Calibration of Statistical Inference for Stochastic Gradient Descent with Infinite Variance

Wenhao Yang

Stanford University

2025.8.7

JSM 2025, Nashville

Joint work with Jose Blanchet, Peter Glynn, Aleksandar Mijatović

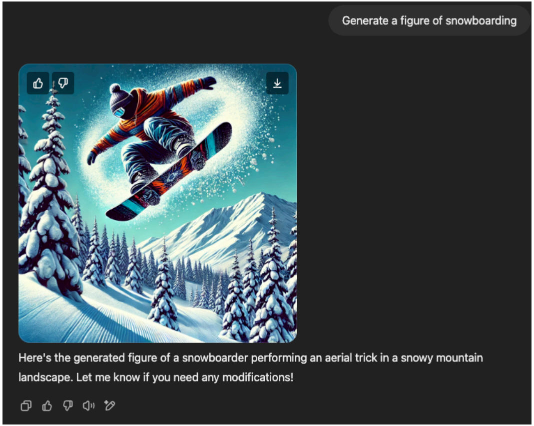
Machine Learning Today

Figure credit to Shutterstock/maxuser

Machine Learning Today

Figure credit to Smith Collection/Gado

Machine Learning Today



ChatGPT

Why Machine Learning Succeeds?

Why Machine Learning Succeeds?

- Deep neural networks. [Vas17, KSH12]

Why Machine Learning Succeeds?

- Deep neural networks. [Vas17, KSH12]
- Stochastic training algorithms. [RM51, RHW86, DHS11, KB14]

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:

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

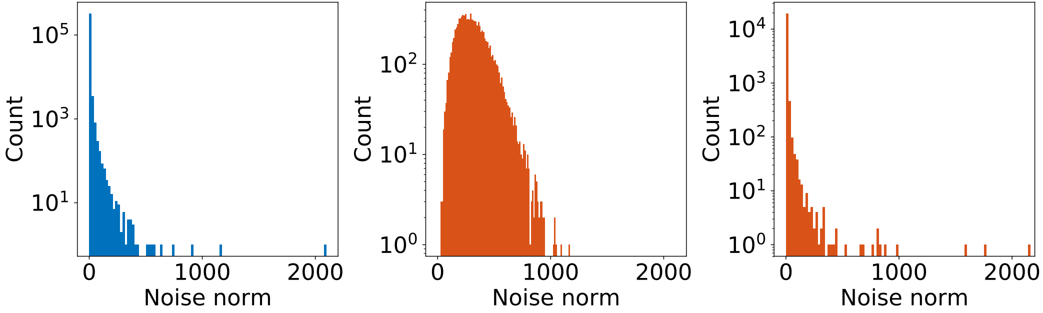
- Stochastic Gradient Descent:

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

- Several ways for uncertainty quantification: $\theta^* \in \mathsf{CI}_n$ with high probability:
    - Plug-in. [CLTZ20]
    - Batch means.[CLTZ20]
    - Random scaling.[LLSS22]
    - ...

Heavy-tail Data in Machine Learning

- The histogram of the norm of the gradient noises computed with AlexNet on Cifar10. [SSG19].



(a) Real (b) Gaussian (c) α-stable

A Challenge for Inference

We need statistical inference methodologies for heavy-tail SGD.

A Challenge for Inference

We need statistical inference methodologies for heavy-tail SGD.

