Central Limit Theorems for Stochastic Gradient Descent

Zehua Lai

University of Chicago, CAM
laizehua@uchicago.edu
joint work with Xi Chen, He Li, Yichen Zhang

April, 2023

Overview

Content

A Quick Introduction to Statistical Inference

If we have n samples of a random variable X, can we estimate the expectation of X?

Answer: $\bar{X}_n := \frac{1}{n}\sum X_i$.

Online updates: $\bar{X}_{n+1} = \frac{1}{n+1}\sum X_i = \frac{1}{n+1}X_{n+1} + \frac{n}{n+1}\bar{X}_n$.

Theorem

$\bar{X}_n$ is unbiased, consistent.

Theorem (Central Limit Theorem, CLT)

$\sqrt{n}(\bar{X}_n - \mathbb{E}X) \Rightarrow N(0, V)$, *where* $V = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top$.

How to estimate V?

Answer: $V_n = \frac{1}{n}\sum(X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top$.

Exercise: try to write down formulas for update it online (iteratively).

Theorem

V_n is biased, consistent.

Now given a direction u, consider $u^\top\sqrt{n}(\bar{X}_n - \mathbb{E}X)$. It converges to $N(0, u^\top Vu)$. So the probability that $u^\top\sqrt{n}(\bar{X}_n - \mathbb{E}X)$ is in the interval $[-2\sqrt{u^\top Vu}, 2\sqrt{u^\top Vu}]$ is around 95%.

$$\Pr\{\sqrt{n}u^\top\mathbb{E}X \in [\sqrt{n}u^\top\bar{X}_n - 2\sqrt{u^\top V_n u}, \sqrt{n}u^\top\bar{X}_n + 2\sqrt{u^\top V_n u}]\} \approx 95\%.$$

A Quick Introduction to Statistical Inference

Consider the following minimization problem,

$$\text{minimize} \;\; L(\theta) := \mathbb{E}_{\zeta \sim P}\left[\ell(\theta; \zeta)\right].$$

We are able to query the sample gradient $\nabla \ell(\theta; \zeta)$, not the full gradient $\nabla L(\theta)$.

What is the *most efficient* way to minimize?

Answer: Stochastic gradient descent with Polyak-Juditsky averaging.

Start from an arbitrary point θ_0. At each step, we compute:

$$\theta_n = \theta_{n-1} - \eta_n \nabla \ell(\theta_{n-1}; \zeta_n),$$

where the step size $\eta_n = \eta_0 n^{-\alpha}$, $\frac{1}{2} < \alpha < 1$.

We can then define the final Polyak-Juditsky averaging estimator as

$$\overline{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i.$$

Theorem (Polyak-Juditsky[1])

Assume the function $L(\theta)$ is strongly convex with global minimizer θ_*. Assume covariance of the sample gradient has enough regularity. Define H, S by

$$H := \nabla^2 L(\theta_*), \quad S := \mathbb{E}\left[\nabla \ell(\theta_*; \zeta) \nabla \ell(\theta_*; \zeta)^\top\right].$$

Then we have a central limit theorem

$$\sqrt{n}\ \left(\overline{\theta}_n - \theta_*\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, H^{-1} S H^{-1}\right), \qquad \text{as} \quad n \to \infty.$$

The choice of η_0, α does not affect the result.

[1] Polyak and Juditsky, SIAM journal on control and optimization (1992).

Consider the case of least square linear regression: $l(\theta) = (y - \theta^\top X)^2$.
$y = \theta_*^\top X + \zeta$.
Now the gradient step is:

$$\theta_n = \theta_{n-1} - \eta_n(XX^\top(\theta_{n-1} - \theta_*) + X\zeta_n).$$

Make the harmless assumption $\theta_* = 0$. This becomes:

$$\theta_n = \theta_{n-1} - \eta_n(A\theta_{n-1} + \xi_n(\theta_{n-1})),$$

where $A = \mathbb{E}[XX^\top], \xi_n(\theta) = (XX^\top - A)\theta + X\zeta_n$.

After some clever algebraic manipulation, we can even prove that

$$\sqrt{n}\bar{\theta}_n = \sqrt{n}H^{-1}\bar{\xi}_n(0) + \text{residual terms}\ldots$$

The residual terms go to 0.

A Quick Introduction to Statistical Inference

Hájek-Le Cam local asymptotic minimax theorem: SGD with Polyak-Juditsky averaging achieves optimal rates asymptotically.

This rate is also called the Cramer-Rao lower bound.

