“Plus/minus the learning rate”: Easy and Scalable Statistical Inference with SGD

Jerry Chee
Cornell University
Department of Computer Science

Hwanwoo Kim
University of Chicago
Department of Statistics

Panos Toulis
University of Chicago
Booth School of Business

Abstract

In this paper, we develop a statistical inference procedure using stochastic gradient descent (SGD)-based confidence intervals. These intervals are of the simplest possible form: $\theta_{N,j} \pm 2\sqrt{\gamma/N}$, where θ_N is the SGD estimate of model parameters θ over N data points, and γ is the learning rate. This construction relies only on a proper selection of the learning rate to ensure the standard SGD conditions for $O(1/n)$ convergence. The procedure performs well in our empirical evaluations, achieving near-nominal coverage intervals scaling up to 20× as many parameters as other SGD-based inference methods. We also demonstrate our method's practical significance on modeling adverse events in emergency general surgery patients using a novel dataset from the Hospital of the University of Pennsylvania. Our code is available on GitHub.

1 Introduction

In recent years, there has been an explosion of interest in large-scale data analysis in the fields of machine learning, statistics, and operations research. (National Research Council, 2013). The method of stochastic gradient descent (SGD) has emerged as the quintessential method in this new domain due to its remarkable simplicity and performance; see Bottou et al. (2018) for a recent review. At the same time, it is increasingly understood that machine learning models should not be employed as just “black boxes” tuned only for prediction. Understanding the model parameters becomes just as important for machine learning systems, which can have far-reaching impact either on specific people or society in general. As a concrete example, in Section 5.2 we present a medical application where the goal is to model adverse events (e.g., surgical complication) upon hospital admission. In this setting, it is not enough to optimize prediction performance. Understanding the model parameters, their signs and magnitudes, gives valuable information about the physiological and sociodemographic factors of the problem, which is necessary to guide effective policies.

However, while optimization methods for machine learning models have made tremendous advances in recent years (Sun et al., 2019), statistical inference methods on model parameters have lagged behind. One key reason for that is that standard techniques for statistical inference do not scale well with large datasets or large models. Recent proposals utilize SGD-based techniques for inference, but are generally challenging to implement—see Section 2 for details. Here, we adopt the framework of likelihood-based inference, where the probability of the data is expressed through a known function of the model parameters. This framework is widely used in statistics (Lehmann and Casella, 2006), and has strong theoretical foundations that we build upon.

To be specific, consider data $(Y, X) \in \mathbb{R}^d \times \mathbb{R}^p$, and a model with negative log-likelihood function ℓ (loss). Let $\theta_\star$ be the model parameters minimizing:

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathrm{E}[\ell(\theta; Y, X)], \tag{1}$$

where $\Theta \subset \mathbb{R}^p$ is a convex Euclidean space. Parameters $\theta_\star$ are unknown and have to be estimated. Typically, the goal of inference, based on i.i.d. data $D_N = \{(Y_i, X_i) : i = 1, \ldots, N\}$, is to construct confidence intervals, $C_{N,j}(D_N)$, such that for every $j = 1, \ldots, p$,

$$\lim \inf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) = 1 - \alpha, \tag{2}$$

for some desired significance level $\alpha \in (0, 1)$. In standard settings, such construction relies on weak convergence results of the form $\sqrt{N}(\widehat{\theta}_N - \theta_\star) \xrightarrow{\mathrm{d}} N(0, F_\star^{-1})$, where $\widehat{\theta}_N$ is the empirical loss minimizer; i.e.,

$$\widehat{\theta}_N = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} \ell(\theta; Y_i, X_i), \tag{3}$$

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

and $F_\star = \mathrm{E}[\nabla\ell(\theta_\star; Y, X)\nabla\ell(\theta_\star; Y, X)^\top]$ is the celebrated Fisher information matrix.

The problem with this construction, however, is that $\widehat{\theta}_N$ cannot be efficiently computed in large data sets. Classical methods, such as Newton-Raphson, the EM algorithm, or quasi-Newton methods scale at a rate of $O(Np^{1+\epsilon})$, at best (Lange, 2010). Moreover, the estimation of $p \times p$ covariance matrix is notoriously hard, especially for large p (Cai et al., 2016; Fan et al., 2016). Standard estimators of covariance matrices, for example, cannot scale particularly well and they are often ill-conditioned or non-invertible (Ledoit and Wolf, 2004).

To address these issues, we could employ the SGD estimator, which is iteratively defined as:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla\ell(\theta_{n-1}; Y_{I_n}, X_{I_n}), \tag{4}$$

where $I_n \sim U\{1, \ldots, N\}$ is a random datapoint, γ_n is the learning rate sequence (typically, $\gamma_n = \gamma_1/n$), and the gradient $\nabla\ell$ is with respect to θ. Classical theory suggests that, under mild conditions, SGD converges to $\widehat{\theta}_N$ as $n \to \infty$ (Benveniste et al., 1990; Borkar, 2008; Robbins and Monro, 1951). If $n = kN$, for $k = 1, 2, \ldots$, then θ_n corresponds to *multipass SGD*—that is, θ_N corresponds to one-pass SGD, θ_{2N} to two-pass SGD, and so on. It follows that θ_{kN} is a viable estimator of $\theta_\star$, for any k. However, to perform inference on $\theta_\star$ through θ_{kN}, we need to understand the statistical properties of θ_{kN}, particularly its $p \times p$ covariance matrix.

1.1 Contributions

In this paper, we propose an inference method based on θ_N, the one-pass SGD estimator. This proposal may be counter-intuitive at first due to the apparent statistical inefficiency of θ_N. The one-pass estimator, however, compensates by two important properties. First, its asymptotic covariance matrix is known in closed form. Second, its covariance matrix can be bounded by a factor that depends only on the learning rate, γ_1. This allows us to construct SGD-based confidence intervals for each component $\theta_{\star,j}$ of the simplest possible form:

$$\theta_{N,j} \pm 2\sqrt{\frac{\gamma_1^*}{N}},\ j = 1, \ldots, p. \tag{5}$$

The key advantage of our proposed construction in Eq. (5) compared to all other SGD-based methods is its simplicity, since it avoids the precise estimation of a large $p \times p$ covariance matrix. Other SGD-based methods also require heavy data-dependent calibration—see Section 2.3 for a detailed discussion. Our method, on the other hand, has a single hyperparameter, γ_1^*, which we can tune with simple data-driven procedures. Importantly, proper selection of γ_1^* relies on conditions that are identical to the standard SGD conditions for $O(1/n)$ convergence. As such, our method does not impose any new constraints. The trade-off is that our confidence intervals are uniform in length, and thus tend to overcover. This turns out not to be a significant issue in the empirical settings we consider. This is because overcoverage happens mainly in ill-conditioned settings, which are problematic for other methods as well, including SGD-based methods and even maximum likelihood (MLE). Moreover, we present extensive empirical evidence to suggest that our confidence intervals remain informative, even in severely ill-conditioned problems. The extent of our empirical evaluations exceeds the state-of-the-art in the existing SGD-based inference literature. We conduct joint inference for up to 4000 parameters (N=$1e5$), 20× more than prior work. We also consider multiple covariate structures, including ill-conditioned cases, whereas most other works mostly used well-conditioned, independent covariate structures.

The rest of the paper is organized as follows. Section 2 serves as motivation, and discusses related work. Section 3 presents our main analysis with results on asymptotic coverage. Section 4 presents our proposed method described in Eq. (5) for constructing confidence intervals. Section 5 presents empirical results on our inference procedure, including simulations and a novel dataset on outpatient healthcare application.

Notation. We use $\xrightarrow{\mathrm{p}}$ and $\xrightarrow{\mathrm{d}}$ to denote convergence in probability and distribution, respectively. For a square matrix A we write $A \succeq 0$ to denote that A is non-negative definite. $||\cdot||$ is understood as the Euclidean norm.

2 Setup and Related Work

2.1 Prediction vs. Inference: A motivating example

Why is statistical inference important for machine learning? Consider a stylized form of the medical application in Section 5.2 on a population of patients. For each patient, we measure comorbidities, X_1, and age, X_2. The outcome model, $Y \sim \mathrm{logistic}\{\theta_\star^\top(1, X_1, X_2)\} \in \{0, 1\}$, captures an adverse event given patient characteristics.

In such settings, it is not enough to optimize prediction performance, e.g., as captured by $\mathrm{E}[(\hat{Y} - Y)^2]$. In particular, the signs of $\theta_\star$ are important indicators of the effects of each factor, while their magnitudes may be important for policy decisions (e.g., triage, treatment allocation). Importantly, the prediction target $\mathrm{E}[(\hat{Y} - Y)^2]$ often behaves very differently from the inferential target, say $\mathrm{E}(||\hat{\theta}_N - \theta_\star||^2)$. To illustrate, suppose $X_1, X_2 \sim N(0, 1)$ marginally, such that $\mathrm{cor}(X_1, X_2) = \rho$. The prediction error, $\mathrm{E}[(\hat{Y} - Y)^2]$, is completely unaffected by ρ. However, regular asymptotics imply that $\mathrm{E}(||\hat{\theta}_N - \theta_\star||^2) = O(||F_\star^{-1}||) = O(1/(1 - \rho^2))$. Thus, inference depends heavily on this nuisance parameter.

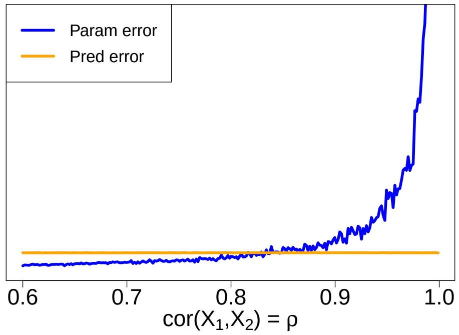


Figure 1: Prediction vs Inference: Multicollinearity degrades parameter estimation error but not prediction error.

See Figure 1 for a numerical simulation with $N = 250$. We see that, as ρ approaches 1, estimating the sign of any component of $\theta_\star$ is essentially a coin flip. This distinction between prediction and inference is not frequently emphasized in the machine learning literature, which tends to focus on the prediction task. A central goal of this paper is to emphasize the inference counterpart, especially in decision-critical applications, and also develop a simple and scalable method for inference based on SGD.

2.2 Background theoretical results

To perform statistical inference using SGD-based estimators, we need weak convergence results similar to the standard results for $\hat{\theta}_N$. Arguably, the most well-known result of this kind comes from the celebrated work of Ruppert (1988); Bather (1989); Polyak and Juditsky (1992) who showed that averaged SGD, namely $\bar{\theta}_N = \frac{1}{N}\sum_{i=1}^{N}\theta_i$, satisfies:

$$\sqrt{N}(\bar{\theta}_N - \theta_\star) \to N_p(0, F_\star^{-1}). \tag{6}$$

Note that $F_\star^{-1}$ is the Cramér-Rao efficiency bound, which is also attained by the empirical risk minimizer, $\hat{\theta}_N$. This efficiency bound cannot be improved by any asymptotically unbiased estimator of $\theta_\star$, which helps explain the popularity of averaged SGD in practice.

On the other hand, standard stochastic approximation results (Ljung et al., 1992, II.8) imply that, under regularity conditions (see Appendix), one-pass SGD satisfies:

$$\sqrt{N}(\theta_N - \theta_\star) \xrightarrow{\text{d}} N_p(0, \Sigma_\star),$$

where

$$\Sigma_\star = \gamma_1^2(2\gamma_1 F_\star - I)^{-1}F_\star. \tag{7}$$

Here, it is assumed that γ_1 is large enough such that $2\gamma_1 F_\star - I \succ 0$. Then, positive-definiteness of $\Sigma_\star$ follows by a diagonalization argument on F_*—see Appendix. A similar result in the context of generalized linear models has been derived by Toulis et al. (2014). It is easy to see that $\Sigma_\star \succeq F_\star^{-1}$, implying that one-pass SGD is not statistically efficient. In fact, the efficiency gap depends on the condition number of $F_\star$, and can be large when the condition number is large.

While both results in Eqs. (6) and (7) can be used for inference on $\theta_\star$ as long as $F_\star$ can be consistently estimated, the aforementioned inefficiency of θ_N has prompted researchers to use averaging for SGD-based inference. In the following section, we give more details on these methods. In Section 3, we present our counter-argument advocating the use of θ_N for inference. The key idea is that $\Sigma_\star$ can easily be bounded above as $\Sigma_\star \preceq \gamma_1^* I$, for a proper selection of the learning rate. This results in a procedure that is significantly simpler and more scalable than alternatives.

