NORMAL APPROXIMATIONS FOR STOCHASTIC ITERATIVE ESTIMATORS (& MARTINGALES)

KRISHNA BALASUBRAMANIAN

DEPT OF STATISTICS, UC DAVIS

REFERENCES AND ACKNOWLEDGEMENT

- *Normal Approximation for Stochastic Gradient Descent via Non-Asymptotic Rates of Martingale CLT*.
  A. Anastasiou, **K. B.** and M. A. Erdogdu.
  Computational Learning Theory (COLT), 2019.

- *Unbiased Normal Approximation for Euler-Discretization of Itô Diffusions*.
  K. B., Hadi Daneshmand and M. A. Erdogdu.
  Under submission to Annals of Statistics, 2019.

OUTLINE

- ▷ Overview
- ▷ Results for linear systems
- ▷ Results for strongly convex setting
- ▷ Results for sampling

Overview

Research Agenda

Provide **algorithm-specific**, **non-asymptotically valid**, and **user-friendly** confidence sets/bands/intervals for parameter estimation and prediction using **stochastic iterative algorithms**.

Population version of M-Estimation (Stochastic Optimization):

$$\theta_* = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ f(\theta) = \mathsf{E}\,_Y[F(\theta, Y)] = \int F(\theta, Y)\, dP(Y) \right\}.$$

- ▷ $F(\theta, Y)$ is a random function depending on the random vector $Y \in \mathbb{R}^d$. Example: negative log-likelihood function.
- ▷ Goal: Provide confidence sets for estimating θ_* , to quantify uncertainty.

- Sample M-estimation (also ERM or Sample-Average Approximation): Get N i.i.d samples Y_i and compute

$$\widehat{\theta}_N = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{N} \sum_{i=1}^{N} F(\theta, Y_i) \right\}$$

- Asymptotic properties of $\widehat{\theta}_N$, in particular consistency and asymptotic normality, have been studied for decades in the math stat literature – useful for asymptotic uncertainty quantification.
- Drawback: Results are typically established for the actual minimizer $\widehat{\theta}_N$ and *not* the computational algorithm used.
- One still needs an algorithm to compute some $\tilde{\theta}_N$ that is close to $\widehat{\theta}_N$.

THE SGD ESTIMATOR

- Online/iterative approach:

  $$\theta_t = \theta_{t-1} - \eta_t g(\theta_t)$$

  where $g(\theta_t)$ is the *stochastic* gradient at iteration t, i.e., a rough estimate of $\nabla f(\theta_{t-1})$, typically based on one random sample Y_t.

- Polyak-Ruppert averaging: $\bar{\theta}_t = t^{-1}(\sum_{i=1}^{t} \theta_i)$

- What is known about this estimator:

  - When η_t is chosen properly, $\|\bar{\theta}_t - \theta_*\| \overset{\text{a.s.}}{=} 0$

  - Non-asymptotic rates on $\|\bar{\theta}_t - \theta_*\|$, (in expectation (most times) or high-probability (sometimes).

WHAT ABOUT OPTIMALITY AND INFERENCE ?

- Polyak and Juditsky (1992) and Ruppert (1998) showed that the following is true:

$$\sqrt{t}\left(\frac{1}{t}\sum_{i=1}^{t}\theta_i - \theta_*\right) \xrightarrow{\mathrm{d}} N(0, V)$$

  where $V = \nabla^2 f(\theta_*)^{-1}\textsc{cov}[\nabla f(\theta_*)]\nabla^2 f(\theta_*)^{-1}$.

- (Asymptotic) Martingale CLT plays a crucial role in proving the above result.

- Related results also by Roger Wets, Alex Shapiro, Vaclav Fabian, Kai Lai Chung, Jerome Sacks and few others.

- Some recent works: Toulis and Airoldi (2017); Duchi and Ruan (2018).

WHAT ABOUT PRACTICAL INFERENCE ?

