Calibration of Statistical Inference for Stochastic Gradient Descent with Infinite Variance

Wenhao Yang

Stanford University

2025.10.27

INFORMS Annual Meeting 2025, Atlanta

Collaborators

Jose Blanchet
Stanford University

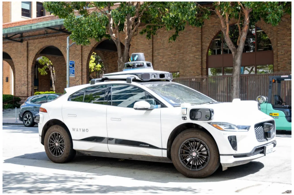
Peter Glynn
Stanford University

Aleksandar Mijatović
University of Warwick

- [BMY] Limit theorems for stochastic gradient descent with infinite variance.
- [BGY] Statistical Inference for the Stochastic Gradient Descent with Infinite Variance.

Machine Learning Today

AlphaGo Zero (cr. Shutterstock/maxuser)



Waymo (cr. Smith Collection/Gado)

ChatGPT

Why Machine Learning Succeeds?

Why Machine Learning Succeeds?

- Deep neural networks. [Vaswani, 2017, Krizhevsky et al., 2012]

Why Machine Learning Succeeds?

- Deep neural networks. [Vaswani, 2017, Krizhevsky et al., 2012]
- Stochastic training algorithms. [Robbins and Monro, 1951, Rumelhart et al., 1986, Duchi et al., 2011, Kingma and Ba, 2014]

Stochastic Gradient Descent (SGD)

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}_{\xi \sim P}[\ell(\theta, \xi)]$$

Stochastic Gradient Descent (SGD)

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}_{\xi \sim P}[\ell(\theta, \xi)]$$

- Stochastic Gradient Descent:

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

Stochastic Gradient Descent (SGD)

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}_{\xi \sim P}[\ell(\theta, \xi)]$$

- Stochastic Gradient Descent:

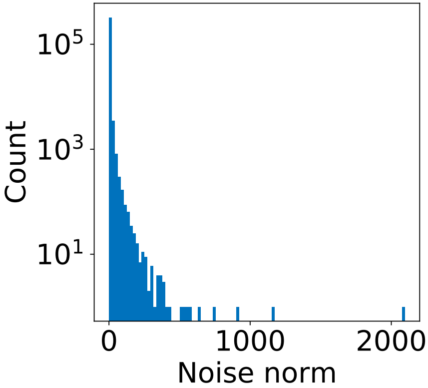
$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \xi_{n+1})$$

- Heavy-tail/infinite-variance stochastic gradient:

$$\mathbb{P}\left(\|\nabla \ell(\theta, \xi)\| > t\right) \sim t^{-\alpha}, \text{ with } \alpha \in (1, 2).$$

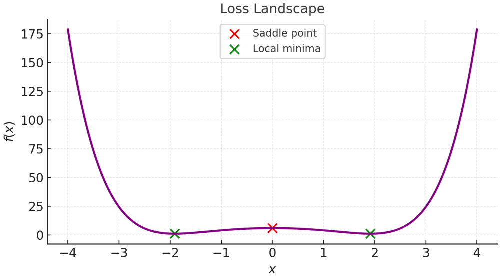
Heavy-tail Stochastic Gradient in Machine Learning

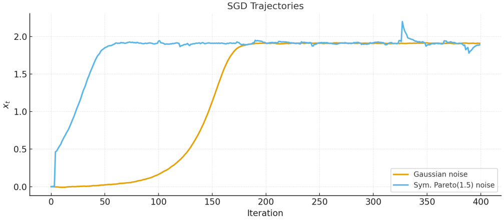
- The histogram of the norm of the gradient noises computed with AlexNet on Cifar10. [Simsekli et al., 2019].



Heavy-tail Benefits in Machine Learning

- Artificially injected heavy-tail noise.





Statistical Inference

- Several ways for uncertainty quantification: $\theta^* \in \mathsf{CI}_n$ with "high probability":
  - Plug-in. [Chen et al., 2020]
  - Batch means.[Chen et al., 2020]
  - Random scaling.[Lee et al., 2022]
  - ...

Statistical Inference

- Several ways for uncertainty quantification: $\theta^* \in \mathrm{CI}_n$ with "high probability":
  - Plug-in. [Chen et al., 2020]
  - Batch means.[Chen et al., 2020]
  - Random scaling.[Lee et al., 2022]
  - ...