2.3 Related work

From the discussion in the previous section, it follows that there are many possible options for inference with SGD-based estimators. Toulis et al. (2017) made a similar point by comparing the distribution of quadratic forms of SGD estimators with their nominal asymptotic χ^2 distribution. Most recent work has exclusively focused on the averaged SGD estimator, $\bar{\theta}_N$, due to its appeal as a statistically optimal estimator. The key idea is to estimate $F_\star^{-1}$, either recursively or through resampling, and then use this estimate for inference (Anastasiou et al., 2019; Chen et al., 2020; Li et al., 2018; Su and Zhu, 2018; Yang et al., 2018).

More specifically, to estimate $F_\star^{-1}$, Chen et al. (2020) follow a clever sub-sampling approach where data are split in batches that are appropriately increased to ensure independence between far apart batch-means. Su and Zhu (2018) split the SGD iterates in a hierarchical tree structure. The total depth of this tree, the number of splits at each junction, and the number of iterations between each split are all parametrized. The covariance matrix is then estimated using the iterates on separate branches of this tree. Yang et al. (2018) propose an online bootstrap resampling scheme. Li et al. (2018) use the empirical covariance of the accumulated averaged estimates. Anastasiou et al. (2019) develop a non-asymptotic multivariate martingale Central Limit Theorem, and use it to prove the rate of convergence of the averaged SGD estimator to a normal random vector. Negrea et al. (2021) show that many stochastic gradient estimators converge weakly to an Ornstein-Uhlenbeck process near local optimum, and with proper tuning the limiting stationary covariance can match the asymptotic distribution of the MLE. Lee et al. (2022b) analyze proximal versions of SGD and estimate the asymptotic covariance matrix in an online fashion. Other works use random scaling methods which can be implemented online as well (Lee et al., 2022a; Li et al., 2022; Chen et al., 2021).

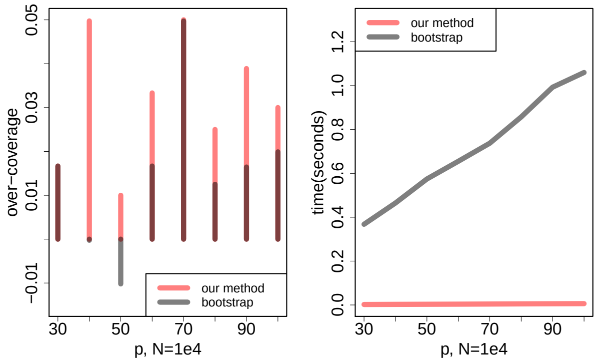
2.4 The need for simple and scalable inference

Most existing SGD-based inference methods thus rely on the averaged SGD estimator (Anastasiou et al., 2019; Chen et al., 2020; Li et al., 2018; Su and Zhu, 2018; Yang et al., 2018). While this estimator comes with elegant theoretical guarantees of optimality (Ruppert, 1988; Polyak and Juditsky, 1992), estimating its $p \times p$ covariance matrix is a daunting task that typically requires heavy manual tuning. This can be an exceptionally challenging task when p is even moderately large, and standard procedures typically suffer by numerical instability and slow convergence rates (Cai et al., 2016; Fan et al., 2016; Ledoit and Wolf, 2004). As such, existing methods of SGD-based inference have been demonstrated on relatively small models and data sizes, in which traditional estimators can scale as well.

As a specific point of comparison, the "batch-means" method of Chen et al. (2020) has a complex set of hyperparameters relating to the convergence of its underlying Markov chain, which includes the number of batches, multiple batch sizes, a decorrelation parameter, and the learning rate. All of these hyperparameters must be manually tuned at costs that cannot be clearly specified. Such difficulties in tuning hyperparameters may cause serious coverage distortion (Lee et al., 2022a). Assuming that such tuning is even possible, Chen et al. (2020) run SGD and construct the full covariance matrix estimate over several batch estimates with complexity $O(N^{\epsilon}p^2 + Np)$.

In contrast, our method employs a single hyperparameter (learning rate), and its selection is fully automatic with small complexity at the order of $O(Np)$. Our method then just needs to run SGD, with the same total complexity of $O(Np)$. Our inference method also empirically scales better than the state-of-the-art; in Section 5 we demonstrate that our method can operate on 20× more model parameters (4000 vs 200) with the same number of samples (N=1e5) compared to other SGD inference methods.

This comparison also extends to more traditional inference procedures, such as the bootstrap (Efron and Tibshirani, 1985; Davison and Hinkley, 1997). Consider, for example, a simplified form of the simulations in Section 5.1, with $Y \sim N(X^\top \theta_\star, 1)$. In Figure 2, we compare our method (formally defined in Algorithm 1) with a procedure that bootstraps the one-pass SGD estimator. In these results, bootstrap appears to be more precise than our method, as shown by the overcoverage rates in Figure 2(a), but quickly becomes impractical as the problem size increases. Compared to the bootstrap, our method overcovers only by a reasonable amount (at most by 5%), and scales without problems. Some recent attempts to scale up bootstrap inadvertently sacrifice its simplicity. The "bag of little bootstraps" (Kleiner et al., 2014), for instance, requires a delicate tuning of its batching and subsampling subroutines, and thus cannot easily scale up to our settings.



(a) Overcoverage Rate

(b) Execution Time

Figure 2: Coverage statistics from confidence intervals generated by bootstrap and our method. Results averaged over 100 replications, with 100 bootstrap samples each.

To address these challenges, our idea in this paper is to trade theoretical optimality with practical simplicity in a controlled manner. We leverage the scalable one-pass SGD estimator, θ_N, and use a simple bound of its asymptotic variance to construct confidence intervals that asymptotically achieve valid coverage. To tune this procedure we only require an appropriate lower bound on the learning rate of SGD. The resulting method is therefore simple and scalable. To compensate for the unavoidable loss in efficiency, we characterize the settings where this loss tends to be worse (Theorem 3.3), and conduct extensive empirical evaluations (Section 5.1) showing that the efficiency loss is relatively small even in heavily ill-conditioned settings. We present the details of our method in the next section.

3 Inference With One-Pass SGD Estimator

The idea to bound $\Sigma_\star$, the covariance matrix of θ_N, can be described as follows. First, note that the eigenvalues of $\Sigma_\star$ in Eq. (7) can be derived in closed form:

$$\mathrm{eigen}(\Sigma_\star) = \left\{ \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} : j = 1, \ldots, p \right\},$$

where λ_j is the j-th eigenvalue of $F_\star$. Note that $\lambda_j > 0$ since $F_\star$ is positive definite. By assumption, the learning rate is large enough such that $2\gamma_1\lambda_j - 1 > 0$ for each component j. Second, note that each eigenvalue of $\Sigma_\star$ asymptotes to $\gamma_1/2$ for large enough γ_1, in the sense that, as γ_1 increases,

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} \Big/ \left(\frac{\gamma_1}{2}\right) \to 1. \tag{8}$$

See Figure 3 for an illustration of the asymptote through the quantity $\lim_{\gamma_1 \to \infty} \mathrm{tr}(\Sigma_\star)/\gamma_1$. Importantly, the limit in Eq. (8) holds for each individual component. It thus implies a uniform bound for $\Sigma_\star$, and, ultimately, a construction of conservative confidence intervals at any desired level.

Theorem 3.1. *Let $\theta_{N,j}$, denote the j-th component of θ_N in Eq.* (4)*, for $j = 1, \ldots, p$. Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$, then $\gamma_1^* I - \Sigma_\star \succ 0$. Define the interval*

$$C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}},\ \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right], \quad (9)$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$ is the critical value of the standard normal. Then, for every $j = 1, \ldots, p$,

$$\liminf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1 - \alpha. \quad (10)$$

The result in Theorem 3.1 shows that we can construct asymptotically valid confidence intervals, *marginally* for every parameter component $\theta_{\star,j}$. Below, we present an additional result for *joint* inference on $\theta_\star$.

Theorem 3.2. *Let θ_N be the one-pass SGD in Eq.* (4)*, and suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Define the following confidence region:*

$$\widehat{\Theta} = \big\{\theta \in \Theta : (1/\gamma_1^*)\ ||\theta - \theta_N||^2 < \chi_{\alpha,p}\big\}, \quad (11)$$

where $\chi_{\alpha,p} = \sup\{x \in \mathbb{R} : P(\chi_p^2 \geq x) \leq \alpha\}$ is the α-critical value of a chi-squared random variable with p degrees of freedom. Then,

$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1 - \alpha. \quad (12)$$

The resulting confidence region for $\theta_\star$ according to Theorem 3.2 is a hypersphere, and can thus be easily computed.

Remark 1 (Choosing γ_1^*). The confidence interval construction in Theorems 3.1 and 3.2 is remarkably simple as it only depends on a user-controlled learning rate (γ_1^*). This avoids estimating $\Sigma_\star$, which is $p\times p$, and only needs a lower bound for its minimum eigenvalue. The resulting bound for γ_1^* is in fact standard for $O(1/n)$ convergence of SGD; e.g., see (Moulines and Bach, 2011, Section 3.1). In Section 4.1, we detail two data-driven procedures for selecting γ_1^* that work well across all our empirical settings.

Remark 2 (Conservativeness). Both our confidence intervals in Eq. (10) and (12) are conservative, and will tend to overcover. This is, unavoidably, the price we have to pay for the computational simplicity of our method. We discuss this issue from a theoretical perspective in the following section. Moreover, in Section 5, we investigate overcoverage through extensive empirical simulations.

Remark 3 (Asymptotics). Our coverage guarantees are only asymptotic. Deriving nonasymptotic guarantees would be challenging. Although non-asymptotic bounds for SGD-based inference have been introduced in Moulines and Bach (2011); Anastasiou et al. (2019), these bounds rely on quantities—e.g., convexity parameters, Lipschitz constants, or condition numbers—that are generally unknown and hard to estimate from available data. We leave this for future work.

Algorithm 1 Scalable inference with one-pass SGD, θ_N.

```
Input: Data $D_N$, SGD procedure of Eq. (4), $\theta_0$, $\alpha \in (0,1)$.
  $\gamma_1^* \leftarrow$ select_gamma$(D_N, \theta_0)$
  $\theta_N \leftarrow$ SGD$(\gamma_1^*, D_N, \theta_0)$
Output: Confidence interval for $\theta_{\star,j}$ from Theorem 3.1:
  $\left(\theta_{N,j} \pm z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/N}\right)$
```

3.1 Quantifying overcoverage

The following result shows that overcoverage in our method depends on the condition number of $F_\star$.

Theorem 3.3. *Let $C_{N,j}(D_N)$ be defined as in Eq.* (9)*, $\lambda_{\min} = \min_j\{\lambda_j\}$, $\lambda_{\max} = \max_j\{\lambda_j\}$, and $\kappa = \lambda_{\max}/\lambda_{\min} \geq 1$. Define $\rho = 1/\gamma_1^*\lambda_{\min}$. Suppose that the learning rate is well specified, such that $\rho \leq 1$. Define*

$$\eta = \max_{j=1,\ldots,p}\left\{\lim\inf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) - (1-\alpha)\right\},$$

the worst-case overcoverage across all $\theta_{\star,j}$. Then,

$$\alpha - 2\Phi(-z_{\frac{\alpha}{2}}\sqrt{2-\rho}) \leq \eta \leq \alpha - 2\Phi(-z_{\frac{\alpha}{2}}\sqrt{2-\rho/\kappa}),$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the critical value of the standard normal distribution.

Theorem 3.3 shows that the amount of overcoverage depends on the condition number of $F_\star$, and the misspecification of the learning rate (ρ). There is no overcoverage when $\rho = 1$ (perfect specification) and $\kappa = 1$. At 5% level, our method may overcover up to 99.4% in a worst-case scenario of large κ (ill-conditioning).

4 Concrete Method and Implementation

Our proposed procedure is summarized in Algorithm 1. The key component of the algorithm is function "select_gamma", which chooses a learning rate so that $\gamma_1^* \geq 1/\lambda_{\min}$, where $\lambda_{\min} = \min_j\{\lambda_j\}$. As discussed before, this is actually the standard SGD condition for $O(1/n)$ convergence (Moulines and Bach, 2011, Section 3.1). So, our method does not introduce any new constraints to practitioners. Sometimes, a good estimate for $\lambda_{\min}$ may be easy to come up with (Cybenko and Van Loan, 1986; Mestre, 2008). In such cases the practitioner may use their own estimate. Here, we contribute two ideas for selecting γ_1^* (Section 4.1). One idea uses an asymptotic result on the eigenvalues of $\Sigma_\star$, and the other idea involves estimating the learning rate condition directly from a crude estimate of $F_\star$. The rest of this section discusses these two approaches, in addition to discussions on simplicity, numerical stability and initialization.

