounds for SGD with constant step size. Heuristically, calibration methods for parameter tuning

[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 $\widehat{G}$.

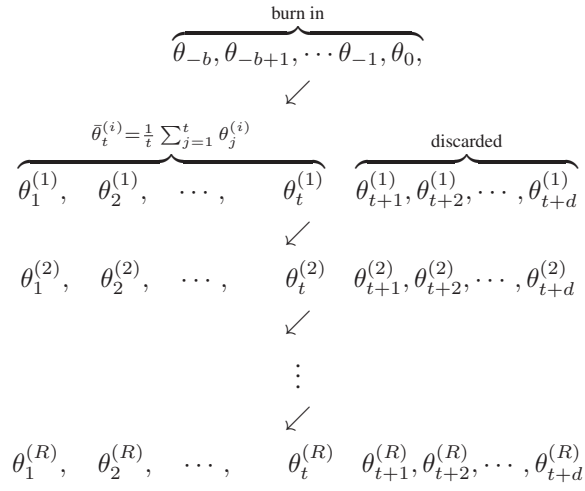



Figure 1: Our SGD inference procedure

in subsampling methods ((Politis, Romano, and Wolf 2012), 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 in the appendix, if we discard $d$ SGD iterates in every segment, the correlation between consecutive $\theta^{(i)}$ and $\theta^{(i+1)}$ is of 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 (Johnson and Zhang 2013)) to find the optimum, and use a relatively small $b$ for SGD to reach stationarity, similar to Markov Chain Monte Carlo burn-in.

*Statistical inference using $\bar{\theta}_t^{(i)}$ and $\widehat{\theta}$:* Similar to ensemble learning (Opitz and Maclin 1999), we use $i = 1, 2, \ldots, R$ estimators for statistical inference:

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

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. Similar to other resampling methods such as bootstrap and subsampling, we use quantiles or variance of $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(R)}$ for statistical inference.

### 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 $H := \nabla^2 f(\widehat{\theta}) = \frac{1}{n} \cdot \sum_{i=1}^{n} \nabla^2 f_i(\widehat{\theta})$. 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:*

- *($G_1$) $\mathbb{E}\left[g_s(\theta) \mid \theta\right] = \nabla f(\theta)$,*
- *($G_2$) $\mathbb{E}\left[\|g_s(\theta)\|_2^2 \mid \theta\right] \leq A\|\theta - \widehat{\theta}\|_2^2 + B$,*
- *($G_3$) $\mathbb{E}\left[\|g_s(\theta)\|_2^4 \mid \theta\right] \leq C\|\theta - \widehat{\theta}\|_2^4 + D$,*
- *($G_4$) $\left\|\mathbb{E}\left[g_s(\theta) g_s(\theta)^\top \mid \theta\right] - G\right\|_2 \leq A_1\|\theta - \widehat{\theta}\|_2 + A_2\|\theta - \widehat{\theta}\|_2^2 + A_3\|\theta - \widehat{\theta}\|_2^3 + A_4\|\theta - \widehat{\theta}\|_2^4$,*

*where $G = \mathbb{E}[g_s(\widehat{\theta}) g_s(\widehat{\theta})^\top \mid \widehat{\theta}]$ and, 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, $\bar{\theta}_t$, satisfies:*

$$\left\| t\mathbb{E}[(\bar{\theta}_t - \widehat{\theta})(\bar{\theta}_t - \widehat{\theta})^\top] - H^{-1} G H^{-1} \right\|_2 \lesssim \sqrt{\eta} + \sqrt{\tfrac{1}{t\eta} + t\eta^2}. \tag{7}$$

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. Further, in the next section, we relate a continuous approximation of SGD to Ornstein-Uhlenbeck process (Robbins and Monro 1951) to give an intuitive explanation of our results.

*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 the appendix.

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.* (11) *below*. When the empirical risk function is not degenerate, (11) satisfies assumptions ($F_1$), ($F_2$), ($F_3$), and ($F_4$). We cannot directly use vanilla SGD to minimize (11), instead we describe a modified SGD procedure for minimizing (11) in Section 3.3, 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 1.