We need statistical inference methodologies for heavy-tail/infinite variance SGD.

Statistical Inference

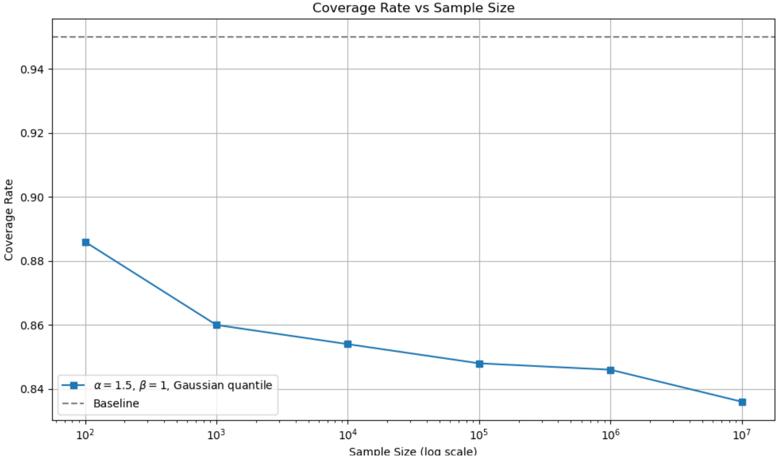
- Several ways for uncertainty quantification: $\theta^* \in \mathsf{CI}_n$ with "high probability":
  - Plug-in. [Chen et al., 2020]
  - Batch means.[Chen et al., 2020]
  - Random scaling.[Lee et al., 2022]
  - ...

We need statistical inference methodologies for heavy-tail/infinite variance SGD.

- What is the challenge?
  - Light-tail central limit theorem (CLT) no longer works.

Example: Apply Classic CLT

- Underlying model: $\nabla \ell(\theta, \xi)$ has infinite variance.
- Confidence interval (95%) for θ^* : $\left[\theta_n - q_1 \frac{\sigma_n}{\sqrt{n}}, \theta_n - q_2 \frac{\sigma_n}{\sqrt{n}}\right]$, Gaussian quantiles.



Challenges for Inference

We need calibrated statistical inference methodologies for heavy-tail SGD.

- Specifically:
  - Limit theorems for heavy-tail SGD.
  - Efficient Statistical inference approach.

Problem Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}[\ell(\theta, \xi)]$$

- SGD:

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

Problem Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}[\ell(\theta, \xi)]$$

- SGD:

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

- $\bar{\ell}(\theta)$ is strongly-convex.

Problem Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}[\ell(\theta, \xi)]$$

- SGD:

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

- $\bar{\ell}(\theta)$ is strongly-convex.
- $\eta_n \propto n^{-\rho}$, $\rho \in (\alpha^{-1}, 1]$.

Problem Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta) := \mathbb{E}[\ell(\theta, \xi)]$$

- SGD:

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

- $\bar{\ell}(\theta)$ is strongly-convex.
- $\eta_n \propto n^{-\rho}$, $\rho \in (\alpha^{-1}, 1]$.
- Heavy-tail assumption $(\alpha \in (1, 2))$: $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > t) = \frac{C}{t^{\alpha}}$

Problem Setup

- Objective:

$$\min_\theta \bar{\ell}(\theta) := \mathbb{E}[\ell(\theta, \xi)]$$

- SGD:

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

- $\bar{\ell}(\theta)$ is strongly-convex.
- $\eta_n \propto n^{-\rho}$, $\rho \in (\alpha^{-1}, 1]$.
- Heavy-tail assumption $(\alpha \in (1,2))$: $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > t) = \frac{C}{t^\alpha}$

Goal: Asymptotic behavior of $n^{?}(\theta_n - \theta^*)$.