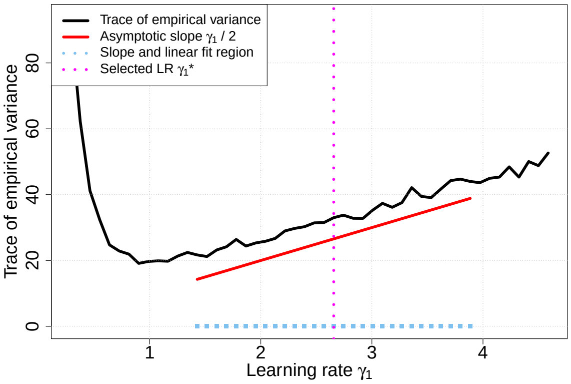


Figure 3: Selection of γ_1^* based on asymptotic results on $\mathrm{eigen}(\Sigma_\star)$. The red line is the line $y = x/2$, which is the asymptote in Eq. (8). The blue line is a "confidence region" for our selection (see Appendix D.1). The vertical purple line marks the heuristically selected γ_1^*.

4.1 Selecting γ_1^*

Linear asymptote in $\Sigma_\star$. At a high level, the variance bound in Theorem 3.1 holds in the regime where the covariance matrix of θ_N is linear with respect to γ_1. One idea is therefore to try and estimate when such regime has been reached. The idea is visualized in Figure 3. Recall from Eq. (8) that the eigenvalues of $\Sigma_\star$ asymptote to $\gamma_1/2$, and so the trace of $\Sigma_\star$ should asymptote to $p\gamma_1/2$, as shown in the figure. The idea is then to slowly increase the learning rate γ_1 and at the same time monitor the trace of $N\mathrm{Var}(\theta_N)$. When γ_1 is large enough for Theorem 3.1 we expect that a linear regression of $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ with respect to γ_1 will give a coefficient around $p/2$ with high confidence. Only a crude estimate of the variance trace is needed, which can be done via bootstrap. See Appendix D.1 for more details, and a practical example.

An eigenvalue bound. In some settings, an estimate $\tilde{F}$ of $F_\star$ exists that may be too crude to be used directly for inference, but may be acceptable for estimating a bound on $\lambda_{\min}$. Then, an alternative way of selecting γ_1^* is to numerically find the maximum eigenvalue of $\tilde{F}^{-1}$, which implies the minimum eigenvalue of $F_\star$ To this end, we propose using inverse power iteration (Trefethen and Bau III, 1997), which is a simple iterative algorithm. More details of this algorithm and its implementation are in Appendix D.2.

4.2 Other implementation details

Numerical stability. The inference procedure in Algorithm 1 depends on selecting a large enough learning rate γ_1^*. However, SGD can be sensitive to the learning rate, and may even diverge if the rate is too large. To resolve such issues, we use SGD with implicit updates (ISGD) (Bertsekas, 2011; Toulis et al., 2014, 2017), revising Eq. (4) as follows:

$$\theta_n = \theta_{n-1} - \frac{\gamma_1}{n}\nabla\ell(\theta_n; Y_{I_n}, X_{I_n}). \tag{13}$$

Note that θ_n appears on both sides of the update, which adds robustness. For instance, in the linear model, ISGD is remarkably stable as it effectively normalizes the learning rate by $||X||^2$; see also (Toulis et al., 2014, Algorithm 1) for efficient computation of Eq. (13) for a large range of models. Furthermore, robustness comes at no cost to efficiency as ISGD has the same asymptotic properties as classical SGD (Toulis et al., 2014, 2016; Bianchi and Hachem, 2016; Asi and Duchi, 2019; Toulis et al., 2021; Lee et al., 2022b). We thus have to use the one-pass ISGD estimator in the numerical experiments that follow. Our inference procedure would not be possible with classical SGD because selecting γ_1^* requires us to explore potentially large learning rates.

Initialization. Choosing θ_0 is important because the one-pass SGD estimator has a limited number of passes to reach the asymptotic regime. Empirically, we have observed that the initialization of θ_0 may have some impact on the quality of the confidence intervals. Several initialization methods were tested: implicit SGD with constant learning rate, decreasing rate $\propto 1/n$ or $1/\sqrt{n}$, and averaged iterates. The best results were achieved by initializing θ_0 with averaged ISGD with a number of $O(\sqrt{p})$ passes over the data.

Other point estimates of $\theta_\star$. It is possible to use other point estimates of $\theta_\star$ to center the confidence intervals in Eq. (9) and (11), e.g., we could use the averaged SGD ($\bar{\theta}_N$), or a multi-pass SGD (θ_{kN}), instead of θ_N. However, the coverage properties of our intervals would not change because we would still have to use our covariance upper bound from Theorem 3.1. That is, despite the swap in the point estimate, the interval length does not change, and so our core asymptotic coverage statement in Theorem 3.1 would remain unchanged as well. While our results are reported with θ_N, we do not observe a meaningful change in the coverage rate when changing the point estimate.

5 Experiments

We have presented a procedure to conduct statistical inference with one-pass SGD. Our explicit design choice has been to trade off theoretical efficiency for a much simpler method. In the following experiments, we demonstrate that we usually strike an effective trade-off. In summary, the amount of overcoverage is mild, and we are able to provide inference results at previously unreported scales. In all of our empirical evaluations, our intervals are always informative and never too wide. Due to space constraints, we report more detailed comparisons under the empirical settings of other works in Appendix E.

5.1 Simulations

Here, we present results from numerical simulations to evaluate the inference procedure in Algorithm 1. Our general approach is to use $M = 500$ confidence intervals $\{\mathrm{CI}_{j,m}\}_{m=1}^{M}$ computed from independently generated data for each run. We evaluate the performance of our inference procedure based on the empirical coverage, defined by $\frac{1}{Mp}\sum_{j=1}^{p}\sum_{m=1}^{M}\mathbf{1}(\theta_{\star,j} \in \mathrm{CI}_{j,m})$, and average interval lengths; see also Chen et al. (2020); Dezeure et al. (2015) for other uses of this metric.[1] A salient subset of our results are given in Tables 1-2 below, with the full set of results in Appendix F.6. We also applied our method on a real-world dataset, based on a preprocessed version of the Adult dataset[2] with 123 binary features and 32,561 samples. These results can be found in Appendix F.2.

Features are sampled as $X_i \sim N_p(0, \Sigma_x)$ under four different structures of the covariance matrix $\Sigma_x \in \mathbb{R}^{p\times p}$:

- Identity (Id): $\Sigma_x = I_p$.
- Toeplitz (T): $\Sigma_x[i,j] = 0.5^{|i-j|}$.
- Equi-Correlation (EC): $\Sigma_x[i,j] = 0.2$ $(i \neq j)$, and $\Sigma_x[i,i] = 1$.
- Ill-Conditioned (IC): $\Sigma_x[i,j] = 0$ $(i \neq j)$, and $\Sigma_x[i,i] = 0.1 + \frac{100-0.1}{p-1}(i-1)$.

As a side note, to the best of our knowledge this is the first paper to consider the challenging ill-conditioned setting.

Configurations of the true parameter $\theta_\star$ include:

- Exponential (Exp): $\theta_{\star,i} = 2(-1)^i e^{-.7i}$.
- Linear (Lin): $\theta_{\star,i} = (i-1)/(p-1)$.

We consider $p \in \{10, 20, 50, 100, 500\}$, and data sizes $N \in \{10^4, 10^5\}$. The model is either a linear normal model or logistic regression; i.e., $Y_i|X_i \sim N(X_i^\top\theta_\star, 1)$ or $Y_i|X_i \sim \mathrm{Bern}(\exp(X_i^\top\theta_\star)/(1+\exp(X_i^\top\theta_\star)))$, respectively. Here, $\mathrm{Bern}(q)$ denotes a Bernoulli random variable with mean q.

Table 1 presents "best-case results" as it assumes we know the true value of $\lambda_{\min}$, and thus set $\gamma_1^* = 1/\lambda_{\min}$. In Table 2, there is no good estimate of $\lambda_{\min}$, and we employ the methods of Section 4.1 to select γ_1^*. We compare to the MLE in Eq. (3), which is statistically optimal under regular conditions. Our implementation of one-pass SGD is 21× faster than the library procedure we use to compute the MLE; e.g., in the R language, this amounts 0.006 seconds (ours) vs 0.128 seconds (mle) for N=1e4, p=100. In Appendix F.4, we also compare to the SGD inference method of (Chen et al., 2020) by referencing results in their paper. In all experiments, the target coverage was set to 95%.

Known $\lambda_{\min}$. In Table 1, where $\lambda_{\min}$ is assumed known, our SGD inference procedure attains an empirical coverage of 96%-98% in the linear regression model, which is close to nominal coverage. The average interval lengths are typically no more than $1.5\times$ the MLE benchmark interval lengths, which indicates that we don't pay a heavy price for the method's computational simplicity. Moreover, higher correlation among X's makes our intervals more conservative. Higher correlation also affects the interval length. For example, with a Toeplitz structure (T), the confidence interval from our procedure almost doubles in length compared to no correlation (Id). Similar patterns emerge from Table 1 for logistic regression as well.

Unknown $\lambda_{\min}$. In, Table 2, $\lambda_{\min}$ is not known, and so we select γ_1^* according the procedures of Section 4.1. We use `SGD-Asym` to refer to the asymptote-based selection procedure, and `SGD-Eig` for the power iteration procedure. We see that `SGD-Asym` attains an empirical coverage of 98%-99%, which is only slightly worse than before. The average interval lengths are typically no more than $2.6\times$ the benchmark MLE interval lengths. Procedure `SGD-Eig` attains the slightly better empirical coverage of 96%-98%, with average interval lengths no more than $1.8\times$ the MLE intervals. These results indicate that improving the estimation of $\lambda_{\min}$ could further improve our method.

Ill-conditioned cases. Our method overcovers more substantially in these cases (∼99.9% in linear model, ∼96-98% in logistic model). Interestingly, the interval lengths increase but not substantially compared to other settings. We emphasize that ill-conditioned settings are also challenging for MLE. Specifically, MLE coverage can even drop to 11% in some ill-conditioned cases, indicating that perhaps inference is just too hard in these particular cases.

Failures. Overall, the coverage results from SGD compare remarkably well with MLE, given that our construction relies only on one single learning rate selection, γ_1^*. One notable failure, however, happens in the logistic model with $p = 500$ and $N = 10^5$, and uncorrelated X's (Id). In Table 1 this setting results in 64% coverage, and in Table 2 the corresponding number is 72%. We believe that the data generating process contributes to this degraded performance, as there is not enough signal to estimate all 500 parameters well enough. This is tentatively confirmed by looking at the empirical distribution of $\exp(X^\top\theta_\star)/(1+\exp(X^\top\theta_\star))$; see Appendix F.3 for details. We obtain additional evidence by looking at the coverage results for MLE, which for these settings are also somewhat problematic (around 88%).

Large-scale settings. The simplicity of our inference procedure allows us to scale to dimensions significantly larger than any other SGD-based inference procedure that

[1] Empirically the coverage rate across coordinates is relatively stable. The component-wise standard deviation is 1.7% for our method, and 2.3% for MLE (linear regression, N=1e4, p=100, Id covariance, Exp $\theta_\star$).

[2] https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

			Linear Regression				Logistic Regression			
			SGD		MLE		SGD		MLE	
$\theta_\star$	Σ_x	p, N	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)
Exp	Id	50, 1e4	96.41	4.33	94.75	3.93	96.71	11.04	95.14	8.91
		500, 1e5	96.95	1.40	95.07	1.24	97.17	3.53	94.92	2.80
	EC	50, 1e4	96.88	4.80	94.86	4.22	96.64	11.56	94.81	9.49
		500, 1e5	96.89	1.52	95.07	1.35	96.91	3.68	94.84	3.01
	T	50, 1e4	98.20	7.08	95.16	5.05	97.54	15.53	95.38	11.00
		500, 1e5	98.17	2.28	94.93	1.60	97.30	4.86	94.91	3.48
	IC	100, 1e5	99.94	4.28	94.82	0.27	96.37	8.52	94.96	0.57
		500, 1e5	99.97	4.89	95.01	0.25	98.78	8.23	94.89	0.50
Lin	Id	50, 1e4	96.76	4.42	94.91	3.93	88.49	34.65	94.48	13.97
		500, 1e5	96.85	1.40	95.01	1.24	63.77	55.77	88.09	7.65
	EC	50, 1e4	96.98	4.79	95.03	4.23	89.56	54.93	94.50	23.19
		500, 1e5	96.87	1.52	95.00	1.35	100.00	222.09	48.40	28.80
	T	50, 1e4	98.17	7.28	95.08	5.05	87.51	42.14	94.58	22.40
		500, 1e5	98.18	2.26	94.99	1.60	85.14	73.68	87.07	12.97
	IC	100, 1e5	99.95	4.30	94.90	0.27	69.91	45.35	86.97	5.74
		500, 1e5	100.00	4.89	94.97	0.25	100.00	249.43	11.00	12.93

Table 1: Linear and Logistic regression and learning rate γ_1^* set to $1/\lambda_{\min}$, where $\lambda_{\min}$ is assumed to be known. The average coverage rate and interval lengths are calculated for a target coverage probability of 95%.

we are aware of. To the best of our knowledge, the largest empirically demonstrated dimension for another SGD-based inference procedure is $p = 200$, $N = 10^5$ in Chen et al. (2020). Here we consider $p \in \{1000, 2000, 4000\}$ and $N = 10^5$ over 100 replications. The results in Table 3 demonstrate the scalability of our method with respect to p. Our method achieves near-nominal coverage ($\sim$98%) even in the largest instances.

