

Large-Scale Statistical Inference with Stochastic Gradient Descent

Hwanwoo Kim

University of Chicago

Joint work with Jerry Chee, Panos Toulis

2023 Joint Statistical Meetings

Aug 8, 2023

Set up

- $D_N = \{(X_i, Y_i)\}_{i=1}^N$: Observed data
- X_i : Covariate of unit i
- Y_i : Outcome variable of interest for unit i
- (X_i, Y_i) : i.i.d data distributed according to p_{θ_\star}
- $\ell(\theta; X, Y) = -\log p_\theta(X, Y)$: negative log-likelihood (loss)

Goal of statistical inference

Based on data, construct a confidence region $C_N(D_N)$ such that

$$\lim_{N \rightarrow \infty} \mathbb{P}(\theta_\star \in C_N(D_N)) \geq 1 - \alpha$$

for some desired significance level $\alpha \in (0, 1)$.

Standard approach: MLE

In standard approach, the construction relies on asymptotic result of the form $\sqrt{N}(\hat{\theta}_N - \theta_*) \xrightarrow{d} N_p(0, F_*^{-1})$, where $\hat{\theta}_N$ is the maximum likelihood estimator (MLE); i.e.,

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

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

Issues in the standard approach

In large data set,

- Computation of $\hat{\theta}_N$ is expensive!
 - ▶ Newton-Raphson, EM algorithm, or quasi-Newton methods scales at a rate of $O(Np^{1+\epsilon})$
- Estimation of F_\star is notoriously challenging!
 - ▶ Standard estimators of covariance matrices do not scale well and are often ill-conditioned or non-invertible.

Stochastic gradient descent (SGD) estimator

Iteratively defined as:

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

where $I_n \sim U\{1, \dots, N\}$ is a random datapoint, γ_n is the learning rate sequence, and the gradient $\nabla \ell$ is with respect to θ .

- Under mild conditions, SGD converges to $\hat{\theta}_N$ as $n \rightarrow \infty$
- $\hat{\theta}_N$, the estimator obtained after N iterations is known as the one-pass SGD estimator.

Asymptotic result of one-pass SGD estimator

Under regularity conditions, one-pass SGD satisfies [Toulis et al., 2017, Ljung et al., 1992]:

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

where

$$\Sigma_{\text{SGD}} = \gamma_1^2(2\gamma_1 F_* - I)^{-1} F_*.$$

(It is assumed that γ_1 is large enough such that $2\gamma_1 F_* - I \succ 0$.)

Key idea

- The asymptotic covariance depends on the initial learning rate γ_1 , which we can choose.
- The eigenvalues of Σ_{SGD} are of the form

$$\frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1},$$

where λ_j is the j th eigenvalue of the F_* .

- Notice

$$\lim_{\gamma_1 \rightarrow \infty} \frac{\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} / \left(\frac{\gamma_1}{2} \right) \rightarrow 1, \quad (1)$$

which implies the uniform bound on Σ_{SGD} .

- (1) is the basis of our key idea.

Backbone theory

Theorem

Let $\theta_{N,j}$, denote the j -th component of θ_N , for $j = 1, \dots, p$. Suppose that $\gamma_1 \geq 1/\min_j\{\lambda_j\}$, then $\gamma_1 I - \Sigma_{SGD} \succ 0$. 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 every $j = 1, \dots, p$,

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

- The key step in constructing confidence intervals is now the estimation of the minimum eigenvalue of F_{\star} .

Plus-minus the learning rate

Data: D_N , Initial state: θ_0 , Significance level: $\alpha \in (0, 1)$.

- 1 $\gamma_1 \leftarrow \text{select_gamma}(D_N, \theta_0)$
- 2 $\theta_N \leftarrow \text{SGD}(\gamma_1, D_N, \theta_0)$
- 3 Confidence interval for $\theta_{*,j}$ is given by

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

- Joint inference on $\{\theta_j\}_{j=1}^p$ is also possible!

How to select γ_1 ?