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* $\eta$ 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 (Robbins and Monro 1951), to give an intuitive explanation of our results. In particular, under regularity conditions, the stochastic process $\Delta_t = \theta_t - \widehat{\theta}$ asymptotically converges to an Ornstein-Uhlenbeck process $\Delta(t)$, (Kushner and Huang 1981; Pflug 1986; Benveniste, Priouret, and Métivier 1990; Kushner and Yin 2003; Mandt, Hoffman, and Blei 2016) that satisfies:

$$\mathrm{d}\Delta(T) = -H\Delta(T)\,\mathrm{d}T + \sqrt{\eta}G^{\frac{1}{2}}\,\mathrm{d}B(T), \tag{8}$$

where $B(T)$ is a standard Brownian motion. Given (8), $\sqrt{t}(\bar{\theta}_t - \widehat{\theta})$ can be approximated as

$$\begin{aligned}\sqrt{t}(\bar{\theta}_t - \widehat{\theta}) &= \tfrac{1}{\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta}) \\ &= \tfrac{1}{\eta\sqrt{t}}\sum_{i=1}^{t}(\theta_i - \widehat{\theta})\eta \approx \tfrac{1}{\eta\sqrt{t}}\int_0^{t\eta}\Delta(T)\,\mathrm{d}T,\end{aligned} \tag{9}$$

where we use the approximation that $\eta \approx \mathrm{d}T$. By rearranging terms in (8) and multiplying both sides by $H^{-1}$, we can rewrite the stochastic differential equation (8) 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

$$\begin{aligned}&\int_0^{t\eta}\Delta(T)\,\mathrm{d}T = \\ &- H^{-1}(\Delta(t\eta) - \Delta(0)) + \sqrt{\eta}H^{-1}G^{\frac{1}{2}}B(t\eta).\end{aligned} \tag{10}$$

After plugging (10) into (9) we have

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

When $\Delta(0) = 0$, the variance $\mathrm{Var}\left[-1/\eta\sqrt{t}\cdot H^{-1}(\Delta(t\eta) - \Delta(0))\right] = O\left(1/t\eta\right)$. 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 Logistic regression

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

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

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

$$f(\theta) = \frac{1}{2}\left(c + \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp(-y_i\theta^\top X_i)\right)\right)^2, \quad \text{where } c > 0 \text{ (e.g., } c = 1). \tag{11}$$

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

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

Therefore, we can compute $g_s = \Psi_s\Upsilon_s$, *using two independent random variables satisfying* $\mathbb{E}[\Psi_s \mid \theta] = \Psi$ and $\mathbb{E}[\Upsilon_s \mid \theta] = \Upsilon$. For $\Upsilon_s$, 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 $I_t^\Upsilon$ are $S_\Upsilon$ indices sampled from $[n]$ uniformly at random with replacement. For $\Psi_s$, 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 $I_t^\Psi$ are $S_\Psi$ 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 $\alpha$ by the generalized self-concordance of logistic regression (Bach 2010; 2014), 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 1.** *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* $\Psi_s$*:*

- *with replacement:* $K_G(\theta) = \frac{1}{S_\Psi}(\frac{1}{n}\sum_{i=1}^{n}(c + k_i(\theta))^2) + \frac{S_\Psi - 1}{S_\Psi}(c + k(\theta))^2$,
- *no 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* $\eta$ *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 $\frac{\left(\nabla^2 k(\widehat{\theta})\right)^{-1}}{n}\left(\frac{\sum_{i=1}^n \nabla k_i(\widehat{\theta})\nabla k_i(\widehat{\theta})^\top}{n}\right)\left(\nabla^2 k(\widehat{\theta})\right)^{-1}$, 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 $S_\Psi$.

# 4 Related work

*Bayesian inference:* First and second order iterative optimization algorithms –including SGD, gradient descent, and variants– naturally define a Markov chain. Based on this principle, 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 SGD (Welling and Teh 2011; Bubeck, Eldan, and Lehec 2015; Mandt, Hoffman, and Blei 2016; 2017). We note that, here as well, the vast majority of the results rely on using a decreasing step size. Very recently, (Mandt, Hoffman, and Blei 2017) 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 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). 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.