Results

- SGD with heavy-tail(α) noise in 1-dimension and learning rate $\eta_n \propto n^{-1}$ [Krasulina 1969]:

$$\eta_n^{\frac{1}{\alpha}-1}(\theta_n - \theta^*) \xrightarrow{d} Z_\alpha.$$

- SGD with heavy-tail(α) noise in d-dimension and learning rate $\eta_n = c \cdot n^{-\rho}$ ($\rho \leq 1$):

Theorem [Blanchet, Mijiatović, and Yang 2024]

$$\eta_n^{\frac{1}{\alpha}-1}(\theta_n - \theta^*) \xrightarrow{d} Z_{\text{final},\rho}.$$

$Z_{\text{final},\rho}$ is the stationary distribution of an Ornstein-Uhlenbeck process driven by Lévy.

$$dX_t = -\left(\nabla^2 \ell(\theta^*) - \mathbb{1}(\rho = 1)\frac{1-\alpha^{-1}}{c}\right) X_t dt + dL_t^\alpha.$$

Results

- SGD with heavy-tail(α) noise in 1-dimension and learning rate $\eta_n \propto n^{-1}$ [Krasulina 1969]:

$$\eta_n^{\frac{1}{\alpha}-1}(\theta_n - \theta^*) \xrightarrow{d} Z_\alpha.$$

- SGD with heavy-tail(α) noise in d-dimension and learning rate $\eta_n = c \cdot n^{-\rho}$ ($\rho \leq 1$):

Theorem [Blanchet, Mijiatović, and Yang 2024]

$$\eta_n^{\frac{1}{\alpha}-1}(\theta_n - \theta^*) \xrightarrow{d} Z_{\mathsf{final},\rho}.$$

$Z_{\mathsf{final},\rho}$ is the stationary distribution of an Ornstein-Uhlenbeck process driven by Lévy.

$$dX_t = -\left(\nabla^2 \ell(\theta^*) - \mathbb{1}(\rho = 1)\frac{1-\alpha^{-1}}{c}\right) X_t dt + dL_t^\alpha.$$

- Fastest rate is achieved when $\rho = 1$, in 1-dimension, optimal constant $c^* = \frac{1}{\ell''(\theta^*)}$.

Polyak-Averaging SGD

- Replace θ_n with Polyak-averaging $\bar{\theta}_n = \frac{\sum_{i=1}^{n} \theta_i}{n}$ [Polyak et al. 1992]:

Theorem [Blanchet, Glynn, and Yang]

When $\eta_n \propto n^{-\rho}$ and $\rho \in (\alpha^{-1}, 1)$:

$$n^{1-\frac{1}{\alpha}} \left(\bar{\theta}_n - \theta^*\right) \xrightarrow{d} Z_{\text{avg}}.$$

- Comparison:

	Finite Variance	Infinite Variance
Z_{final}	High variance	High "scale"
Z_{avg}	Low variance	Low "scale"

Brief Sum-up

- Limit theorems for SGD with infinite variance still hold.

$$\eta_n^{\frac{1}{\alpha}-1}\left(\theta_n-\theta^*\right) \xrightarrow{d} Z_{\mathsf{final},\rho} \tag{1}$$

$$n^{1-\frac{1}{\alpha}}\left(\bar{\theta}_n-\theta^*\right) \xrightarrow{d} Z_{\mathsf{avg}}. \tag{2}$$

- Unknown parameters:
  - Index α.
  - Quantiles of limit distributions.

Self-normalization

Theorem [Blanchet, Glynn, and Yang]

When $\eta_n \propto n^{-\rho}$ with $\rho \in (\alpha^{-1}, 1)$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_i, \xi_i)\nabla\ell(\theta_i, \xi_i)^\top$:

$$\left(n^{1-\frac{1}{\alpha}}(\bar{\theta}_n - \theta^*), n^{\frac{1}{2}-\frac{1}{\alpha}}\sigma_n\right) \xrightarrow{d} (Z_{\mathsf{avg}}, W).$$

- Self-normalization:

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathrm{Tr}(\sigma_n^2)}} \xrightarrow{d} \frac{\|Z_{\mathsf{avg}}\|_\infty}{\sqrt{\mathrm{Tr}(W)}}.$$

- Benefits:
  - No estimation on α.