Two variant: gradient-free optimization and contextual bandit optimization.

Content

Gradient-free Optimization

What is Gradient-free Optimization/Zeroth Order Optimization?

Again, we still want to solve the problem

$$\text{minimize} \;\; L(\theta) := \mathbb{E}_P \left[\ell(\theta; \zeta) \right].$$

However, we now only have the access to the sample function value $\ell(\theta; \zeta)$, not the sample gradient $\nabla \ell(\theta; \zeta)$.

Gradient-free Optimization

Let us consider several different cases:

(i) At each step, we can evaluate the function once and get $\ell(\theta; \zeta_n)$.

(ii) At each step, we can evaluate the function twice and get $\ell(\theta_1; \zeta_n), \ell(\theta_2; \zeta_n)$.

(iii) For each step, we can evaluate the function m times and get $\ell(\theta_1; \zeta_n), \ldots, \ell(\theta_m; \zeta_n)$.

Gradient-free Optimization

(i) In this case, CLT rate is impossible. [2]

(ii) (iii) Those two cases are similar. They will be our main focus.

[2] Flaxman et al., SODA (2005). Agarwal et al., NeuIPS (2011).

Gradient-free Optimization

The simplest idea to solve (ii) (iii) is the Kiefer-Wolfowitz (KW) algorithm (finite difference).

We can use finite difference to approximate the gradient.

Given a direction v, a spacing parameter h, our two-point estimator of gradient is

$$\widehat{g}_{h,v}(\theta;\zeta) = \frac{1}{h}\Delta_{h,v}\ell(\theta;\zeta)v := \frac{\ell(\theta+hv;\zeta)-\ell(\theta;\zeta)}{h}v.$$

This is almost the same thing as $vv^\top\nabla\ell(\theta;\zeta)$.

Gradient-free Optimization

Given m direction $v_1, \ldots, v_m$, a spacing parameter h, our $m+1$-point estimator of gradient is

$$\widehat{g}_{h,v}(\theta;\zeta) = \sum_{i=1}^{m} \frac{1}{h} \Delta_{h,v_i} \ell(\theta;\zeta) v_i.$$

Gradient-free Optimization

This finite difference idea does not work if we can only evaluate once per iteration.

We need to specify two hyperparameters: the spacing parameter h and the distribution of the finite difference direction v.

Choosing h is easy. We can use $h_n = h_0 n^{-\gamma}, \frac{1}{2} < \gamma < 1$.

Choosing v is tricky. Obvious requirement: $\mathbb{E}v = 0, \mathbb{E}vv^\top = I_d$.

Gradient-free Optimization

(G) Gaussian: $v \sim \mathcal{N}(0, I)$.

(S) Spherical: v is sampled from the uniform distribution on the sphere $\|v\|^2 = d$.

(I) Uniform in a coordinate basis: v is sampled uniformly from $\left\{\sqrt{d}\boldsymbol{e}_1, \sqrt{d}\boldsymbol{e}_2, \ldots, \sqrt{d}\boldsymbol{e}_d\right\}$, where $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ is the natural coordinate basis of $\mathbb{R}^d$.

Gradient-free Optimization

Multipoint:

Previous ones+i.i.d..

Uniform without replacement from the coordinate basis (at most $d + 1$ point evaluation).

Uniformly sample orthonormal basis (uniform in Stiefel manifold).

Content

The CLT

What would happen if we combine the KW estimators with Polyak-Juditsky averaging? What is the CLT for those estimators?

Recall CLT for usually SGD with averaging:

$$\sqrt{n}\ \left(\overline{\theta}_n - \theta^*\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, H^{-1} S H^{-1}\right), \qquad \text{as} \quad n \to \infty.$$

Theorem (Xi Chen, Z-L, He Li, Yichen Zhang (2021))

The CLT for KW with averaging is

$$\sqrt{n}\ \left(\overline{\theta}_n - \theta^*\right) \Rightarrow \mathcal{N}\left(\mathbf{0}, H^{-1} Q H^{-1}\right), \qquad \text{as} \quad n \to \infty,$$

where $Q = \mathbb{E}\left[vv^\top S vv^\top\right]$.

The CLT

(G) Gaussian: $Q^{(\mathtt{G})} = (2S + \mathrm{tr}(S) I_d)$.

(S) Spherical: $Q^{(\mathtt{S})} = \frac{d}{d+2} (2S + \mathrm{tr}(S) I_d)$.

