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

Jerry Chee, Hwanwoo Kim, & Panos Toulis   Contact: hwkim@uchicago.edu

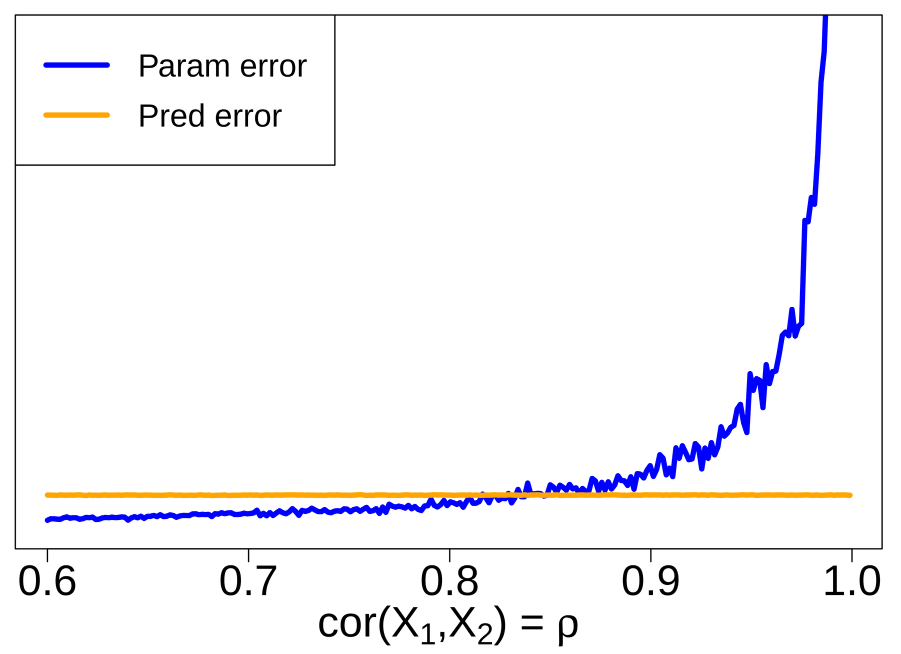
Overview

- We develop a **statistical inference** procedure using **stochastic gradient descent** (SGD)-based confidence intervals.
- These intervals are of the **simplest form**:

$$\theta_{N,j} \pm 2\sqrt{\gamma/N}.$$

- This construction is **simple** as it relies only on properly selecting the learning rate (γ).
- The procedure achieves **near-nominal coverage intervals** scaling up to **20×** **more parameters** than other SGD-based methods.

Motivation: Prediction vs Inference



Multicollinearity degrades parameter estimation error but not prediction error.

Background

Statistical inference setup. Consider data $(Y, X) \in \mathbb{R}^d \times \mathbb{R}^p$, negative log-likelihood ℓ, and unknown model parameters:

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

The empirical loss minimizer $\widehat{\theta}_N = \arg\min_{\theta\in\Theta} \sum_{i=1}^N \ell(\theta; Y_i, X_i)$ admits weak convergence results of the form

$$\sqrt{N}(\widehat{\theta}_N - \theta_\star) \xrightarrow{d} N_p(0, F_\star^{-1}), \quad (1)$$

where $F_\star$ is the Fisher information matrix. This can be used to construct 95% confidence intervals (CIs):

$$\hat{\theta}_{N,j} \pm 2\sqrt{F^{-1}_{\star,jj}/N}. \quad \text{[MLE-based inference]} \quad (2)$$

But, $\widehat{\theta}_N$ cannot be efficiently computed in large data sets.

SGD: A scalable approach. Instead, we may look at SGD:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla\ell(\theta_{n-1}; Y_{I_n}, X_{I_n}), \quad (3)$$

with $I_n \sim U\{1\ldots N\}$ is a random datapoint, $\gamma_n = \gamma_1/n$ the learning rate. There are two potential choices. Which one should we use?

Choice 1: Averaged SGD, $\bar{\theta}_N$

$\bar{\theta}_N = \frac{1}{N}\sum_{i=1}^N \theta_i$ has optimal weak convergence of Eq. (1). Most SGD-based inference uses the theoretical optimality of $\bar{\theta}_N$, and construct CI as in Eq. (2).

Other methods are not simple. Practically, these methods require significant data-dependent calibration. For example, Chen et al. 2020 requires tuning their (a) number of batches, (b) multiple batch sizes, (c) decorrelation parameter, and (d) learning rate.

Choice 2: One-Pass SGD, θ_N (Our Method)

We propose an inference method based on θ_N in Algorithm 1. Under regularity conditions (Toulis et al., 2017, Ljung et al., 1992, II.8),

$$\sqrt{N}(\theta_N - \theta_\star) \xrightarrow{d} N_p(0, \Sigma_\star), \text{ where } \Sigma_\star = \gamma_1^2(2\gamma_1 F_\star - I)^{-1}F_\star. \quad (4)$$

Our method is simple.

1. The asymptotic variance $\Sigma_\star$ is known in closed form in Eq. (4).
2. We can bound $\Sigma_\star \preceq \gamma_1^* I$ which only depends on the learning rate.

Pros. We only need to estimate a *single* parameter (γ_1^*), instead of the $p \times p$ covariance matrix.
Cons. Our CIs are conservative and exhibit some overcoverage. We still need to select γ_1^*.