- ▷ Variance V is asymptotically optimal (i.e., achieves Cramer-Rao lower bound) but requires knowledge of θ_*.

- ▷ Bootstrap style algorithms have been proposed recently by Fang et al. (2018); Su and Zhu (2018).

- ▷ What about the validity of confidence sets based on asymptotic normality ? – Justifiable only asymptotically. Similar to traditional normality or bootstrap results for M-estimators.

- ▷ This work: non-asymptotic uncertainty quantification.

INFORMAL RESULT

Informally speaking, we prove the following non-asymptotic normal approximation result in [ABE,'19]:

$$\sup_{A \in \mathcal{A}} |P(\bar{\theta}_t \in A) - P(Z \in A)| \leq C \frac{d^2}{\sqrt{t}}$$

where $\mathcal{A}$ is the set of all measurable convex subsets of $\mathbb{R}^d$ and $Z \sim N(0, I_d)$ and some constant $C > 0$.

Results for linear setting

- ▷ Consider a simpler problem: Given a positive definite matrix $A \in \mathbb{R}^{d \times d}$, find the θ_* which is the solution to the system $A\theta = b$.

- ▷ Stochastic Iterative update:

$$\theta_t = \theta_{t-1} - \eta_t y_t, \qquad y_t = A\theta_{t-1} - b + \zeta_t,$$

$$\bar{\theta}_t = \frac{1}{t} \sum_{i=0}^{t-1} \theta_i.$$

ASSUMPTION

- Let the noise ζ_t is a martingale difference sequence with

$$\mathsf{E}\left[\zeta_t\zeta_t^\top|\mathcal{F}_{t-1}\right] \overset{\text{a.s.}}{=} V \quad \text{and} \quad \mathsf{E}\left[\|\zeta_k\|_2^3\right] \leq \gamma d^{3/2}$$

- Assume that A and V are such that

$$\alpha I \preceq \left[A^{-1}VA^{-1}\right] \preceq \beta I.$$

Theorem [ABE19]: Let $\bar{\Delta}_t = \bar{\theta}_t - \theta_*$. If $\eta_t = \eta t^{-c_3}$, $c_3 \in (0,1)$, we have

$$\mathrm{E}\left[\left|h(\sqrt{t}\ \bar{\Delta}_t) - h(A^{-1}V^{1/2}Z)\right|\right]$$

$$\leq \frac{2.36\ \gamma\sqrt{\beta}}{\alpha^2} M_2(h) \frac{d^2}{\sqrt{t}} + K_4\ M_1(h)\ \sqrt{\frac{d}{t}} + K_5\ M_2(h)\ \frac{d}{t},$$

where K_4 & K_5 are some absolute constants, and $M_1(h)$ & $M_2(h)$ are constants depending only on function h.

- From the previous result, we can get Berry-Esseen bound by approximating indicator functions with carefully constructed twice-differentiable functions (following the idea of Bentkus (2003)) and obtain:

$$\sup_{A \in \mathcal{A}} \left| P(\bar{\Delta}_t \in A) - P(Z \in A) \right| \leq C \frac{d^2}{\sqrt{t}}$$

PROOF IDEA

- ▷ The following decomposition for $\bar{\Delta}_t$ is easy to obtain:

$$\sqrt{t}\bar{\Delta}_t = \underbrace{\frac{1}{\sqrt{t}\eta_0} B_t \Delta_0}_{I_1} + \underbrace{\frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} A^{-1} \zeta_j}_{I_2} + \underbrace{\frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} W_j^t \zeta_j}_{I_3},$$

where B_t and W_j^t are matrices that are functions of A and η_j.