(I) Uniform in a natural coordinate basis: $Q^{(\mathtt{I})} = d\,\mathrm{diag}(S)$.

The CLT

Easy observations: Gaussian is worse than spherical.

Other methods are generally incomparable. We cannot say for sure $Q^{(\mathtt{S})} \succeq Q^{(\mathtt{I})}$ or $Q^{(\mathtt{S})} \preceq Q^{(\mathtt{I})}$.

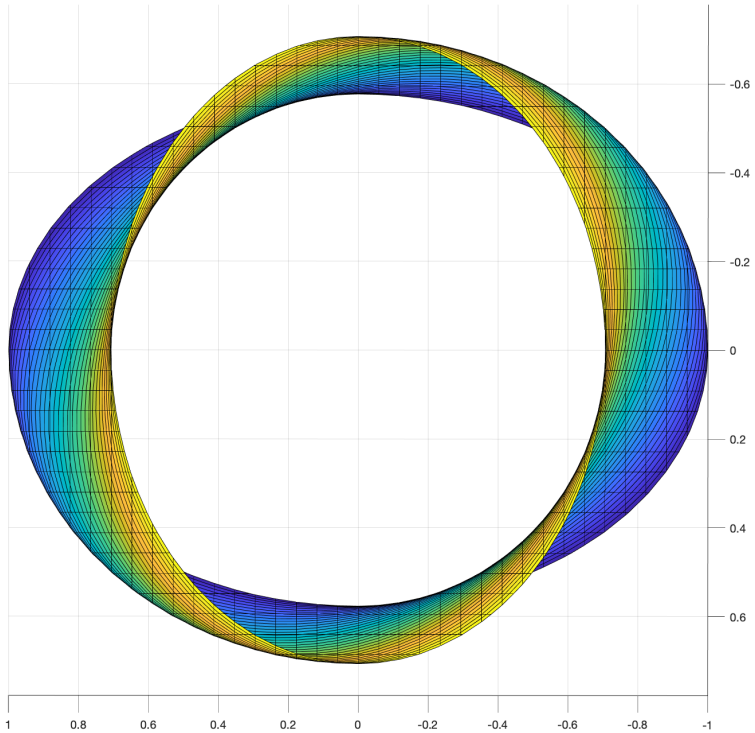

Figure: Comparison of Q matrices under different direction distributions

The CLT

Dimension factor appears.

Example: $S = I$. Spherical $Q^{(\mathrm{S})} = dI = dS$. Uniform in a natural coordinate basis: $Q^{(\mathrm{I})} = dI = dS$.

Zeroth order estimator is worse than the optimal first order estimator by a factor of d.

The CLT

This difference is not superfluous. The two point function evaluation gives us only partial ($\frac{1}{d}$) information of the gradient. We should not expect it behaves as good as the full gradient.

In fact, for any algorithm with two function evaluations, the convergence rate of $\|\hat{\theta}_n - \theta^*\|^2$ has a lower bound[3] $\Omega(\frac{d}{n})$. So the estimator is optimal asymptotically up to a constant factor.

[3] Duchi, John, et al. IEEE Transactions on Information Theory (2015).

The CLT

The rough bound indicates that there is a significant difference between zeroth order and first order optimization.

If we just want to make the variance in one direction e as small as possible. We can simply choose v to be in the direction e with high probability and in other direction with low probability. The variance in the direction e can be arbitrarily near the Cramer-Rao lower bound. But it will make the variance in other direction arbitrarily large and make the non-asymptotic convergence arbitrarily slow.

The CLT

We can derived the CLT $m+1$-point estimator. Again, the final form matches the rough lower bound derived by Duchi et al.

For the final two estimator (uniform without replacement/uniform in Stiefel manifold), $d+1$ point estimator achieves the exact lower bound of Cramer-Rao.

Online Statistical Inference

We can now implement one of the most important application of CLT: statistical inference.

More specifically, if there is a consistent estimator of $H^{-1}QH^{-1}$, then we can use the normal distribution to construct, for example, a 95% confidence inteval for θ^*.

Online Statistical Inference

How to estimate $H^{-1}QH^{-1}$? We can use same estimators as in SGD. There are many ways to do this:

One way is bootstraping[4]. We can simply use the trajectory of $\overline{\theta}_n$ to construct the covariance matrix of θ^*.

We can even prove a functional CLT: the trajectory converges to a Brownian motion. Thus, we can construct a "fixed-b" estimator based on the Brownian motion[5].

The "plug-in" method is to estimate H, Q separately.