Comparison to other SGD-based methods. Despite its simplicity, our method compares favorably to other SGD-based inference methods as well. One notable characteristic is that our method does not suffer from undercoverage, which is generally considered worse than overcoverage from a statistical perspective. Undercoverage seems to be common in methods that require extensive tuning, such as Lee et al. (2022a) and Chen et al. (2020). For instance, the method of Chen et al. (2020), following the implementation of Lee et al. (2022a), may undercover down to 64% even in “easy” instances of logistic regression with N=1e5 and $p < 200$. See Appendix E for further discussion.

5.2 A novel healthcare application

In this section, we use our method to analyze a large medical dataset from the University of Pennsylvania Perelman School of Medicine. In this application, medical researchers want to model adverse events—death, surgical complication, or prolonged hospital stay—among emergency general surgey (EGS) patients. While prediction is one goal, it is also important to understand the driving physiological and sociodemographic factors in order to create effective policies to reduce such adverse events. The scale of this problem warrants the development of large-scale inference methods. EGS medical conditions (Gale and Crystal, 2014; Shafi, 2014) account for more than 800,000 operations in the United States and affect an estimated 3–4 million patients per year (Gale and Crystal, 2014; Havens and Salim, 2015; Ogola and Shafi, 2015; Scott and Havens, 2016).

Our dataset includes patient sociodemographics, insurance types, and comorbidities (Elixhauser and Coffey, 1998). The data analyzed here consists of 22,000 rows and 83 features (1-hot encoded). This scale allows us to compare the confidence intervals generated by our inference procedure with those generated by MLE. We provide both quantitative and qualitative evidence that our inference procedure generates informative confidence intervals. Our workwhorse model was logistic regression.

Quantitatively, 85.7% of the marginal intervals created by our SGD-based procedure share non-zero overlap with MLE. Moreover, 6 out of 8 significant features found by MLE are also found by SGD, while there is only one conflict in coefficient sign. Qualitatively, the positive and negative significant features which SGD and MLE agree on generally make medical sense. For example, both methods agree that performing surgery or high risk conditions increases the probability of an adverse event. Or that a lower number of comorbidities indicates a higher probability of not having an adverse event. We present the full results in Appendix G.

			Linear Regression				Logistic Regression			
			SGD-Asym		SGD-Eig		SGD-Asym		SGD-Eig	
$\theta_\star$	Σ_x	p, N	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)	CovRate (%)	AvgLen ($\times 10^{-2}$)
Exp	Id	50, 1e4	98.89	6.92	96.74	4.38	99.00	17.69	96.68	11.17
		500, 1e5	98.86	2.14	96.81	1.40	99.02	5.84	97.31	3.55
	EC	50, 1e4	98.75	6.73	96.98	4.75	98.84	18.58	96.98	11.61
		500, 1e5	99.00	2.51	96.91	1.52	98.75	5.47	96.99	3.74
	T	50, 1e4	99.18	11.77	97.96	7.09	98.98	23.87	97.30	15.30
		500, 1e5	99.29	3.95	98.26	2.28	99.00	8.16	97.26	4.86
	IC	100, 1e5	100.00	9.65	99.96	4.28	96.57	6.83	96.36	8.45
		500, 1e5	100.00	10.61	100.00	4.95	99.05	16.49	98.77	8.13
Lin	Id	50, 1e4	98.99	6.89	96.76	4.39	94.84	51.51	86.30	31.44
		500, 1e5	98.98	2.31	96.91	1.40	71.83	110.39	64.10	54.57
	EC	50, 1e4	98.94	7.87	96.44	4.74	93.09	79.9	89.64	49.49
		500, 1e5	99.02	2.56	96.84	1.52	100.00	414.28	100.00	212.58
	T	50, 1e4	99.14	10.29	97.74	7.08	90.48	51.52	87.34	40.68
		500, 1e5	99.37	4.20	98.24	2.28	94.73	146.82	85.26	74.51
	IC	100, 1e5	100.00	9.52	99.90	4.28	93.43	94.07	69.46	47.79
		500, 1e5	100.00	10.72	100.00	4.95	100.00	567.26	100.00	236.25

Table 2: Linear and Logistic regression and SGD with the first γ_1^* selection method (SGD-Asym) and second method (SGD-Eig). The average coverage rate and interval lengths are calculated for a target coverage probability of 95%.

$\theta_\star$	Σ_x	p, N	CovRate (%)	AvgLen ($\times 10^{-2}$)
Exp	Id	1e3, 1e5	97.18	1.47
		2e3, 1e5	97.82	1.59
		4e3, 1e5	98.36	1.78

Table 3: High dimensional linear regression with inverse power method to select γ_1^*. The average coverage rate and interval lengths are calculated for a target coverage probability of 95%.

6 Concluding Remarks

We developed a statistical inference procedure using SGD-based confidence intervals. These intervals are of the simplest possible form as they depend only on a cautious selection of the learning rate. Through numerical experiments of up to $p = 4000$, and $N = 10^5$ we demonstrated that our intervals have good coverage properties even compared to MLE. We also showed our method's practical significance on modeling adverse events in EGS. This example illustrated the practical need for scalable inference in modern machine learning as prediction performance alone does not capture how well models describe the real world. The key idea in our method is to trade off statistical efficiency for computational simplicity, and is guided by our belief that simplicity trumps theoretical optimality in real-world applications. The resulting inference procedure is remarkably simple, and achieves the desired coverage, albeit conservatively, in a wide range of large-scale empirical evaluations.

Acknowledgements

This work was partially funded by the National Science Foundation under award DGE-1922551.

References

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

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. *arXiv preprint arXiv:1904.02130*, 2019.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

JA Bather. *Stochastic approximation: A generalisation of the Robbins-Monro procedure*, volume 89. Mathematical Sciences Institute, Cornell University, 1989.

Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive algorithms and stochastic approximations*. Springer-Verlag New York, Inc., 1990.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

Pascal Bianchi and Walid Hachem. Dynamical behavior of a stochastic forward–backward algorithm using random monotone operators. *Journal of Optimization Theory and Applications*, 171(1):90–120, 2016.

Vivek S Borkar. Stochastic approximation. *Cambridge Books*, 2008.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Reviews*, 60(2):223–311, 2018. URL `http://leon.bottou.org/papers/bottou-curtis-nocedal-2018`.

T Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

Xi Chen, Zehua Lai, He Li, and Yichen Zhang. Online statistical inference for stochastic optimization via kiefer-wolfowitz methods. *arXiv preprint arXiv:2102.03389*, 2021.

George Cybenko and Charles Van Loan. Computing the minimum eigenvalue of a symmetric positive definite toeplitz matrix. *SIAM journal on scientific and statistical computing*, 7(1):123–131, 1986.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

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

Bradley Efron and Robert Tibshirani. The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17):1–35, 1985.

Steiner C. Harris D. R. Elixhauser, A. and R. M. Coffey. Comorbidity measures for use with administrative data. *Medical Care*, 36:8–27, 1998.

Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

Jianqing Fan, Yuan Liao, and Han Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

Shafi S. Dombrovskiy V. Y. Arumugam D. Gale, S. C. and J. S. Crystal. The public health burden of emergency general surgery in the united states: a 10-year analysis of the nationwide inpatient sample—2001 to 2010. *Journal of Trauma and Acute Care Surgery*, 77:202–208, 2014.

Peetz A. B. Do W. S. Cooper Z. Kelly E. Askari R. Reznor G. Havens, J. M. and A. Salim. The excess morbidity and mortality of emergency general surgery. *Journal of Trauma and Acute Care Surgery*, 78:306–311, 2015.

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

Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

Kenneth Lange. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Association for the Advancement of Artifical Intelligence*, 2022a.

Yoonhyung Lee, Sungdong Lee, and Joong-Ho Won. Statistical inference with implicit sgd: proximal robbins-monro vs. polyak-ruppert. In *39th International Conference on Machine Learning*, 2022b.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using sgd. In *32nd AAAI Conference on Artificial Intelligence*, 2018.

Xiang Li, Jiandong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via local sgd. In *35th Annual Conference on Learning Theory*, 2022.

Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic approximation and optimization of random systems*, volume 17. Springer, 1992.

Xavier Mestre. Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11):5113–5129, 2008.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

National Research Council. *Frontiers in massive data analysis*. National Academies Press, 2013.

Jeffrey Negrea, Jun Yang, Haoyue Feng, Daniel M Roy, and Jonathan H Huggins. Statistical inference with stochastic gradient algorithms, 2021. URL `http://utstat.toronto.edu/~negrea/media/scaling-limit-sgld.pdf`.

Gale S. C. Haider A. Ogola, G. O. and S. Shafi. The financial burden of emergency general surgery: national estimates 2010 to 2060. *Journal of Trauma and Acute Care Surgery*, 79:444–448, 2015.

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

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

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

David J Sakrison. Stochastic approximation: A recursive method for solving regression problems. In *Advances in communication systems*, volume 2, pages 51–106. Elsevier, 1966.

Olufajo O. A. Brat G. A. Rose J. A. Zogg C. K. Haider A. H. Salim A. Scott, J. W. and J. M. Havens. Use of national burden to define operative emergency general surgery. *JAMA surgery*, 151:e160480–e160480, 2016.

Aboutanos M. B. Agarwal Jr S. Brown C. V. Crandall M. Feliciano D. V. Guillamondegui O. Haider A. Inaba K. Osler T. M. Shafi, S. Emergency general surgery: definition and estimated burden of disease. *Journal of Trauma and Acute Care Surgery*, 74:1092–1097, 2014.

S. S. Shapiro and Wilk. M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611, 1965.

Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, 50(8): 3668–3681, 2019.

Panos Toulis, Jason Rennie, and Edoardo Airoldi. Statistical analysis of stochastic gradient methods for generalized linear models. In *31st International Conference on Machine Learning*, 2014.

Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298. PMLR, 2016.

Panos Toulis, Edoardo M Airoldi, et al. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Panos Toulis, Thibaut Horel, Edoardo M Airoldi, et al. The proximal robbins–monro method. *Journal of the Royal Statistical Society Series B*, 83(1):188–212, 2021.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Yusaku Yamamoto. On the optimality and sharpness of laguerre's lower bound on the smallest eigenvalue of a symmetric positive definite matrix. *Applications of Mathematics*, 62(4):319–333, 2017.

Yixin Yang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient estimator. *The Journal of Machine Learning Research*, 19:1–21, 2018.

Appendix

A A general result on the sampling properties of SGD-based estimators.

Here, we present a general result on the asymptotic distribution of SGD-based estimators. This result is general enough to apply to many variants, including mini-batch and variance-reduced SGD, which we discuss next. The result is stated on the following general procedure:

$$\theta_n = \theta_{n-1} - \frac{\gamma_1}{n} S_n(\theta_{n-1}). \tag{14}$$

Here, $S_n(\theta)$ is a random function, and it is understood that there exists a non-decreasing sequence of σ-fields , $\mathcal{F}_n$, such that S_n is measurable by $\mathcal{F}_n$ for every n. In practice, $S_n(\theta)$ will correspond to some form of stochastic gradient, so that it is zero in expectation at $\theta_\star$. For the asymptotic distribution of θ_n in Eq. (14) we will use the following assumptions.

Assumption A.1. There exists $s : \mathbb{R}^p \to \mathbb{R}^p$ such that $s(\theta) = \mathrm{E}[S_n(\theta)]$, for all n. The Jacobian, $J_s(\theta)$, of $s()$ exists and is bounded for every θ. Also, define:

$$\sigma^2_{n,\epsilon} = \mathrm{E}\big(\mathbb{I}\{||S_n(\theta) - s(\theta)||^2 > \epsilon n\}||S_n(\theta) - s(\theta)||^2\big).$$

The following conditions hold:

(i) There is a unique $\theta_\star \in \Theta$ for which $\mathrm{E}[S_n(\theta_\star)] = 0$;

(ii) $J_s(\theta)$ is Lipschitz and $s(\theta)$ is smooth element-wise (with bounded high-order derivatives);

(iii) $\gamma_1 J_s^\star - I/2 \succ 0$, where $J_s^\star = J_s(\theta_\star)$;

(iv) There exists $V_s^\star \succ 0$ such that
$$||\mathrm{E}\big(S_n(\theta_\star)S_n(\theta_\star)^\top\big) - V_s^\star|| \to 0;$$

(v) $\sum_{j=1}^n \sigma^2_{j,\epsilon} = o(n)$, for every $\epsilon > 0$.