Two approaches were considered:

① Heuristic based on asymptotics:

- ▶ For γ_1 large, $\Sigma_{\text{SGD}} \approx \frac{\gamma_1}{2} I$.
- ▶ For N large, $\Sigma_{\text{SGD}} \approx \text{Var}(\sqrt{N}\theta_N)$.
- ▶ Combining both

$$\frac{p}{2}\gamma_1 \approx \text{Tr}(\text{Var}(\sqrt{N}\theta_N))$$

- ▶ Linear regression of $\text{Tr}(\text{Var}(\sqrt{N}\theta_N))$ with respect to γ_1 will give a coefficient around $p/2$ with high confidence.
- ▶ Slowly increase γ_1 till regression coefficients is stabilized around $p/2$.

② Inverse power iteration: Estimate the maximum eigenvalue of the inverse of the Fisher information matrix F_{\star}^{-1} without explicitly computing the inverses.

- Future work: more efficient learning rate selection!

Simulation Studies 1: Coverage Rate

- Features are sampled as $X_i \sim N_p(0, I_p)$
- True parameter θ_* are set to be $\theta_{*,i} = 2(-1)^i e^{-.7i}$.

Linear Model	Sample (n)/Dimension (p)	One-pass SGD	MLE
Coverage Rate (%)	$n = 10^4, p = 50$	96.74	94.75
	$n = 10^5, p = 500$	96.81	95.07
Average Length ($\times 10^{-2}$)	$n = 10^4, p = 50$	4.38	3.93
	$n = 10^5, p = 500$	1.40	1.24

Logistic Model	Sample (n)/Dimension (p)	One-pass SGD	MLE
Coverage Rate (%)	$n = 10^4, p = 50$	96.68	95.14
	$n = 10^5, p = 500$	97.31	94.92
Average Length ($\times 10^{-2}$)	$n = 10^4, p = 50$	11.17	8.91
	$n = 10^5, p = 500$	3.55	2.80

Simulation Studies 2: Scalability and Computational cost





Sample (n)/Dimension (p)	Cov Rate (%)	Avg Length ($\times 10^{-2}$)
$n = 10^5, p = 2 \cdot 10^3$	97.82	1.59
$n = 10^5, p = 4 \cdot 10^3$	98.36	1.78

- Standard MLE library in the R doesn't scale to the above setting.
- One-pass SGD is $21\times$ faster than the MLE; e.g., in the R language, this amounts 0.006 seconds (SGD) vs 0.128 seconds (MLE) for $N = 10^4, p = 100$.

Further Comments

- The methodology performs well in our empirical evaluations, achieving near-nominal coverage intervals scaling up to $20\times$ as many parameters as other SGD-based inference methods.
- The main weakness of the methodology is that it tends to overcover. In particular, the larger the condition number of F_{\star} , the worse the over-coverage rates are \Rightarrow However, over-coverage can be bounded. In the case of significance level of 0.05 (or 0.1), the maximum over-coverage is 0.044 (or 0.08). [Chee et al., 2023].
- More experiments with diverse features, parameter configurations, and comparisons with other SGD-based inferences are available in the paper [Chee et al., 2023].
- All introduced methodologies can be used with an implicit SGD algorithm for stability [Toulis et al., 2017].
- R code available at: <https://github.com/jerry-chee/SGDIInference>.

References

-  Chee, J., Kim, H., and Toulis, P. (2023).
“plus/minus the learning rate”: Easy and scalable statistical inference with sgd.
In International Conference on Artificial Intelligence and Statistics, pages 2285–2309. PMLR.
-  Ljung, L., Pflug, G., and Walk, H. (1992).
Stochastic approximation and optimization of random systems, volume 17.
Springer.
-  Toulis, P., Airoidi, E. M., et al. (2017).
Asymptotic and finite-sample properties of estimators based on stochastic gradients.
The Annals of Statistics, 45(4):1694–1727.
-  Toulis, P., Rennie, J., and Airoidi, E. (2014).
Statistical analysis of stochastic gradient methods for generalized linear models.
In 31st International Conference on Machine Learning.