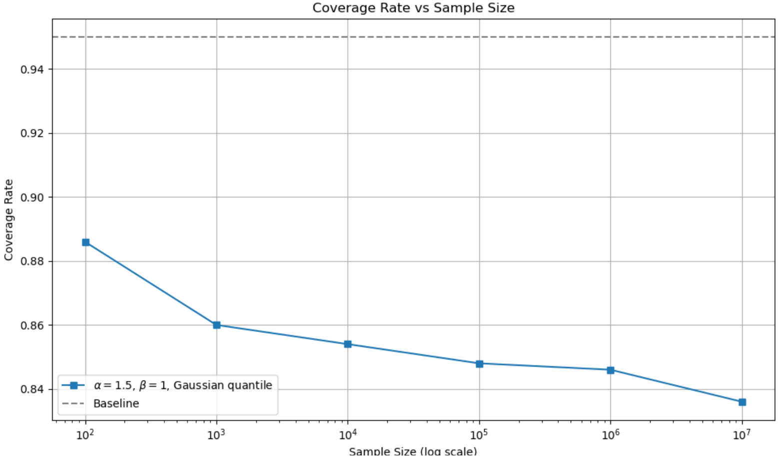
- What is the challenge?
  - Light-tail central limit theorem (CLT) no longer works.

Example: Apply Classic CLT

- Underlying model: $\nabla \ell(\theta, \xi) \sim \text{Stable}(\alpha, \beta)$. $(\alpha = 1.5, \beta = 1)$
- 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.

Example: Apply Classic CLT

- Underlying model: $\nabla \ell(\theta, \xi) \sim \text{Stable}(\alpha, \beta)$. $(\alpha = 1.5, \beta = 1)$
- 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.
- $\eta_n \propto n^{-\rho}$, $\rho \in (0, 1]$.

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.
- $\eta_n \propto n^{-\rho}$, $\rho \in (0, 1]$.
- Heavy-tail assumption $(\alpha \in (0, 1))$: $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > t) = \frac{C}{t^{\alpha}}$

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.
- $\eta_n \propto n^{-\rho}$, $\rho \in (0, 1]$.
- Heavy-tail assumption $(\alpha \in (0,1))$: $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > t) = \frac{C}{t^{\alpha}}$

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 [BMY24]

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

$Z_{\text{final},\rho}$ 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 [BMY24]

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

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

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

- 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 [BGY]

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

$$n^{1-\frac{1}{\alpha}}\left(\bar{\theta}_n - \theta^*\right) \xrightarrow{d} \widetilde{Z}_{\mathsf{avg}}.$$

- If symmetric, the tails of $Z_{\mathsf{final},\rho}$ and $\widetilde{Z}_{\mathsf{avg}}$ satisfy:

$$\mathbb{P}(|Z_{\mathsf{final},\rho}| \geq x) \approx \frac{C_1}{x^\alpha},\ \mathbb{P}(|\widetilde{Z}_{\mathsf{avg}}| \geq x) \approx \frac{C_2}{x^\alpha}$$

then $C_1 \geq C_2$. Equality is chosen when $c = c^*, \rho = 1$ for final iterate SGD.

Brief Sum-up

- Limit theorems for SGD with infinite variance still hold.

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

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

Self-normalization

- Inspiring from i.i.d. infinite variance mean-estimation [Logan et al. 1973]

Theorem [BGY]

When $\eta_n \propto n^{-1}$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \ell'(\theta_i, \xi_i)^2$:

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

- Self-normalization:

$$\frac{\sqrt{n}\,(\theta_n - \theta^*)}{\sigma_n} \xrightarrow{d} \frac{Z_{\mathsf{final}}}{W}.$$

- Benefits:
  - No estimation on α.

Self-normalization

- Inspiring from i.i.d. infinite variance mean-estimation [Logan et al. 1973]

Theorem [BGY]

When $\eta_n \propto n^{-1}$ and $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \ell'(\theta_i,\xi_i)^2$:

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

- Self-normalization:

$$\frac{\sqrt{n}\,(\theta_n - \theta^*)}{\sigma_n} \xrightarrow{d} \frac{Z_{\mathsf{final}}}{W}.$$

- Benefits:
  - No estimation on α.
- Left to do: quantiles of Z_{final}/W.