Remarks. Assumption 1(i) is necessary for stochastic convergence of the iterations in Eq. (14). As mentioned earlier, this holds when $S_n(\theta)$ is defined through the stochastic gradient (see following section). Assumptions 1(ii) and 1(iii) are necessary for $O(1/n)$ convergence. Assumption 1(v) is the classical Feller-Lindeberg condition for asymptotic normality.

Theorem A.2. *For the procedure in Eq.* (14)*, suppose that $\theta_n \xrightarrow{\mathrm{p}} \theta_\star$, and all conditions in Assumption A.1 hold. Then,*

$$\sqrt{n}(\theta_n - \theta_\star) \xrightarrow{\mathrm{d}} N_p(0, \Sigma),$$

where $\gamma_1^2 V_s^\star = (\gamma_1 J_s^\star - I/2)\Sigma + \Sigma(\gamma_1 J_s^\star - I/2)$.

Proof. Define $w_n = \theta_n - \theta_\star$ and rewrite the SGD iteration as follows:

$$w_n = w_{n-1} - \frac{\gamma_1}{n}[s(\theta_{n-1}) + \epsilon_n],$$

where $\epsilon_n = S_n(\theta_{n-1}) - s(\theta_{n-1})$. Note that the definition of ϵ is possible by Assumption A.1. We make a Taylor approximation on $s(\theta)$ around $\theta_\star$: $s(\theta) = s(\theta_\star) + J_s(\theta_\star)(\theta - \theta_\star) + O(||\theta - \theta_\star||^2)$. We write the above iteration as:

$$w_n = [I - \frac{\gamma_1}{n}J_s(\theta_\star)]w_{n-1} - \frac{\gamma_1}{n}[\epsilon_n + s(\theta_\star) + O(||\theta_{n-1} - \theta_\star||^2)].$$

Then,

- $O(||\theta_{n-1} - \theta_\star||^2) \xrightarrow{\mathrm{p}} 0$ since we assume consistency of θ_n, i.e., $\theta_n \xrightarrow{\mathrm{p}} \theta_\star$.
- $s(\theta_\star) = 0$ by Assumption 1(i);
- $2\gamma_1 J_s(\theta_\star) - I$ is positive definite by Assumption 1(iii);
- $||\mathrm{E}(\epsilon_n\epsilon_n^\top - V_s^\star)|| \xrightarrow{\mathrm{p}} 0$ by the continuous mapping theorem and Assumption 1(iv);

- The Lindenberg-Feller conditions hold for ϵ_n by definition of ϵ and Assumption 1(v);

The result then follows from Theorem 1 of Fabian (1968). □

Remarks. Theorem A.2 requires that θ_n is consistent for $\theta_\star$. The conditions for such form of convergence are, of course, weaker than the conditions in Theorem A.2, and have been extensively studied. One classical set of assumptions, for example, is strong convexity of $s(\theta)$ and uniformly bounded second moments, i.e., $\mathrm{E}(||S_n(\theta)||^2) = O(1)$ w.p. 1 for every θ; see Moulines and Bach (2011) for more details. Almost sure consistency is also possible when $S_n(\theta)$ is uniformly bounded; see, for example, Borkar (2008, Section 1).

We illustrate Theorem A.2 for the statistical model of Eq. (1), where a dataset, $D_N = \{(x_j, y_j) : j = 1, \ldots, N\}$, is given. In this context, Theorem A.2 may be used under either a "fixed data" regime where N is finite, or a "streaming data" regime where $N = \infty$. In the former regime, $\theta_\star$ as defined in Assumption A.1 corresponds to $\widehat{\theta}_N$ in Eq. (3), since all probability statements in Assumption A.1 and Theorem A.2 are with respect to the empirical data distribution. In the latter regime, $\theta_\star$ as defined in Assumption A.1 corresponds to the population parameter $\theta_\star$ as defined in Eq. (1). The streaming regime is suitable for the statistical inference task of Eq. (2), which is the main goal of this paper, but below we will briefly discuss both regimes.

We now illustrate how Theorem A.2 covers a variety of SGD procedures.

Example 1: SGD, fixed data ($N < \infty$). Let $G_N(\theta)$ be the $p \times N$ matrix where the j-th column is $\nabla \ell(\theta; y_j, x_j)$ and Z_N denotes a binary (column) vector of length N with only one nonzero element chosen at random. With these definitions, the SGD procedure of Eq. (4) corresponds to procedure (14) with $S_n(\theta) = G_N(\theta) Z_N$. Since $\mathrm{E}(Z_N) = (1/N)\mathbf{1}_N$, where $\mathbf{1}_N$ is the vector of N ones, it follows that $\mathrm{E}(G_N(\theta)Z_N) = (1/N)\sum_{j=1}^N \nabla \ell(\theta; y_j, x_j)$ is the full-data gradient. As mentioned earlier, the minimizer $\widehat{\theta}_N$ defined in Eq. (3) plays the role of $\theta_\star$ (as defined in Assumption A.1 and Theorem A.2), and all probability statements are with respect to the empirical data distribution. Specifically, Theorem A.2 shows that multi-pass SGD, θ_{kN}, is asymptotically normal around $\widehat{\theta}_N$ as $k \to \infty$.

Example 2: Mini-batch SGD, fixed data $(N < \infty)$**.** Instead of sampling one datapoint per iteration, m data points can be sampled and their gradients averaged. In this case, $S_n(\theta) = (1/m) G_N(\theta) Z_{N,m}$ where $Z_{N,m}$ is a N-length binary vector with only m nonzero elements chosen at random without replacement. Averaging the gradients does not change their expectation, and so $\mathrm{E}(S_n(\theta))$ is the full gradient as in Example 1. The asymptotic variance of mini-batch SGD is $1/m$ as that of plain SGD, but this comes at a cost of processing m times more samples. It follows that mini-batch SGD is statistically equivalent to plain SGD.

Example 3: Variance reduction, fixed data ($N < \infty$). For variance reduction methods (Johnson and Zhang, 2013, SVRG) the procedure is typically defined as in Eq. (14) with

$$S_n(\theta) = G_N(\theta) Z_N - G_N(\widetilde{\theta}) Z_N + (1/N) G_N(\widetilde{\theta}) \mathbf{1}_N,$$

where $\widetilde{\theta}$ is an additional iterate obtained by periodically running SGD over all the data. By construction, $\mathrm{E}(S_n(\theta))$ is again the same as in Example 1. However, the limit variance, $V_s^\star$ of $S_n(\theta)$ is vanishing at rate $O(1/n)$ if both $\theta_n \xrightarrow{\mathrm{p}} \theta_\star$ and $\widetilde{\theta}_n \xrightarrow{\mathrm{p}} \theta_\star$ (as before, $\theta_\star \equiv \widehat{\theta}_N$ in this setting). This means that θ_n from the SVRG procedure is a *super-efficient* estimator of $\widehat{\theta}_N$.

Example 4: One-pass SGD, streaming data ($N = \infty$). In this setting, we define $S_n(\theta) = \nabla \ell(\theta; Y_n, X_n)$, where (Y_n, X_n) is the n-th datapoint sampled independently from the population. As mentioned earlier, $\theta_\star$ in Theorem A.2 now corresponds to the target population parameter in Eq. (1).

An important and useful simplification happens if we assume that the loss in Eq. (1) and Eq. (3) is *well-specified.* We introduce the following assumption.

Assumption A.3. For the statistical model of Eq. (1) let $Y|X \sim f(Y|X, \theta_\star)$, where f is a known density family parameterized in Θ, but the model parameter $\theta_\star$ is unknown. The loss function is well-specified when it corresponds to the negative log-likelihood, i.e., $\ell(\theta; y, x) = -\log f(y|x, \theta)$.

Under Assumption A.3 and typical regularity conditions on the model f and space Θ (Van der Vaart, 2000, Section 5) it holds that $\mathrm{E}\{\nabla \ell(\theta_\star; Y, X)\} = 0$, and the expected Hessian of the loss and its gradient variance are related:

$$\mathrm{E}\{\nabla^2 \ell(\theta_\star; Y, X)\} = \mathrm{Var}\{\nabla \ell(\theta_\star; Y, X)\} \triangleq F_\star \succ 0. \tag{15}$$

Matrix $F_\star$ is known as the Fisher information matrix and plays a crucial role in statistical estimation (Casella and Berger, 2002, Section 7). The following result gives a more refined asymptotic variance under a well-specified loss.

Corollary A.4. *In the streaming data regime, consider procedure Eq.* (14) *with* $S_n(\theta) = \nabla\ell(\theta; Y_n, X_n)$ *where* (Y_n, X_n) *is the* n*-th datapoint sampled i.i.d.. Suppose also that Assumption A.1, and the loss is well-specified as described in Assumption A.3. Then, the one-pass SGD estimator satisfies*

$$\sqrt{N}(\theta_N - \theta_\star) \xrightarrow{\text{d}} N_p(0, \Sigma_\star),$$

where $\Sigma_\star = \gamma_1^2(2\gamma_1 F_\star - I)^{-1} F_\star$.

Proof. In this case, by definition, $J_s^\star = V_s^\star = F_\star$, where $F_\star$ is the Fisher information matrix. The variance formula of Theorem A.2 can be written as:

$$\gamma_1^2 F_\star = (\gamma_1 F_\star - I/2)\Sigma + \Sigma(\gamma_1 F_\star - I/2). \tag{16}$$

This equation has a unique solution. To see this, consider two solutions, Σ_1 and Σ_2. Let $\Delta = \Sigma_1 - \Sigma_2$ and $\gamma_1 F_\star = F \succ 0$ for simplicity. Then, by simple subtraction:

$$\begin{aligned}(F - I/2)\Delta + \Delta(F - I/2) = 0\\ F\Delta + \Delta F = \Delta.\end{aligned}$$

Let λ be the minimum eigenvalue of F and v the corresponding eigenvector, such that $Fv = \lambda v$, and $\lambda > 0$. Note that $2\lambda - 1 > 0$ by Assumption 1. Multiply with Δv from the left and with v from the right to obtain:

$$\begin{aligned}(\Delta v)^\top F\Delta v + (\Delta v)^\top \Delta(Fv) = (\Delta v)^\top \Delta v\\ (\Delta v)^\top F\Delta v + \lambda||\Delta v||^2 = ||\Delta v||^2\\ \lambda||\Delta v||^2 + \lambda||\Delta v||^2 \leq ||\Delta v||^2\\ (2\lambda - 1)||\Delta v||^2 \leq 0.\\ ||\Delta v||^2 \leq 0.\end{aligned}$$

Hence, $\Delta = 0$ and so the solution must be unique if it exists. Indeed, a solution exists since we can verify that $\Sigma = \gamma_1^2(2\gamma_1 F_\star - I)^{-1} F_\star$ satisfies Eq. (16). □

Example 5: Adaptive methods, streaming data ($N = \infty$**).** In the streaming regime, define $S_n(\theta) = \Gamma\nabla\ell(\theta; Y_n, X_n)$ where Γ is a $p \times p$ positive definite matrix, and $\gamma_1 = 1$ without loss of generality. Under well-specified loss, we can use Eq. (15) to obtain $J_s^\star = \Gamma \mathrm{E}(\nabla^2\ell(\theta; Y, X))\Gamma^\top = \Gamma F_\star \Gamma^\top$. Also, $\mathrm{Var}(S_n(\theta_\star)) = \Gamma F_\star \Gamma^\top$. If we choose $\Gamma = F_\star^{-1}$, Corollary A.4 implies that the one-pass SGD estimator satisfies:

$$N\mathrm{Var}(\theta_N) \to F_\star^{-1}.$$

Matrix $F_\star^{-1}$ is the celebrated Cramér-Rao bound in statistics, and is the fundamental estimation bound under regularity conditions; that is, no other consistent estimator of $\theta_\star$ can achieve smaller variance than $F_\star^{-1}/N$ in the limit (Casella and Berger, 2002, Section 10.1).

Most adaptive methods therefore try to approximate $F_\star$, one way or another. For example, Sakrison (1966) shows that the scheme $S_n(\theta) = F_n^{-1}\nabla\ell(\theta; Y_n, X_n)$ is also efficient when $F_n \xrightarrow{\text{p}} F_\star$. Amari (1998) refers to such $S_n(\theta)$ as the "natural gradient", and makes a similar argument. Other adaptive procedures such as AdaGrad Duchi et al. (2011), Adam Kingma and Ba (2014), or RMSProp, are variations of this idea.