*Connection with Bootstrap methods:* The classical approach for statistical inference is to use the bootstrap (Efron and Tibshirani 1994; Shao and Tu 2012). Bootstrap samples are generated by replicating the entire data set by resampling, and then solving the optimization problem on each generated set of the data. 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., (Ljung, Pflug, and Walk 2012; Polyak and Juditsky 1992), and more recently, (Toulis and Airoldi 2014; Chen et al. 2016); and also for *fixed step size*, e.g., (Benveniste, Priouret, and Métivier 1990), 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, (Benveniste, Priouret, and Métivier 1990) provides results only when the (constant) step-size approaches 0 (see Section 4.4 and 4.6, and in particular Theorem 7 in (Benveniste, Priouret, and Métivier 1990)). Similarly, while (Kushner and Yin 2003) 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.

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 (Toulis and Airoldi 2014) and (Chen et al. 2016), which follow up on (Polyak and Juditsky 1992). Both of these rely on decreasing step size, for reasons mentioned above. The work in (Chen et al. 2016) uses SGD with decreasing step size for estimating an $M$-estimate's covariance. Work in (Toulis and Airoldi 2014) studies implicit SGD with decreasing step size and proves results similar to (Polyak and Juditsky 1992), however it does not use SGD to compute confidence intervals.

Overall, 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).

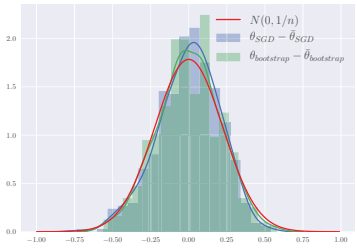
# 5 Experiments

## 5.1 Synthetic data

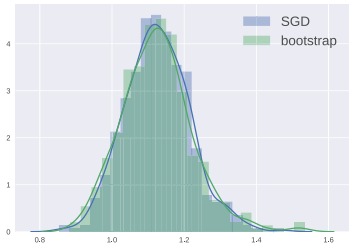
The coverage probability is defined as $\frac{1}{p}\sum_{i=1}^p \mathbb{P}[\theta_i^\star \in \hat{C}_i]$ where $\theta^\star = \operatorname{argmin}_\theta \mathbb{E}[f(\theta, X)] \in \mathbb{R}^p$, and $\hat{C}_i$ is the estimated confidence interval for the $i^{\text{th}}$ coordinate. The average 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 $i^{\text{th}}$ 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. For results given as a pair $(\alpha, \beta)$, it usually indicates (coverage probability, confidence interval length).

**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. Q-Q plots in the appendix show of samples from our SGD inference procedure.

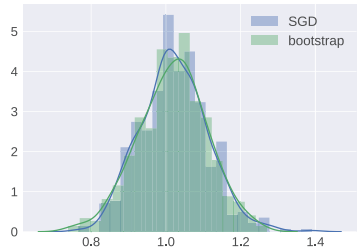
*Normal distribution mean estimation:* Figure 2a compares 500 samples from SGD inference procedure and Bootstrap versus the distribution $\mathcal{N}(0, 1/n)$, using $n = 20$ i.i.d. samples from $\mathcal{N}(0, 1)$. We used mini batch SGD. For the parameters, we used $\eta = 0.8$, $t = 5$, $d = 10$, $b = 20$, 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).




(a) Normal.




(b) Exponential.




(c) Poisson.

Figure 2: Estimation in univariate models.

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 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)

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 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.

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 0.1 | (0.872, 0.204) | (0.937, 0.249) | (0.939, 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)

| $\eta$ | $t = 100$ | $t = 500$ | $t = 2500$ |
|---|---|---|---|
| 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.

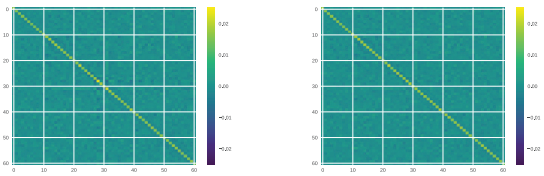
*Exponential distribution parameter estimation:* Figure 2b compares 500 samples from inference procedure and Bootstrap, using $n = 100$ samples from an exponential distribution with PDF $\lambda e^{-\lambda x}$ where $\lambda = 1$. We used SGD for MLE with mini batch sampled with replacement. For the 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).