- ▷ Using the triangle inequality,

$$\begin{aligned} &\left| \mathsf{E}\left[h(\sqrt{t}\bar{\Delta}_t) \right] - \mathsf{E}\left[h(A^{-1}V^{1/2}Z) \right] \right| \\ \leq &\underbrace{\left| \mathsf{E}\left[h(I_2) \right] - \mathsf{E}\left[h(A^{-1}V^{1/2}Z) \right] \right|}_{\textit{martingle clt}} + \left| \mathsf{E}\left[h\left(\sqrt{t}\bar{\Delta}_t\right) - h(I_2) \right] \right| \end{aligned}$$

Martingale Results

MULTIVARIATE MARTINGALE DEFINITIONS

- Let $X_1, X_2, ..., X_n \in \mathbb{R}^d$ be a martingale difference sequence adapted to a filtration $\mathcal{F}_0, \mathcal{F}_1, ..., \mathcal{F}_n$ with almost surely

$$\Sigma_k = \mathsf{E}\,[X_k X_k^\top | \mathcal{F}_{k-1}].$$

- Let $S_n = \sum_{i=1}^n X_i$, and the variance of the summation be $\Sigma_n = \textsc{var}(S_n)$.

- For $k \in \{1, \ldots, n\}$, partial covariance is $P_k = \sum_{i=k}^n \Sigma_i$.

Theorem [ABE19]: If we assume that $P_1 = \Sigma_n$ almost surely, then for $Z \sim N(0, I_d)$ and $h : \mathbb{R}^d \to \mathbb{R}$ a twice differentiable function, we have

$$\left| \mathsf{E}\, h(\Sigma_n^{-1/2} S_n) - \mathsf{E}\, h(Z) \right|$$

$$\leq \frac{3\pi}{8} \sqrt{d} M_2(h) \sum_{k=1}^{n} \mathsf{E}\, \| \Sigma_n^{1/2} P_k^{-1} \Sigma_n^{1/2} \|_2^{1/2} \| \Sigma_n^{-1/2} X_k \|_2^3.$$

Corollary [ABE19]: For a martingale difference sequence satisfying

$$\alpha I \preceq \Sigma_k \preceq \beta I$$

almost surely for all $k \in [n]$ and

$$\mathsf{E}\left[\|X_k\|_2^3\right] \leq \gamma d^{3/2},$$

we have

$$\left|\mathsf{E}\, h(\Sigma_n^{-1/2} S_n) - \mathsf{E}\, h(Z)\right| \leq \frac{3\pi\gamma\sqrt{\beta}}{4\alpha^2} M_2(h) \frac{d^2}{\sqrt{n}}.$$

- ▷ Proof technique is based on a combination of standard Lindeberg's telescoping sum along with Stein's method.

- ▷ The following two assumptions could be relaxed: (i) the eigenvalues of the conditional covariances Σ_k are bounded away from 0, i.e., $\Sigma_k \succeq \alpha I$, and (ii) the summation of the conditional covariances are deterministic, i.e., $P_1 \overset{\text{a.s.}}{=} \Sigma_n$.

Corollary [ABE19]: Assume there are constants β and δ such that almost surely

$$\mathsf{E}\left[\|X_k\|_2^3 | \mathcal{F}_{k-1}\right] \leq \beta \vee \delta \textsc{trace}(V_k)$$

then, we have

$$\left|\mathsf{E}\, h(\Sigma_n^{-1/2} S_n) - \mathsf{E}\, h(Z)\right| \leq 2\frac{M_1(h)}{\sqrt{n}} \textsc{trace}(\frac{1}{n}\Sigma_n)^{1/2}$$
$$+ \frac{3\pi}{4}\delta\sqrt{d} n M_2(h) \|\Sigma_n^{-1/2}\|_2^3 \left[\textsc{trace}(\tfrac{1}{n}\Sigma_n) + \beta^{2/3}\right]$$

Results in strongly convex setting

STOCHASTIC GRADIENT DESCENT

Statistical Estimation or Stochastic Optimization:

$$\theta_* = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ f(\theta) = \mathrm{E}_Y[F(\theta, Y)] = \int F(\theta, Y)\, dP(Y) \right\}.$$