B Details on Figure 2

We set $\theta_\star$ to the "Linear" setting as described in the simulation results. We assume that we have a good estimate of $\lambda_{\min}$, the minimum eigenvalue of the Fisher Information. Thus, the timing results are comparing the multiple SGD runs required for the bootstrap estimate with our method which only needs to run SGD once to compute the confidence interval.

C Results for Section 3

Theorem 3.1. *Let $\theta_{N,j}, j = 1, \ldots, p$ denote the j-th component of θ_N in Eq.* (4). *Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$ so that $\gamma_1^* I - \Sigma_\star \succ 0$. Define the interval*

$$C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}},\ \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right],$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$ is the critical value of the standard normal. Then, for any $j = 1, \ldots, p$,

$$\liminf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1-\alpha.$$

Proof. Let $\Sigma_\star = Q\Lambda Q^\top$ be the eigendecomposition of $\Sigma_\star$. The j-th diagonal element of Λ satisfies $\Lambda[j,j] = (\gamma_1^*)^2\lambda_j/(2\gamma_1^*\lambda_j - 1)$, with λ_j the j-th eigenvalue of $F_\star$. The eigenvalues of $\Sigma_\star$ can be decomposed as

$$\frac{(\gamma_1^*)^2\lambda_j}{2\gamma_1^*\lambda_j - 1} = \frac{\gamma_1^*}{2} + \frac{1}{4\lambda_j} + \frac{1}{4\lambda_j(2\gamma_1^*\lambda_j - 1)}.$$

Since $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$, it follows that for all $j = 1, \ldots, p$,

$$\gamma_1^* - \frac{(\gamma_1^*)^2\lambda_j}{2\gamma_1^*\lambda_j - 1} > 0 \Rightarrow \gamma_1^* I - \Sigma_\star \succ 0. \tag{17}$$

This establishes the first claim in the theorem. Now, we prove the second claim. The technique is straightforward, and follows the joint inference proof in Section A.1 Specifically, for $j = 1, \ldots, p$ let

$$\tilde{C}_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_j^2}{N}}, \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_j^2}{N}}\right],$$

where σ_j^2 is the j-the diagonal element of $\Sigma_\star$. By the asymptotic normality of θ_N in Corollary A.4, we obtain

$$\lim \inf_{N\to\infty} P\big(\theta_{\star,j} \in \tilde{C}_{N,j}(D_N)\big) = 1-\alpha,$$

for a desired significance level α. By the previous bound in Eq. (17), $\text{diag}(Q\Lambda Q^\top) = \text{diag}(\Lambda Q Q^\top) \leq \gamma_1^* I$, where "$\leq$" here denotes element-wise comparison. Then, for any fixed $j \in \{1, \ldots, p\}$, we have

$$\tilde{C}_{N,j}(D_N) \subseteq C_{N,j}(D_N).$$

This implies that C_N are conservative confidence intervals, i.e.,

$$\lim \inf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1-\alpha,$$

□

Theorem 3.2. *Let θ_N be the one-pass SGD in Eq.* (4), *and suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Define the following confidence region:*

$$\widehat{\Theta} = \{\theta \in \Theta : (1/\gamma_1^*)\ ||\theta - \theta_N||^2 < \chi_{\alpha,p}\},$$

where $\chi_{\alpha,p} = \sup\{x \in \mathbb{R} : P(\chi_p^2 \geq x) \leq \alpha\}$ is the α-critical value of a chi-squared random variable on p degrees of freedom. Then,

$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1-\alpha.$$

Proof. Note that by Corollary A.4, we have

$$(\theta_N - \theta_\star)^\top \Sigma_\star^{-1} (\theta_N - \theta_\star) \xrightarrow{\mathrm{d}} \chi_p^2.$$

By Theorem 3.1, we get $\Sigma_\star^{-1} - (1/\gamma_1^*) I \succeq 0$, and so

$$(\theta_N - \theta_\star)^\top \Sigma_\star^{-1} (\theta_N - \theta_\star) - (1/\gamma_1^*)\|\theta_N - \theta_\star\|^2 \geq 0.$$

It follows that

$$P(\theta_\star \notin \widehat{\Theta}) = P\{(1/\gamma_1^*)\|\theta_N - \theta_\star\|^2 \geq \chi_{\alpha,p}\} \leq P\{(\theta_N - \theta_\star)^\top \Sigma_\star^{-1}(\theta_N - \theta_\star) \geq \chi_{\alpha,p}\} = \alpha.$$

Thus, by the Portmanteau Theorem (Kallenberg, 1997), we get

$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1 - \limsup_{N\to\infty} P(\theta_\star \notin \widehat{\Theta}) \geq 1 - \alpha.$$

□

Theorem 3.3. *Let* $C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}, \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right]$*, for* $j = 1, \ldots, p$*. Let* $\lambda_{\min} = \min_j \lambda_j$ *and* $\lambda_{\max} = \max_j \lambda_j$*, and define* $\rho = 1/\gamma_1^* \lambda_{\min}$*. Suppose that the learning is well specified, such that* $\rho \leq 1$*. Define*

$$\eta = \max_{j=1,\ldots,p} \left\{\lim \inf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) - (1-\alpha)\right\}$$

as the worst-case overcoverage across all components of $\theta_\star$*. Let* $\kappa = \lambda_{\max}/\lambda_{\min} \geq 1$ *be the condition number of* $F_\star$*. Then,*

$$\alpha - 2\Phi\bigg(-z_{\frac{\alpha}{2}}\sqrt{2-\rho}\bigg) \leq \eta \leq \alpha - 2\Phi\bigg(-z_{\frac{\alpha}{2}}\sqrt{2-\rho/\kappa}\bigg),$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ *is the critical value of standard normal.*

Proof. Fix some component j. Denote σ_j^2 as the j-th diagonal element of $\Sigma_\star$. By Corollary A.4 the confidence interval

$$\left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\sigma_j^2/N}, \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\sigma_j^2/N}\right]$$

has the nominal $100(1-\alpha)\%$ level in the limit. Our goal is to understand how much we overcover when we use the same γ_1^* on all coordinates. The coverage for component j in this case is:

$$P\left(\left|\frac{\theta_{N,j} - \theta_{\star,j}}{\sqrt{\gamma_1^*/N}}\right| \leq z_{\frac{\alpha}{2}}\right) = P\left(\left|\frac{\theta_{N,j} - \theta_{\star,j}}{\sqrt{\sigma_j^2/N}}\right| \leq z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/\sigma_j^2}\right) = 1 - 2\Phi\left(-z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/\sigma_j^2}\right).$$

The amount of overcoverage for component j is therefore equal to

$$1 - 2\Phi\left(-z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/\sigma_j^2}\right) - (1-\alpha) = \alpha - 2\Phi\left(-z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/\sigma_j^2}\right).$$

It follows that

$$\eta = \max_j \left\{\alpha - 2\Phi\left(-z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/\sigma_j^2}\right)\right\} = \alpha - 2\Phi\left(-z_{\frac{\alpha}{2}}\sqrt{\max_j \gamma_1^*/\sigma_j^2}\right).$$

By construction (see Theorem 3.1), and since $\sigma_j^2 \leq \max_j \frac{\gamma_1^{*2}\lambda_j}{2\gamma_1^*\lambda_j - 1}$, it follows that:

$$\max_j \frac{\gamma_1^*}{\sigma_j^2} \geq \gamma_1^* \frac{1}{\max_j \frac{\gamma_1^{*2}\lambda_j}{2\gamma_1^*\lambda_j - 1}} = \gamma_1^* \min_j \frac{2\gamma_1^*\lambda_j - 1}{\gamma_1^{*2}\lambda_j} = 2 - \max_j \frac{1}{\gamma_1^*\lambda_j} = 2 - \frac{1}{\gamma_1^*}\frac{1}{\lambda_{\min}} = 2 - \rho.$$

Therefore, $\eta \geq \alpha - 2\Phi(-z_{\frac{\alpha}{2}}\sqrt{2-\rho})$.

For an upper bound, we can use the fact $\sigma_j^2 \geq \min_j \frac{\gamma_1^{*2}\lambda_j}{2\gamma_1^*\lambda_j - 1}$ and get

$$\max_j \frac{\gamma_1^*}{\sigma_j^2} \leq \gamma_1^* \frac{1}{\min_j \frac{\gamma_1^{*2}\lambda_j}{2\gamma_1^*\lambda_j - 1}} = \gamma_1^* \max_j \frac{2\gamma_1^*\lambda_j - 1}{\gamma_1^{*2}\lambda_j} = 2 - \min_j \frac{1}{\gamma_1^*\lambda_j} = 2 - \frac{1}{\gamma_1^*}\frac{1}{\lambda_{\max}} = 2 - \rho/\kappa.$$

Therefore, we can upper bound $\eta \leq \alpha - 2\Phi(-z_{\frac{\alpha}{2}}\sqrt{2-\rho/\kappa})$.

□

Algorithm 2 lin_asym: Selecting γ_1^* based on the linear asytmptote in $\Sigma_\star$ (see Section 4.1)

```
Input: Data $D_N$, SGD procedure of Eq. (4), $\theta_0$, $(u, l)$ bounds on $\lambda_{\min}$, kReps, kGammas
 1: $g \leftarrow$ vector$(0.5/u,\ \ldots,\ 2/l,$ length $=$ kGammas$)$
 2: $t \leftarrow$ vector() {calculate empirical covariance}
 3: for $\gamma$ in $g$ do
 4:   $T \leftarrow$ matrix()
 5:   for $i$ in 1 to kReps do
 6:     $T$.rowstack( SGD($\gamma$, bootstrap($D_N$)) )
 7:   end for
 8:   $t$.append( $\mathrm{tr}(\mathrm{covar}(T))$ )
 9: end for
10: $l \leftarrow 0$
11: $M \leftarrow$ matrix() {search for largest region matching expected linear fit}
12: for $i$ in 1 to kGammas $- 1$ do
13:   for $j$ in $(i+1)$ to kGammas do
14:     if $j - i > 4$ then
15:       $f \leftarrow$ lm($t[i:j] \sim g[i:j]$) {linear regression trace vs gamma}
16:       $c \leftarrow$ confint($f$.slope)
17:       if $p/2 \in c$ and shapiro(res($f$)) $> 0.1$) then
18:         $M$.rowstack( vector($i, j, j - i$) )
19:         $l \leftarrow$ max($l,\ j - i$)
20:       end if
21:     end if
22:   end for
23: end for
24: if $l < 5$ then
25:   [Warning]: There is no good learning rate for this problem.
26: end if
27: $m \leftarrow$ whichmax($M[:,2]/\sqrt{M[:,0]}$) {select largest and lowest linear fit region}
28: $i \leftarrow$ int( median($M[m,0], M[m,1]$) )
Output: $\gamma_1^* \leftarrow g[i]$
```

D Full details on selection of γ_1^*

D.1 Linear asymptote in $\Sigma_\star$

Full details of this heuristic are shown in Algorithm 2. Lines 3 to 9 empirically estimate $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ for a range of candidate learning rate values. In our experiments we set the number of candidate learning rate values to kGammas $= 30$ and ran SGD for each learning rate kReps $= 100$. From lines 11 to 23, we check for the correct slope and for linear fits on all continuous subsets of the learning rate and covariance trace. We use the normality test by Shapiro and B. (1965) on the linear regression residuals. The best subset region is selected by weighting the largest length, and inversely weighting the square root of the leftmost point. The middle point of this "confidence region" is returned as the estimate for γ_1^*.

To illustrate how this heuristic is expected to work in practice, we run a simulation with $p = 20$, $N = 10,000$, and $X \sim N_p(0, I)$. The outcomes are generated as $Y_i = X_i^\top \theta_\star + \varepsilon_i$, with $\varepsilon_i \sim N(0,1)$. Figure 3 depicts $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ of the one-pass SGD procedure as a function of the learning rate (x-axis). In this particular example, $F_\star = \mathrm{E}(XX^\top) = I$, and so the optimal selection for the learning rate would be $\gamma_1 = 1$ in terms of minimum asymptotic variance. Indeed, we see that the trace is minimized around this value in Figure 3. Our heuristic procedure calculates $\gamma_1^* \approx 2.6$, which is when the trace of $N\mathrm{Var}(\theta_N)$ appears to become linear to the learning rate. This is indicated by the vertical purple dashed line in Figure 3.