[5] Zhu et al., JASA (2021).
[5] Lee et al., arXiv preprint (2021).

How to estimate H, the Hessian? Again, take a spacing parameter h, two random directions u, v, and then calculate the second-order difference.

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i^2} \Delta_{h_i,u_i} \Delta_{h_i,v_i} f(\theta_i;\zeta_i) u_i v_i^\top.$$

How to estimate Q, the variance of our gradient estimator? This is easier, we can simply calculate the empirical covariance.

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_{h_i,v_i}(\theta_i;\zeta_i)\widehat{g}_{h_i,v_i}(\theta_i;\zeta_i)^\top.$$

Theorem

Plug-in estimators are consistent. Furthermore, we have

$$\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \leq Cn^{-\alpha/2}.$$

Online Statistical Inference

Numerical experiments for linear and logistic regression:

100,000 samples. $d = 20$. Population design matrix $\Sigma = I$. Project to random direction to construct the 95% confidence inteval.

Online Statistical Inference

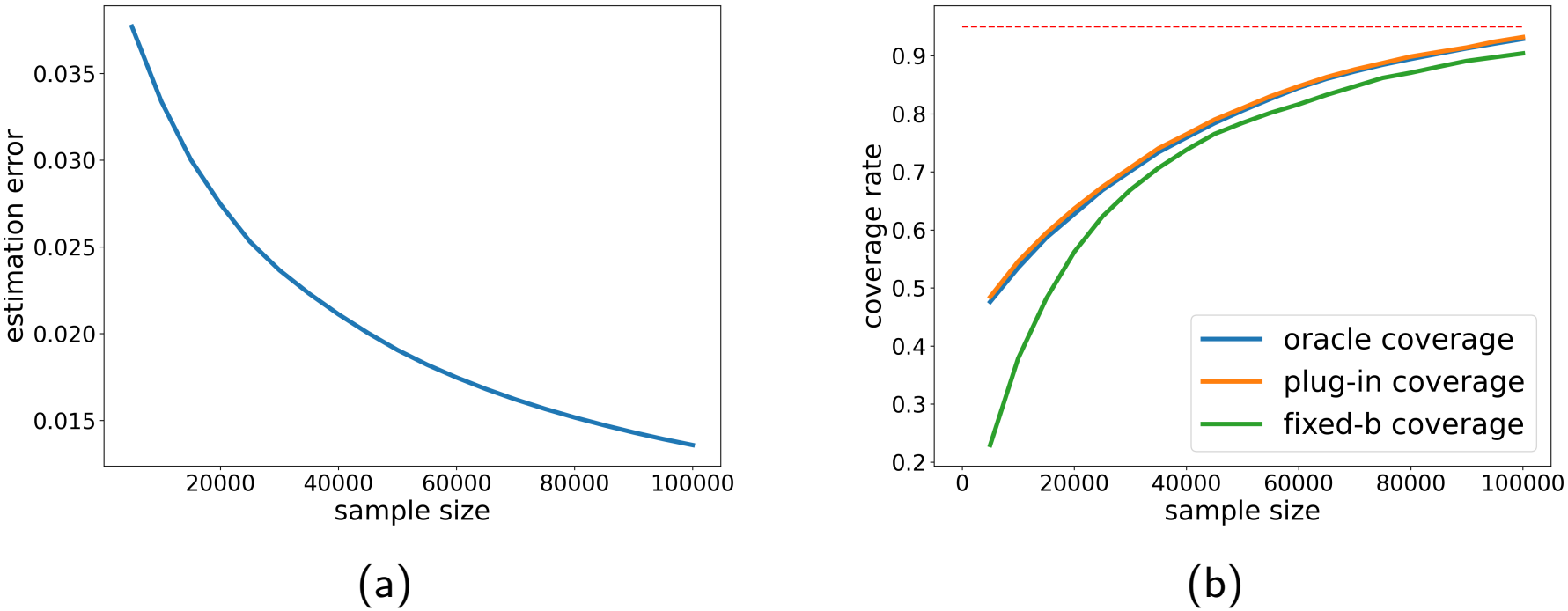

Figure: Convergence of the parameter estimation error $\|\overline{\theta}_n - \theta^\star\|$ and coverage rates v.s. the sample size n. Plots show the cases of linear regression. Dashed lines in plots (b) correspond to the nominal 95% coverage.

Content

Contextual Bandit Optimization

What is contextual bandit optimization?