SGD algorithm: Given $\theta_0 \in \mathbb{R}^d$

$$\theta_t = \theta_{t-1} - \eta_t g(\theta_t)$$

Here, ζ_t is mean-zero i.i.d random noise vector and

$$g(\theta_t) = \nabla f(\theta_{t-1}) + \zeta_t.$$

STRONGLY-CONVEX SETTING: DEFINITIONS

- ▷ $X_k := [\mathsf{E}\left[\nabla f(\theta_{k-1})\right] - \nabla f(\theta_{k-1}) - \zeta_k]$ is a martingale difference sequence. Let $\mathsf{E}\left[\|X_k\|_2^2 | \mathcal{F}_{k-1}\right] \leq K_d$.

- ▷ $\Sigma_t := \sum_{k=1}^{t} V_k$ where V_k corresponds to the covariance matrix of X_k:

$$V_k \overset{\text{a.s.}}{=} \mathsf{E}\left[X_k X_k^\top | \mathcal{F}_{k-1}\right], \qquad \forall k \in \{1, \dots, t\}.$$

- ▷ Define

$$\varrho(\eta, t) := \sum_{i=1}^{t-1} \left\{ e^{-2c_1 \sum\limits_{i=1}^{t-1} \eta_i} + \left[\frac{C'}{\eta_j^{1-c_2}} \sum_{i=j}^{t} m_j^t e^{-\lambda m_j^t} (m_j^i - m_j^{i-1}) \right]^2 \right\}$$

STRONGLY-CONVEX SETTING: MAIN RESULT

Theorem [ABE19]: For a twice differentiable function $h : \mathbb{R}^d \to \mathbb{R}$, we have the following non-asymptotic bound,

$$
\begin{aligned}
&\mathrm{E}\left[\left|h(\Sigma_t^{-1/2}\ \bar{\Delta}_t) - h(Z)\right|\right] \\
&\leq C\sqrt{d}\sum_{k=1}^{t} \mathrm{E}\left[\left\|\Sigma_t^{1/2} P_k^{-1} \Sigma_t^{1/2}\right\|_2^{1/2} \left([\nabla^2 f(\theta^*)]^{-1} \Sigma_t [\nabla^2 f(\theta^*)]^{-1}\right)^{-1/2} X_k\right\|_2^3\right] \\
&\quad + C_2 \frac{\|\Sigma_t^{-1/2}\|_2}{t}\left[\frac{[\mathrm{E}\,\|\Delta_0\|_2]}{\eta_0} + \frac{L_H \sum_{j=1}^{t-1}\sqrt{\eta_j}}{\sqrt{2\mu}} + \sqrt{K_d\ \varrho(\eta, t)}\right] \\
&\quad + \frac{C_3\|\Sigma_t^{-1/2}\|_2^2}{t^2}\left[\frac{\left[\mathrm{E}\,\|\Delta_0\|_2^2\right]}{\eta_0^2} + \frac{L_H^2 \sum_{j=1}^{t-1}\eta_j}{2\mu} + K_d\ \varrho(\eta, t)\right].
\end{aligned}
$$

STRONGLY-CONVEX SETTING: REMARKS

- ▷ Proof is similar to the linear setting (with an extra term to handle).

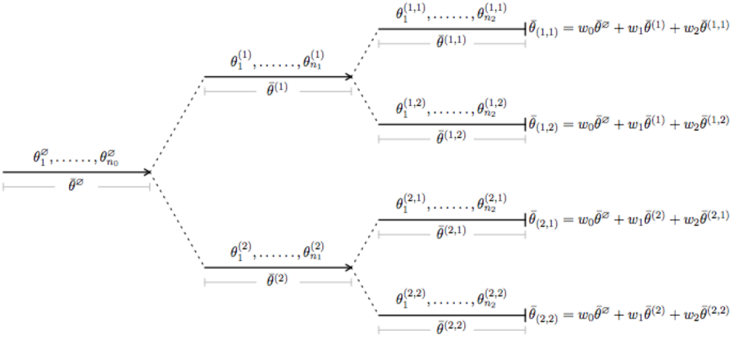
- ▷ Theorem is stated for any step-size sequence η_j.