D.2 Selecting γ_1^* from estimating an eigenvalue bound

Full details of the selection procedure of $\gamma_\star$ using the inverse power iteration is provided in Algorithm 3. Line 1 and 2, respectively, correspond to obtaining SGD estimate of $\theta_\star$ and constructing a finite sample approximation of $F_\star$. The

Algorithm 3 `eig_bound`: Selecting γ_1^* by estimating an eigenvalue bound of $F_\star$ (see Section 4.1)

```
Input: Data $D_N$, Error Threshold $\epsilon$, $\theta_0$
$\theta_1 \leftarrow \texttt{SGD}(0.01, D_N, \theta_0)$
$F_n \leftarrow \frac{1}{N}\sum_{i=1}^{N} \nabla\ell(\theta_1; Y_i, X_i)\nabla\ell(\theta_1; Y_i, X_i)^T$
Sample $x_0 \sim \mathcal{N}(0, I)$.
$v_c \leftarrow \frac{x_0}{||x_0||}$
$v_p \leftarrow 0$
ERR $\leftarrow ||v_c - v_p||$
while ERR $> \epsilon$ do
    $v_p = v_c$
    Solve $F_n q = v_c$
    $v_c \leftarrow \frac{q}{||q||}$
    ERR $\leftarrow ||v_c - v_p||$
end while
$\lambda = (v_c)^T F_n v_c$
Output: $\frac{1}{\lambda}$
```

subsequent procedures comprises the standard inverse power iteration given in Trefethen and Bau III (1997).

E Comparing empirical settings to other SGD inference methods

Comparison to Chen et al. (2020)

- Our method can operate on 20X more model parameters (4000 vs 200) with the same number of samples ($N = 1e5$). Standard MLE packages can mostly handle the data sizes of Chen et al. (2020) experiments.

- We can also see the practical difficulty of tuning the batch-means methods of Chen et al. (2020) (implemented in Lee et al. (2022a)). Lee's implementation under covers by about 20%, in part due to the difficulty of tuning the hyperparameters (k batch sizes) to ensure the weak correlation requirement.

Comparison to Lee et al. (2022a)

- Their experiments consider a parameter vector dimension of 800 for the inference of a single parameter. For joint inference, their paper's maximum parameter dimension is 200. In comparison we consider a dimension of 4000 for all components of the parameter.

- Lee et al. (2022a).'s method still requires estimating the covariance matrix, albeit in an online manner. Our method constructs confidence intervals with only the scalar learning rate.

- Lee et al. (2022a) only consider independent covariate structures, while we showcase our method for four different scenarios including ill-conditioned cases.

F Supplemental experimental results

F.1 Simulation details

Our SGD-Asym heuristic for γ_1^* described in Section 4.1 requires a search region, which we set to $(0.5/\tilde{\lambda}, 2/\tilde{\lambda})$, where $\tilde{\lambda}$ was a crude bound for $\lambda_{\min}$ directly calculated from data, and using bounds from Yamamoto (2017).

Configurations of the true parameter $\theta_\star$:

- Exponential (Exp): $\theta_{\star,i} = 2(-1)^i e^{-.7i}$.
- Linear (Lin): $\theta_{\star,i} = (i-1)/(p-1)$.

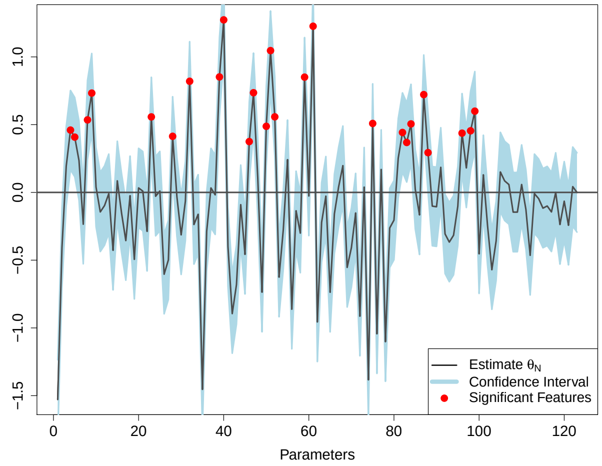


Figure 4: One-pass SGD estimates θ_N with confidence intervals from our simple inference procedure on the UCI Adult dataset. Statistically significant features marked with red circles.

F.2 Adult dataset

Real data experiments were performed on a preprocessed version of the Adult dataset[3] which has 123 binary features and 32,561 samples. One-pass SGD is used to fit logistic regression. In Figure 4 the estimates θ_N of one-pass SGD are plotted along with their confidence intervals. There are several features marked in red circles which which are statistically significant, in that their confidence intervals do not contain zero.

Bounds from Yamamoto (2017) were used to calculate bounds on the search region for the learning rate. The heuristic learning rate selection method in Algorithm 2 chose $\gamma_1^* = 725.397$. Here the numerical stability of implicit SGD plays an important role, such a high learning rate would easily make SGD with classical updates diverge.

Interestingly, the `glm()` function in R which has been used to compute benchmark confidence intervals and which uses iteratively re-weighted least squares was unable to converge, even when increasing the max number of iterations $4\times$ past the default setting, or when relaxing the convergence tolerance from 10^{-8} up to 10^{-3}.

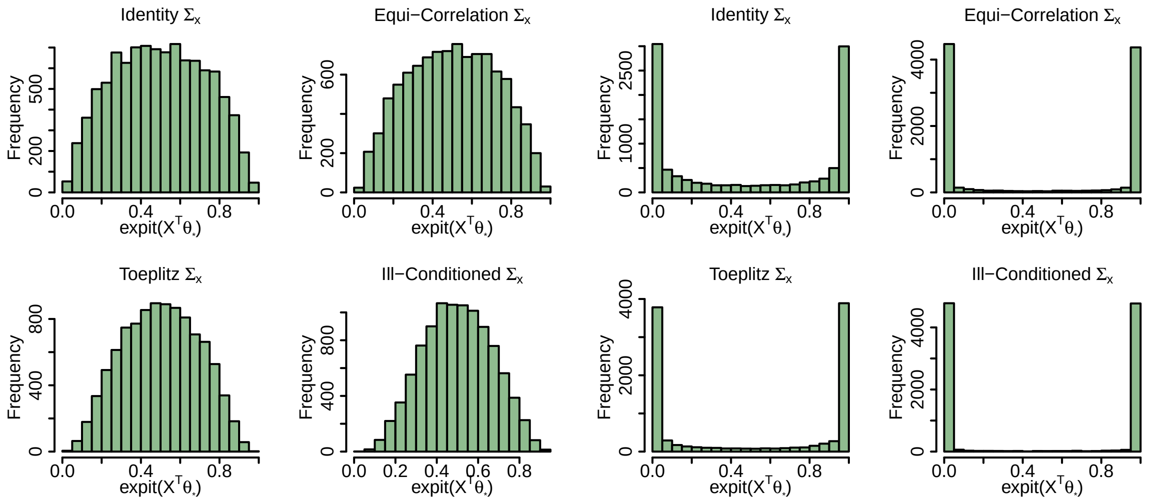
F.3 Logistic regression and Linear $\theta_\star$ setting

We observed that for logistic regression and Linear configuration of $\theta_\star$, both the SGD and benchmark MLE inference procedures performed comparably worse than other experimental settings. We tentatively believe the data generating process contributed to this decrease in performance. We observed a difference in distributions of $\text{expit}(X^\top \theta_\star)$ which we believe affected the inference procedures due to an increase in the variance of $Y|X \sim \text{Bern}(\text{expit}(X^\top \theta_\star))$ based on the data generating process $X \sim N_p(0, \Sigma_x)$. In Figure 5a, the Exponential configuration of $\theta_\star$ is used to plot histograms of $\text{expit}(X^\top \theta_\star)$. Unimodal distributions are observed for all configurations $X \sim N_p(0, \Sigma_x)$ of the data generating process. However, in Figure 5b for the Linear configuration of $\theta_\star$ we see a bimodal distribution with high concentration of probability at 0 and 1, for all configurations of Σ_x. We believe this increased variance in $\text{expit}(X^\top \theta_\star)$ adds additional noise between the covariates X and the response Y.

F.4 Direct comparisons to Chen et al. (2020).

We include experimental results from Chen et al. (2020) to compare to another SGD-based inference procedure. Our simulation setup is modified from theirs, and we include their results in experimental settings where we also have results. We see that the confidence intervals of Chen et al. (2020) tend to under cover, whereas our intervals tend to over cover. In addition, in the $(p < N)$ regime the largest problem they tackle is $p = 200$, $N = 10^5$. Empirically, we have demonstrated in Section 5.1 the greater scalability of our method. With the same number of samples, our SGD-based inference procedure has been shown to provide near-nominal confidence intervals for up to $20\times$ more parameters. In addition, Section 2.4 also highlights the extent of heavy manual tuning required to make this method work.

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



(a) Exponential configuration of $\theta_\star$.

(b) Linear configuration of $\theta_\star$.

Figure 5: Histograms of $\mathrm{expit}(X^\top\theta_\star)$ for all four configurations of $X \sim N_p(0, \Sigma_x)$ of the data generating process, with $N = 10^4$, $p = 100$. Logistic regression model.

$\theta_\star$	Σ_x	p, N	Plug-in CovRate (%)	Plug-in AvgLen ($\times 10^{-2}$)	BM($n^{0.25}$) CovRate (%)	BM($n^{0.25}$) AvgLen ($\times 10^{-2}$)
Lin	Id	20, 1e5	94.99	1.44	93.92	1.41
		100, 1e5	95.04	1.41	93.15	1.35
	EC	20, 1e5	95.10	1.59	93.66	1.54
		100, 1e5	94.93	1.56	93.19	1.52
	T	20, 1e5	94.84	1.81	93.75	1.78
		100, 1e5	95.01	1.77	91.83	1.67

Table 4: Linear regression results from Chen et al. (2020) with matching (p, N), and nominal coverage probability 95%. For brevity we include their plug-in and batch means ($M = n^{0.25}$) estimators.

F.5 Compute budget for experiments.

Experiments were conducted on a university compute cluster and personal laptop. The cluster uses Intel E5-2680v4 2.4GHz CPUs, and the laptop is an Apple 2020 Macbook Pro M1. Each simulation ran 500 independent trails. For the SGD-Asym selection method: 30 learning rate values were evaluated, and for each SGD was run 100 times to estimate the variance of the estimate. To initialize θ_0, we run $10 * \sqrt{p}$ passes over averaged SGD with learning rate 1. The SGD-Eig selection method is run until the error is below 1e-3. After γ_1^* has been selected, we run (implicit) SGD for 1 epoch.

$\theta_\star$	Σ_x	p, N	Plug-in CovRate (%)	Plug-in AvgLen ($\times 10^{-2}$)	BM($n^{0.25}$) CovRate (%)	BM($n^{0.25}$) AvgLen ($\times 10^{-2}$)
Lin	Id	20, 1e5	95.00	3.79	90.22	3.46
		100, 1e5	94.69	5.21	90.84	4.87
	EC	20, 1e5	94.54	5.37	90.64	4.77
		100, 1e5	94.79	10.24	90.27	9.75
	T	20, 1e5	95.17	5.74	90.39	5.22
		100, 1e5	94.91	8.47	90.83	7.71

Table 5: Logistic regression results from Chen et al. (2020) with matching (p, N), and nominal coverage probability 95%. For brevity we include their plug-in and batch means ($M = n^{0.25}$) estimators.

F.6 Full set of simulation results

Finally, we include a full set of simulation results, which we include a informative subset in the main paper.