Self-normalization

Theorem [Blanchet, Glynn, and Yang]

When $\eta_n \propto n^{-\rho}$ with $\rho \in (\alpha^{-1}, 1)$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta_i, \xi_i)\nabla\ell(\theta_i, \xi_i)^\top$:

$$\left(n^{1-\frac{1}{\alpha}}(\bar{\theta}_n - \theta^*), n^{\frac{1}{2}-\frac{1}{\alpha}}\sigma_n\right) \xrightarrow{d} (Z_{\mathsf{avg}}, W).$$

- Self-normalization:

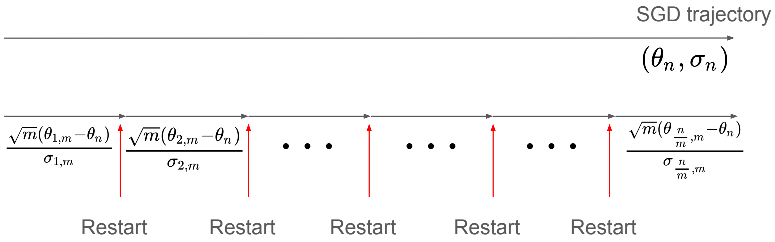
$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Tr}(\sigma_n^2)}} \xrightarrow{d} \frac{\|Z_{\mathsf{avg}}\|_\infty}{\sqrt{\mathsf{Tr}(W)}}.$$

- Benefits:
  - No estimation on α.
- Left to do: quantiles of $\|Z_{\mathsf{avg}}\|_\infty / \sqrt{\mathsf{Tr}(W)}$.

Sub-sampling for SGD

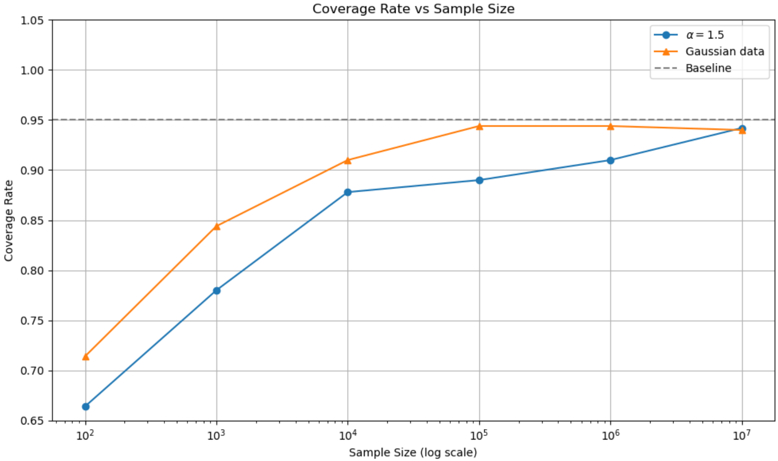
- Assume simulator for sampling.
- Sub-sample size: $m = \sqrt{n}$
- The approximate 95% confidence interval:

$$\left[\theta_n - \widehat{q}_{0.975}\frac{\sigma_n}{\sqrt{n}}, \theta_n - \widehat{q}_{0.025}\frac{\sigma_n}{\sqrt{n}}\right]$$

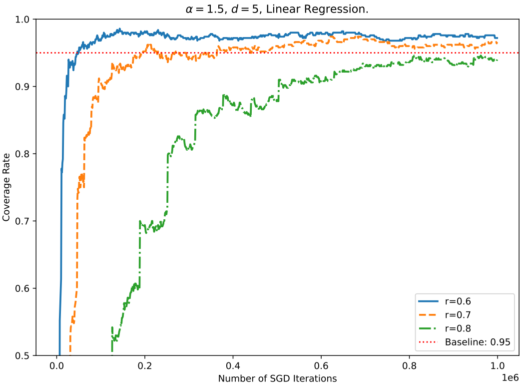


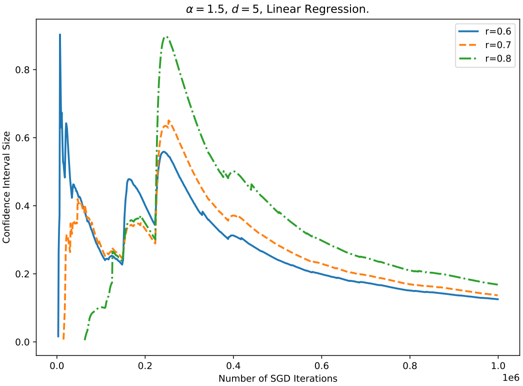
Simulation

- Blue: Sub-sampling + heavy tail, Orange: Sub-sampling + Gaussian data:



Simulation: Linear Regression





Takeaways

- Heavy-tailed SGD:
  - Weak convergence for final iterate and Polyak-Averaging.
  - Self-normalization + sub-sampling for inference.
- Application:
  - Stopping criteria: monitor the confidence interval.