Main Idea

Let λ_j be the j-th eigenvalue of $F_\star$. Then, the corresponding eigenvalue of $\Sigma_\star$ is $\gamma_1^2\lambda_j/(2\gamma_1\lambda_j - 1)$, and thus satisfies:

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

The limit implies a uniform bound on $\Sigma_\star$, and a construction of conservative confidence intervals.

Theorem 3.1. Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Then, $\gamma_1^* I - \Sigma_\star \succ 0$. For every $j = 1, \ldots, p$, the confidence intervals $C_{N,j}$ in Algorithm 1 satisfy:

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

Remark 1. The bound for γ_1^* is standard for $O(1/n)$ convergence of SGD; e.g., see Section 3.1 of Moulines and Bach, 2011.

Remark 2. Joint inference for all or a subset of components of $\theta_\star$ is also possible. See Thm 3.2 in paper.

Remark 3. Overcoverage depends on the condition number of $F_\star$, and misspecification of the learning rate (ρ). See Thm. 3.3 in paper.

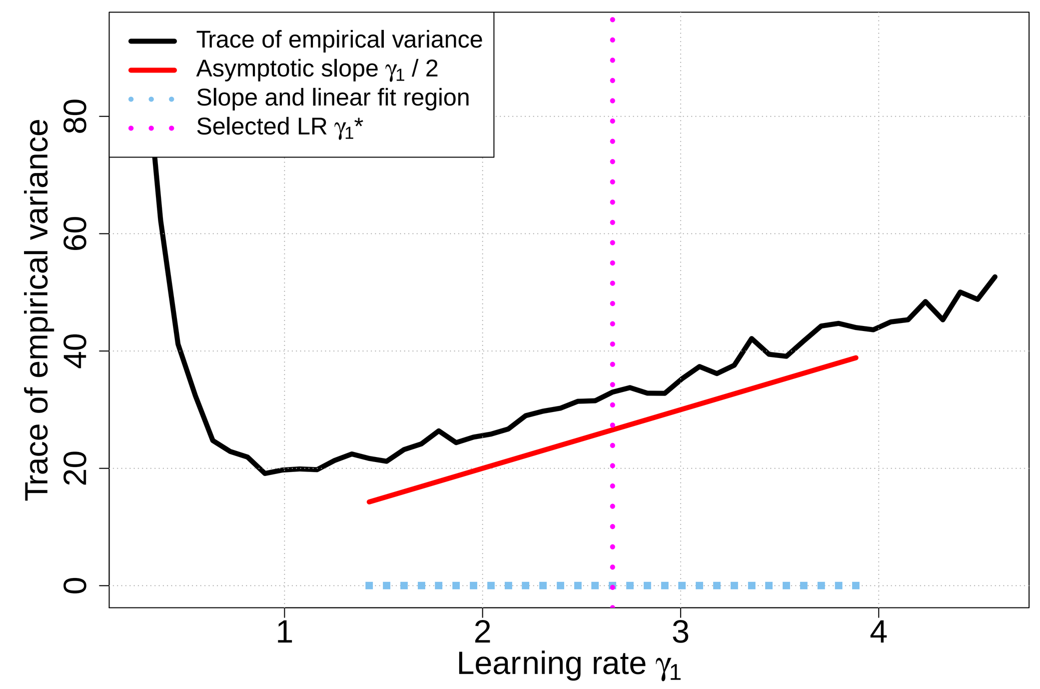
Concrete Method & Implementation

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

Input: Data D_N, SGD procedure of Eq. (3), θ_0, $\alpha \in (0, 1)$.
$\gamma_1^* \leftarrow \texttt{select\_gamma}(D_N, \theta_0)$
$\theta_N \leftarrow \texttt{SGD}(\gamma_1^*, D_N, \theta_0)$
Output: Confidence interval for $\theta_{\star,j}$ with

$$C_{N,j}(D_N) = \left(\theta_{N,j} \pm z_{\frac{\alpha}{2}}\sqrt{\gamma_1^*/N}\right),$$

and $z_{\frac{\alpha}{2}}$ is the critical value of the standard normal.



Selecting γ_1^*. (a) A good estimate of $\min_j\{\lambda_j\}$ already exists. (b) Selection of γ_1^* based on asymptotic results on $\mathrm{eigen}(\Sigma_\star)$. (c) Selection based on inverse power iteration of $F_\star^{-1}$.

Results

Code: github.com/jerry-chee/SGDInference

model	Σ_x	One-Pass SGD CovRate (%)	One-Pass SGD AvgLen ($\times 10^{-2}$)	MLE CovRate (%)	MLE AvgLen ($\times 10^{-2}$)	Avg SGD[a] CovRate (%)	Avg SGD[a] AvgLen ($\times 10^{-2}$)
linear	Id	96.01	1.31	95.05	1.24	93.15	1.35
	EC	96.12	1.42	94.97	1.34	93.19	1.52
	T	98.02	2.18	95.02	1.60	90.83	7.71
logistic	Id	97.34	3.47	94.89	2.80	90.84	4.87
	EC	97.45	3.67	94.99	2.99	90.27	9.75
	T	97.73	4.75	95.05	3.47	90.83	7.71

Selected simulations(500 trials,p=100,N=1e5). Assume $\min_j\{\lambda_j\}$ is known.

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

Large-scale simulations (500 trials). Inverse power iteration to select γ_1^*.

[a] Chen et al. 2020