Sub-sampling for Mean Estimation [Romano et al. 1999]

- Self-normalization:

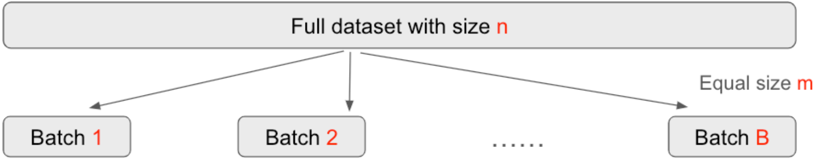
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma_n} \xrightarrow{d} \frac{Z}{W}$$

Sub-sampling for Mean Estimation [Romano et al. 1999]

- Self-normalization:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma_n} \xrightarrow{d} \frac{Z}{W}$$

- Sub-sampling:



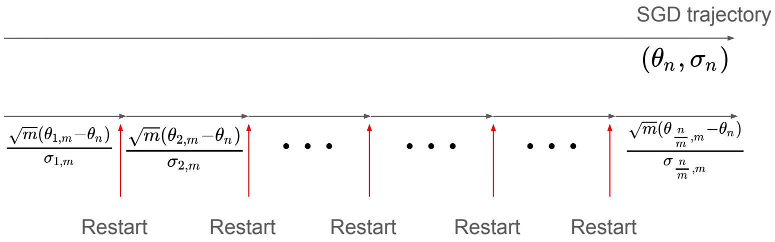
$$\frac{\sqrt{m}(\bar{X}_{1,m} - \bar{X}_n)}{\sigma_{1,m}} \qquad \frac{\sqrt{m}(\bar{X}_{2,m} - \bar{X}_n)}{\sigma_{2,m}} \qquad \frac{\sqrt{m}(\bar{X}_{B,m} - \bar{X}_n)}{\sigma_{B,m}}$$

$$\text{Empirical distribution} \approx \frac{Z}{W}$$

Sub-sampling for SGD

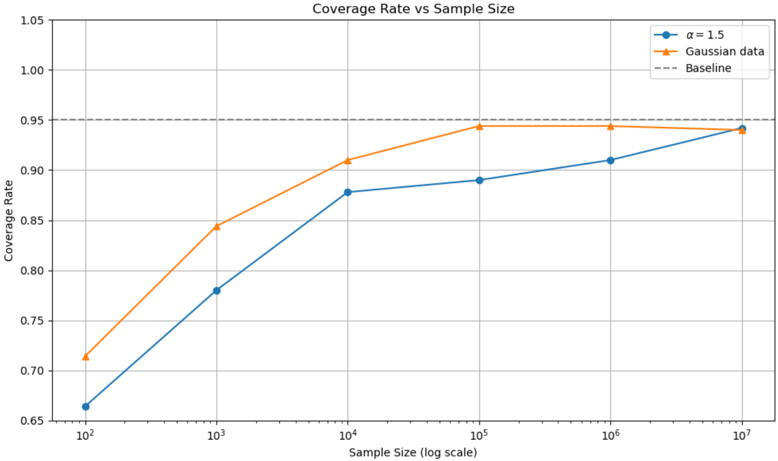
- Assume simulator for sampling.
- E.g. $m = \sqrt{n}$
- The approximate 95% confidence interval:

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



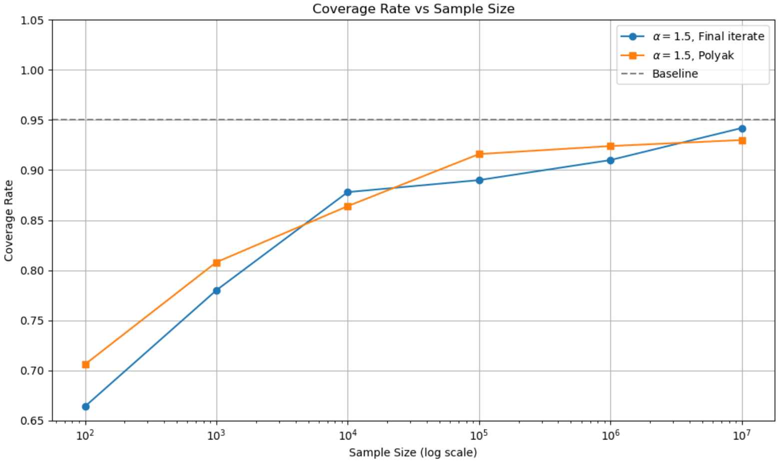
Simulation

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



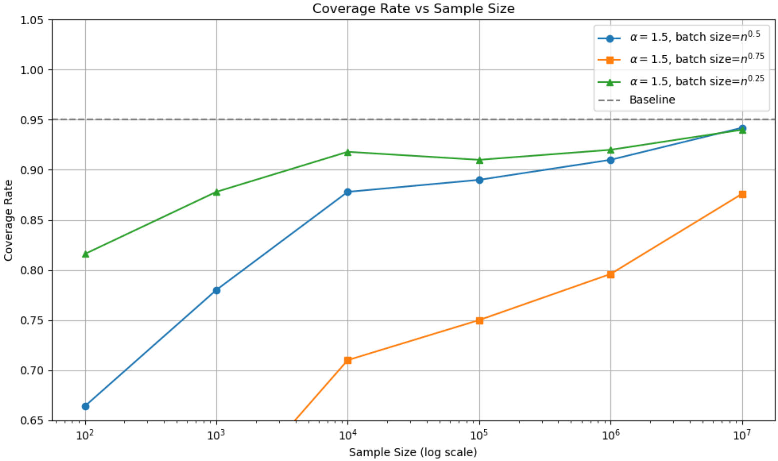
Simulation: Asymmetric

- Blue: Final iterate, Orange: Polyak-Averaging: $\rho = .7$



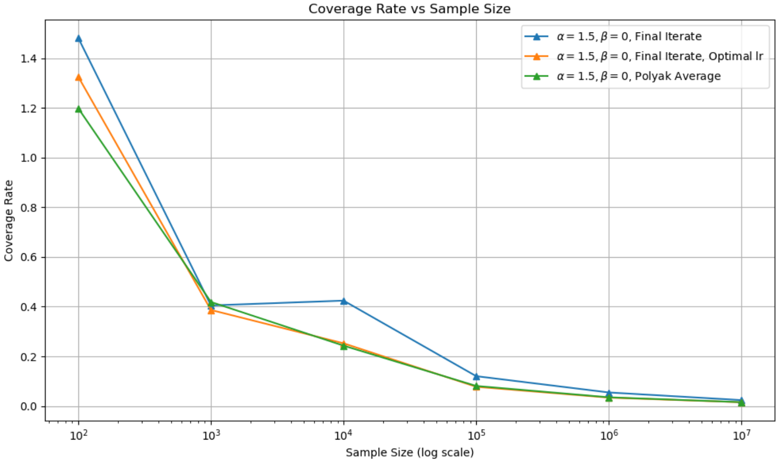
Simulation: Asymmetric

- Blue: batch size $n^{0.5}$, Orange: batch size $n^{0.75}$, Green: batch size $n^{0.25}$:



Simulation: Optimality comparison for SGD and Polyak-averaging

- Symmetric noise, Blue: Final iterate, Orange: Final iterate + Optimal learning rate, Green: Polyak-Averaging $\rho = .7$



Takeaways

- Heavy-tailed SGD [BMY24] [BGY]:
  - Weak convergence for final iterate and Polyak-Averaging.
  - Self-normalization + sub-sampling for inference.
- Application [BGY]:
  - Stopping criteria: monitor the confidence interval.

[BMY24] Jose Blanchet, Aleksandar Mijatović, and Wenhao Yang. Limit theorems for stochastic gradient descent with infinite variance. *arXiv preprint arXiv:2410.16340*, 2024.

[CLTZ20] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. 2020.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[LLSS22] Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7381–7389, 2022.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[SSG19] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[Vas17] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.