$\theta_\star$	Σ_x	p	N	SGD Cov Rate (%)	SGD Avg Len ($\times 10^{-2}$)	MLE Cov Rate (%)	MLE Avg Len ($\times 10^{-2}$)
Exponential	Identity	10	1e4	95.82	4.16	94.46	3.92
	Identity	20	1e4	96.13	4.21	94.62	3.92
	Identity	50	1e4	96.41	4.33	94.75	3.93
	Identity	100	1e5	96.01	1.31	95.05	1.24
	Identity	500	1e5	96.95	1.40	95.07	1.24
	Equi-Corr	10	1e4	96.60	4.46	95.38	4.11
	Equi-Corr	20	1e4	96.23	4.47	95.10	4.17
	Equi-Corr	50	1e4	96.88	4.80	94.86	4.22
	Equi-Corr	100	1e5	96.12	1.42	94.97	1.34
	Equi-Corr	500	1e5	96.89	1.52	95.07	1.35
	Toeplitz	10	1e4	97.88	6.71	94.74	4.95
	Toeplitz	20	1e4	97.95	6.87	94.88	5.01
	Toeplitz	50	1e4	98.20	7.08	95.16	5.05
	Toeplitz	100	1e5	98.02	2.18	95.02	1.60
	Toeplitz	500	1e5	98.17	2.28	94.93	1.60
	Ill-Cond	10	1e4	99.84	13.61	94.64	1.79
	Ill-Cond	20	1e4	99.95	14.51	94.92	1.25
	Ill-Cond	50	1e4	100.00	15.79	94.99	0.94
	Ill-Cond	100	1e5	99.94	4.28	94.82	0.27
	Ill-Cond	500	1e5	99.97	4.89	95.01	0.25
Linear	Identity	10	1e4	96.68	4.18	95.46	3.92
	Identity	20	1e4	96.18	4.18	95.00	3.93
	Identity	50	1e4	96.76	4.42	94.91	3.93
	Identity	100	1e5	96.04	1.31	95.08	1.24
	Identity	500	1e5	96.85	1.40	95.01	1.24
	Equi-Corr	10	1e4	96.36	4.45	95.06	4.12
	Equi-Corr	20	1e4	96.85	4.66	95.36	4.17
	Equi-Corr	50	1e4	96.98	4.79	95.03	4.23
	Equi-Corr	100	1e5	95.98	1.42	94.92	1.34
	Equi-Corr	500	1e5	96.87	1.52	95.00	1.35
	Toeplitz	10	1e4	97.78	6.67	95.36	4.96
	Toeplitz	20	1e4	98.10	6.87	95.23	5.01
	Toeplitz	50	1e4	98.17	7.28	95.08	5.05

	Toeplitz	100	1e5	98.04	2.19	95.08	1.60
	Toeplitz	500	1e5	98.18	2.26	94.99	1.60
	Ill-Cond	10	1e4	99.78	13.53	95.20	1.79
	Ill-Cond	20	1e4	99.96	14.45	94.55	1.25
	Ill-Cond	50	1e4	100.00	15.97	95.06	0.94
	Ill-Cond	100	1e5	99.95	4.30	94.90	0.27
	Ill-Cond	500	1e5	100.00	4.89	94.97	0.25

Table 6: Linear regression and learning rate γ_1^* set to $1/\lambda_{\min}$, where $\lambda_{\min}$ is assumed to be known. The average coverage rate and interval lengths are calculated for a target coverage probability of $1-\alpha=.95$.

				SGD	SGD	MLE	MLE
$\theta_\star$	Σ_x	p	N	Cov Rate	Avg Len	Cov Rate	Avg Len
				(%)	$(\times 10^{-2})$	(%)	$(\times 10^{-2})$
Exponential	Identity	10	1e4	97.14	10.82	95.32	9.05
	Identity	20	1e4	97.30	11.13	94.92	8.95
	Identity	50	1e4	96.71	11.04	95.14	8.91
	Identity	100	1e5	97.34	3.47	94.89	2.80
	Identity	500	1e5	97.17	3.53	94.92	2.80
	Equi-Corr	10	1e4	97.22	11.49	95.22	9.40
	Equi-Corr	20	1e4	97.05	11.60	95.02	9.42
	Equi-Corr	50	1e4	96.64	11.56	94.81	9.49
	Equi-Corr	100	1e5	97.45	3.67	94.99	2.99
	Equi-Corr	500	1e5	96.91	3.68	94.84	3.01
	Toeplitz	10	1e4	97.64	14.88	94.72	10.93
	Toeplitz	20	1e4	97.35	15.01	94.64	10.96
	Toeplitz	50	1e4	97.54	15.53	95.38	11.00
	Toeplitz	100	1e5	97.73	4.75	95.05	3.47
	Toeplitz	500	1e5	97.30	4.86	94.91	3.48
	Ill-Cond	10	1e4	86.80	33.11	94.68	5.01
	Ill-Cond	20	1e4	89.84	30.27	94.66	3.10
	Ill-Cond	50	1e4	93.78	27.85	94.68	2.10
	Ill-Cond	100	1e5	96.37	8.52	94.96	0.57
	Ill-Cond	500	1e5	98.78	8.23	94.89	0.50
Linear	Identity	10	1e4	97.48	14.88	95.08	10.52
	Identity	20	1e4	97.07	20.29	94.82	11.67
	Identity	50	1e4	88.49	34.65	94.48	13.97
	Identity	100	1e5	82.62	16.52	94.76	5.09
	Identity	500	1e5	63.77	55.77	88.09	7.65
	Equi-Corr	10	1e4	96.80	15.87	95.08	12.17
	Equi-Corr	20	1e4	95.89	22.48	95.21	15.38
	Equi-Corr	50	1e4	89.56	54.93	94.5.0	23.19
	Equi-Corr	100	1e5	80.20	29.41	94.55	10.00
	Equi-Corr	500	1e5	100.00	222.09	48.40	28.80
	Toeplitz	10	1e4	98.26	20.31	95.82	14.99
	Toeplitz	20	1e4	95.57	24.15	94.48	17.74
	Toeplitz	50	1e4	87.51	42.14	94.58	22.40
	Toeplitz	100	1e5	79.18	20.15	95.04	8.31
	Toeplitz	500	1e5	85.14	73.68	87.07	12.97
	Ill-Cond	10	1e4	50.70	84.39	94.74	16.61
	Ill-Cond	20	1e4	58.44	92.99	94.32	15.49
	Ill-Cond	50	1e4	100.00	130.84	81.35	17.82
	Ill-Cond	100	1e5	69.91	45.35	86.97	5.74

Ill-Cond	500	1e5	100.00	249.43	11.00	12.93

Table 7: Logistic regression and learning rate γ_1^* set to $1/\lambda_{\min}$, where $\lambda_{\min}$ is assumed to be known. The average coverage rate and interval lengths are calculated for a target coverage probability of $1-\alpha=.95$.

$\theta_\star$	Σ_x	p	N	SGD-Asym	SGD-Asym	MLE	MLE
				Cov Rate	Avg Len	Cov Rate	Avg Len
				(%)	$(\times 10^{-2})$	(%)	$(\times 10^{-2})$
Exponential	Identity	10	1e4	98.70	6.26	95.32	3.92
	Identity	20	1e4	98.83	6.06	95.30	3.92
	Identity	50	1e4	98.89	6.92	95.02	3.93
	Identity	100	1e5	98.24	1.69	95.24	1.24
	Identity	500	1e5	98.86	2.14	95.03	1.24
	Equi-Corr	10	1e4	98.70	6.36	94.84	4.12
	Equi-Corr	20	1e4	98.35	5.93	95.07	4.17
	Equi-Corr	50	1e4	98.75	6.73	95.22	4.22
	Equi-Corr	100	1e5	98.97	2.44	94.98	1.34
	Equi-Corr	500	1e5	99.00	2.51	95.04	1.35
	Toeplitz	10	1e4	99.02	10.04	94.78	4.96
	Toeplitz	20	1e4	99.27	10.94	95.37	5.01
	Toeplitz	50	1e4	99.18	11.77	94.93	5.05
	Toeplitz	100	1e5	99.03	3.13	94.94	1.60
	Toeplitz	500	1e5	99.29	3.95	94.97	1.60
	Ill-Cond	10	1e4	99.98	29.91	94.86	1.79
	Ill-Cond	20	1e4	100.00	31.42	95.11	1.25
	Ill-Cond	50	1e4	100.00	34.63	94.95	0.94
	Ill-Cond	100	1e5	100.00	9.65	95.03	0.27
	Ill-Cond	500	1e5	100.00	10.61	94.98	0.25
Linear	Identity	10	1e4	98.98	6.23	94.72	3.92
	Identity	20	1e4	98.88	6.87	94.80	3.92
	Identity	50	1e4	98.99	6.89	95.24	3.93
	Identity	100	1e5	98.91	2.30	95.15	1.24
	Identity	500	1e5	98.98	2.31	94.91	1.24
	Equi-Corr	10	1e4	98.56	6.23	95.14	4.12
	Equi-Corr	20	1e4	98.75	7.09	95.27	4.17
	Equi-Corr	50	1e4	98.94	7.87	94.84	4.23
	Equi-Corr	100	1e5	98.88	2.34	95.07	1.34
	Equi-Corr	500	1e5	99.02	2.56	94.97	1.35
	Toeplitz	10	1e4	99.12	10.37	95.66	4.96
	Toeplitz	20	1e4	99.22	9.95	94.94	5.01
	Toeplitz	50	1e4	99.14	10.29	94.98	5.05
	Toeplitz	100	1e5	99.22	3.69	94.92	1.60
	Toeplitz	500	1e5	99.37	4.20	95.13	1.60
	Ill-Cond	10	1e4	100.00	30.82	95.06	1.79
	Ill-Cond	20	1e4	100.00	33.00	94.68	1.25
	Ill-Cond	50	1e4	100.00	32.78	94.90	0.94
	Ill-Cond	100	1e5	100.00	9.52	94.91	0.27
	Ill-Cond	500	1e5	100.00	10.72	94.96	0.25

Table 8: Linear regression and SGD-Asym γ_1^* selection method (see Section 4.1, Appendix D.1). The average coverage rate and interval lengths are calculated for a target coverage probability of $1-\alpha=.95$.

				SGD-Asym	SGD-Asym	MLE	MLE
$\theta_\star$	Σ_x	p	N	Cov Rate	Avg Len	Cov Rate	Avg Len
				(%)	($\times 10^{-2}$)	(%)	($\times 10^{-2}$)
Exponential	Identity	10	1e4	98.92	17.60	95.04	9.05
	Identity	20	1e4	98.93	17.41	94.87	8.95
	Identity	50	1e4	99.00	17.69	94.94	8.91
	Identity	100	1e5	99.07	5.90	95.07	2.80
	Identity	500	1e5	99.02	5.84	95.00	2.80
	Equi-Corr	10	1e4	98.96	17.62	94.70	9.40
	Equi-Corr	20	1e4	99.15	17.88	95.48	9.41
	Equi-Corr	50	1e4	98.84	18.58	94.86	9.48
	Equi-Corr	100	1e5	98.98	5.65	94.95	2.99
	Equi-Corr	500	1e5	98.75	5.47	94.85	3.01
	Toeplitz	10	1e4	98.96	23.3	94.66	10.93
	Toeplitz	20	1e4	99.01	22.89	94.92	10.96
	Toeplitz	50	1e4	98.98	23.87	94.80	11.00
	Toeplitz	100	1e5	99.10	7.42	94.78	3.47
	Toeplitz	500	1e5	99.00	8.16	94.95	3.48
	Ill-Cond	10	1e4	87.28	33.37	94.46	5.02
	Ill-Cond	20	1e4	89.22	59.82	95.22	3.10
	Ill-Cond	50	1e4	94.26	55.78	95.14	2.11
	Ill-Cond	100	1e5	96.57	6.83	94.92	0.57
	Ill-Cond	500	1e5	99.05	16.49	94.95	0.50
Linear	Identity	10	1e4	98.92	20.12	95.50	10.52
	Identity	20	1e4	99.04	30.35	95.36	11.67
	Identity	50	1e4	94.84	51.51	94.77	13.97
	Identity	100	1e5	95.51	23.66	94.81	5.09
	Identity	500	1e5	71.83	110.39	88.52	7.65
	Equi-Corr	10	1e4	99.18	24.62	95.02	12.16
	Equi-Corr	20	1e4	98.92	35.03	95.00	15.35
	Equi-Corr	50	1e4	93.09	79.90	94.25	23.19
	Equi-Corr	100	1e5	89.88	43.19	94.79	10.00
	Equi-Corr	500	1e5	100.00	414.28	47.88	28.88
	Toeplitz	10	1e4	98.90	29.17	95.50	15.01
	Toeplitz	20	1e4	98.86	36.97	95.13	17.74
	Toeplitz	50	1e4	90.48	51.52	94.49	22.44
	Toeplitz	100	1e5	92.88	29.40	94.85	8.31
	Toeplitz	500	1e5	94.73	146.82	86.73	12.97
	Ill-Cond	10	1e4	59.36	158.56	94.72	16.66
	Ill-Cond	20	1e4	80.28	196.15	93.65	15.50
	Ill-Cond	50	1e4	100.00	290.55	81.20	17.75
	Ill-Cond	100	1e5	93.43	94.07	88.03	5.74
	Ill-Cond	500	1e5	100.00	567.26	10.95	12.97

Table 9: Logistic regression and SGD-Asym γ_1^* selection method (see Section 4.1, Appendix D.1). The average coverage rate and interval lengths are calculated for a target coverage probability of $1 - \alpha = .95$.

G Results for medical application

Significant predictive features for an adverse EGS event		
SGD Only	Both	MLE Only
+ 8 Elixhauser comorbidities - Age - {Commercial, Self Pay, Other} insurance - General abdominal - Hepato-pancreatico-biliary - Upper GI	+ {Emergency, Urgent} admission + Medicaid insurance + Surgery performed (Yes) + High risk disability score + High risk Angus sepsis score - {0-4} Elixhauser comorbidities	- Other race (not w,b) - 6 Elixhauser comorbidities

Table 10: Significant predictive features for adverse EGS events.