- ▷ Typically, the dominant term is the first term, arising from the martingale rates.

- ▷ Similar to the linear setting, one can get Berry-Esseen bound by approximating indicators by twice-differentiable functions.

- ▷ Similar result could be derived for online bootstrap algorithms (Fang et al. (2018); Su and Zhu (2018)), to get practically computable and non-asymptotically valid confidence intervals for SGD.

Online bootstrap: HiGrad [Su and Zhu (2018)]



$\theta_1^{\varnothing}, \ldots\ldots, \theta_{n_0}^{\varnothing}$
$\bar{\theta}^{\varnothing}$
$\theta_1^{(1)}, \ldots\ldots, \theta_{n_1}^{(1)}$
$\bar{\theta}^{(1)}$
$\theta_1^{(2)}, \ldots\ldots, \theta_{n_1}^{(2)}$
$\bar{\theta}^{(2)}$
$\theta_1^{(1,1)}, \ldots\ldots, \theta_{n_2}^{(1,1)}$
$\bar{\theta}^{(1,1)}$
$\bar{\theta}_{(1,1)} = w_0\bar{\theta}^{\varnothing} + w_1\bar{\theta}^{(1)} + w_2\bar{\theta}^{(1,1)}$
$\theta_1^{(1,2)}, \ldots\ldots, \theta_{n_2}^{(1,2)}$
$\bar{\theta}^{(1,2)}$
$\bar{\theta}_{(1,2)} = w_0\bar{\theta}^{\varnothing} + w_1\bar{\theta}^{(1)} + w_2\bar{\theta}^{(1,2)}$
$\theta_1^{(2,1)}, \ldots\ldots, \theta_{n_2}^{(2,1)}$
$\bar{\theta}^{(2,1)}$
$\bar{\theta}_{(2,1)} = w_0\bar{\theta}^{\varnothing} + w_1\bar{\theta}^{(2)} + w_2\bar{\theta}^{(2,1)}$
$\theta_1^{(2,2)}, \ldots\ldots, \theta_{n_2}^{(2,2)}$
$\bar{\theta}^{(2,2)}$
$\bar{\theta}_{(2,2)} = w_0\bar{\theta}^{\varnothing} + w_1\bar{\theta}^{(2)} + w_2\bar{\theta}^{(2,2)}$

Figure 3: Graphical illustration of the HiGrad algorithm. Here we have three levels and at the end of each level, each segment is split into two segments. Averages are obtained for each level and at each leaf a weighted average is calculated. The weights w_j are detailed in Section 2.3.

RESEARCH AGENDA

- Research agenda: Provide **algorithm-specific**, **non-asymptotically valid**, and **user-friendly** confidence sets/intervals for parameter estimation and prediction done via stochastic iterative algorithms.

- Some take-home messages (at least to me!):
  - **Asymptotics hides a lot**! Be non-asymptotic in your analyses.
  - SGD is not an algorithm to compute mle. **SGD is the estimator**.
  - Establish **non-asymptotic posterior results** (concentration or confidence bands) for the **actual sampling or approximate inference algorithm** used.

SUMMARY

- In this talk, we discussed
  - Non-asymptotic normality of SGD algorithm.
  - Non-asymptotic normality of Euler-Discretization of Itô diffusions.
- Ongoing work:
  - Establish results for other stochastic iterative estimators (both optimization and sampling).
  - Leverage structure to obtain faster rates: Often times we are interested in only specific functionals of the estimated parameter (e.g., max-norm of the SGD estimator). Ongoing work with Miles Lopes.