**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 $1_p = [1 \quad 1 \quad \dots \quad 1]^\top \in \mathbb{R}^p$. Additional simulations comparing covariance matrix computed with different methods are given in the appendix.

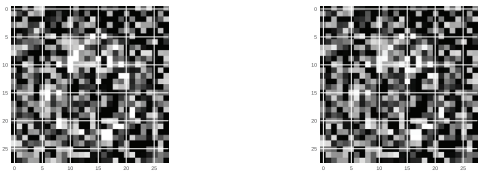
**Linear regression:** *Experiment 1:* Results for the case where $X \sim \mathcal{N}(0, I) \in \mathbb{R}^{10}$, $Y = w^{*T} X + \epsilon$, $w^* = 1_p/\sqrt{p}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 10^2)$ with $n = 100$ samples is given in Table 1a. Bootstrap gives (0.941, 4.14), and confidence 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. 3.3) are given in the appendix. *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 $\mathbb{P}[Y = +1] =$



(a) Bootstrap (b) SGD inference covariance

Figure 3: Splice data set



(a) Original "0": logit -46.3, CI (-64.2, -27.9) (b) Adversarial "0": logit 16.5, CI (-10.9, 30.5)

Figure 4: MNIST

$\mathbb{P}[Y = -1] = 1/2$, $X \mid Y \sim \mathcal{N}(0.01Y1_p/\sqrt{p}, \Sigma) \in \mathbb{R}^{10}$, $\Sigma_{ij} = 0.2^{|i-j|}$ 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 matrices computed using our SGD inference procedure, bootstrap, and inverse Fisher information matrix on the LIBSVM Splice data set, and we observe that they have similar statistical properties.

**Splice data set** The Splice data set [3] contains 60 distinct features with 1000 data samples. This is a classification problem between two classes of splice junctions in a DNA sequence. We use a logistic regression model trained using vanilla SGD.

In Figure 3, 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.

**MNIST** Here, we train a binary logistic regression classifier to classify 0/1 using a noisy MNIST data set, and demonstrate that adversarial examples produced by gradient attack (Goodfellow, Shlens, and Szegedy 2015) (perturbing an image in the direction of loss function's gradient with respect to data) can be detected using prediction intervals. We flatten each $28 \times 28$ image into a 784 dimensional vector, and train a linear classifier using pixel values as features. To add noise to each image, where each original pixel is either 0 or 1, we randomly changed 70% pixels to random numbers uniformly on $[0, 0.9]$. Next we train the classifier on the noisy MNIST data set, and generate adversarial examples using this noisy MNIST data set. Figure 4 shows each image's logit value ($\log \frac{\mathbb{P}[1|\text{image}]}{\mathbb{P}[0|\text{image}]}$) and its 95% confidence interval (CI) computed using quantiles from our SGD inference procedure.

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

### 5.3 Discussion

In our experiments, we observed that using a larger step size $\eta$ produces accurate results with significantly accelerated convergence time. This might imply that the $\eta$ 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 $\eta$ 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 $n \cdot (\text{matrix inversion complexity } = \Omega(p^2)) \cdot (\text{number of Newton iterations } t)$ operations to produce a sample. The experiments therefore suggest that our SGD inference procedure produces results similar to Bootstrap while using far fewer operations.

## 6 Acknowledgments

This work was partially supported by NSF Grants 1609279, 1704778 and 1764037, and also by the USDoT through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. A.K. is supported by the IBM Goldstine fellowship. We thank Xi Chen, Philipp Krähenbühl, Matthijs Snel, and Tom Spangenberg for insightful discussions.

## References

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

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

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

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

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

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

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

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

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

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

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

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

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

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

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

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

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

Pflug, G. 1986. Stochastic minimization with constant stepsize: asymptotic laws. *SIAM Journal on Control and Optimization* 24(4):655–666.

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

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

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

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

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

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

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

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