The observed data at each decision point t is a triplet $\zeta_t = (X_t; A_t; Y_t(A_t))$, consisting of covariate X_t, action A_t, and reward $Y_t(A_t)$. We assume the contextual bandit environment is stochastic, in which $\{X_t, Y_t(a)\}$ is i.i.d. Here $Y_t(a)$ corresponds to the reward Y_t given a fixed action a regardless of the realized action A_t. $Y_t(a)$ is observed for $a = A_t$ only, but not observed for any other a.

Contextual Bandit Optimization

Now let us add some SGD flavor.

The model is parametrized by some parameter θ and loss function $l(\theta, \zeta)$.

$$\theta^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}_{\mathbb{P}_{Y|X}} \left[\ell\left(\theta; \zeta\right) \mid X, A = a \right].$$

Contextual Bandit Optimization

The SGD update with weight scheme:

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell(\theta_{t-1}; \zeta_t).$$

Actions A_t are selected according to some policy $A_t \sim \pi_t(\theta_{t-1}; X_t)$, which defines action distribution. Weights are also computed using previous iterates $w_t(\theta_{t-1}; A_t, X_t)$.

Contextual Bandit Optimization

Examples of policies:

Softmax: $\pi_t = softmax(Cs(\theta_{t-1}; X_t, A_t))$. s is some score function.

ϵ-greedy: $\mathbb{P}(A_t = a) = (1-\epsilon)1_{a \in argmax\{s(\theta_{t-1}; X_t, a, Y_t))\}} + \epsilon/|A|$.

Contextual Bandit Optimization

Examples of weights:

Constant: $\boldsymbol{w}_t = w$.

Inverse probability weighting: $\boldsymbol{w}_t = 1/\mathbb{P}(A_t|X_t, \theta_{t-1})$.

Contextual Bandit Optimization

A more concrete example - contextual least square linear regression:

$$s = \mathbb{E}[Y_t \mid A_t, X_t] = X_t^\top \theta^*_{A_t} = (1 - A_t)\big(X_t^\top \theta^*_{[1:p]}\big) + A_t\big(X_t^\top \theta^*_{[p+1:2p]}\big),$$

$$\ell(\theta; \zeta_t) = \tfrac{1}{2}(1 - A_t)\left(Y_t - X_t^\top \theta_{[1:p]}\right)^2 + \tfrac{1}{2} A_t \left(Y_t - X_t^\top \theta_{[p+1:2p]}\right)^2.$$

Content

CLT

What is difficulty of CLT in contextual bandit?

$$\theta_t = \theta_{t-1} - \eta_t w_t \nabla \ell(\theta_{t-1}; \zeta_t).$$

$w_t \nabla \ell(\theta_{t-1}; \zeta_t)$ is not the gradient of a loss function, we cannot apply previous result directly. Instead, we need to define the function:

$$L_{\theta'}(\theta) = \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\pi(X,\theta')} \left(w(\theta'; X, A) \ell(\theta; X, A, Y) \mid X \right) \right].$$

We must treat θ', θ as two separate variables and denote the partial derivative as

$$\nabla L_{\theta'}(\theta) := \frac{\partial}{\partial \theta} L_{\theta'}(\theta), \nabla^2 L_{\theta'}(\theta) := \frac{\partial^2}{\partial \theta^2} L_{\theta'}(\theta).$$

CLT

Now we can define the two matrix $H = \nabla^2 L_{\theta^*}(\theta^*)$ and S to be the covariance matrix of $w(\theta^*; X, A)\nabla l(\theta^*; X, A, Y)$.

Theorem (Xi Chen, Z-L, He Li, Yichen Zhang (2022))

We have the CLT

$$\sqrt{t}(\bar{\theta}_t - \theta^*) \to \mathcal{N}(0, H^{-1}SH^{-1}).$$

Furthermore, we can define consistent plug-in estimators of H, S and construct confidence intervals using CLT.

CLT

Further consideration: How to choose π and w to achieve a small asymptotic covariance matrices?

Corollary (Xi Chen, Z-L, He Li, Yichen Zhang (2022))

For contextual least square linear regression with any fixed π, $w_t = w$ achieves the smallest covariace matrices.

At least in this simple setting, there is no need to do any weighting scheme.

Statistical inference

Same. Just plug-in.

Summary

- CLTs in various machine learning settings are largely unexplored and a detailed analysis can often provide us new insights.
- Minimax bounds are much harder and much more interesting. Can we generalize the Hájek-Le Cam local asymptotic minimax theorem beyond the standard SGD setting?

Thank you