Inference for Stochastic Gradient Descent: Beyond Finite Variance

Wenhao Yang
Department of Management Science and Engineering
Stanford University

February 2, 2026
The Department of Statistics
University of Georgia

Collaborators

Jose Blanchet
Stanford University

Peter Glynn
Stanford University

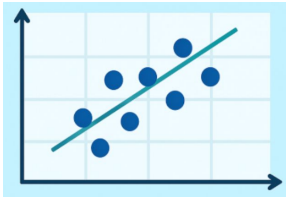
Aleksandar Mijatović
University of Warwick

- [BMY, 2024] Limit Theorems for Stochastic Gradient Descent with Infinite Variance.
- [BGY, Prep] Statistical Inference for Stochastic Gradient Descent with Infinite Variance.
- Key words: **Stochastic Gradient Descent**, **Inference**, **Infinite Variance**.

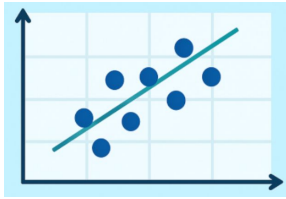
Background 1: Stochastic Gradient Descent (SGD)

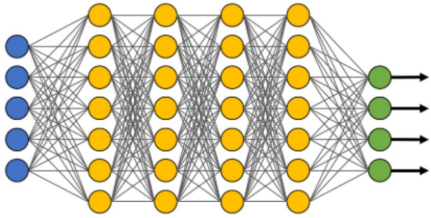
- SGD, Inference, Infinite Variance.

Applications of SGD

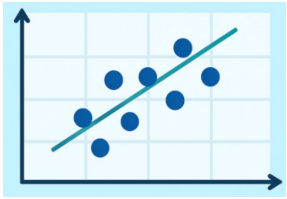


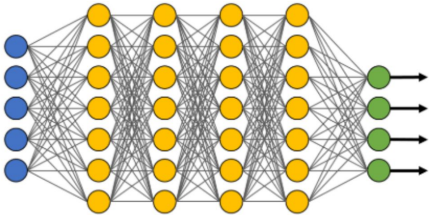
Applications of SGD

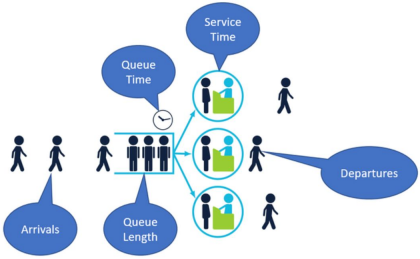




Applications of SGD







SGD setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

SGD setup

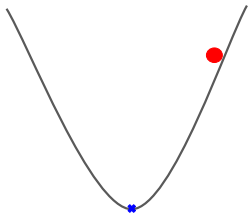
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



SGD setup

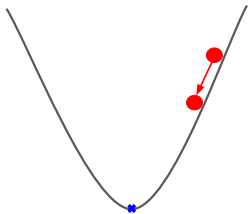
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



SGD setup

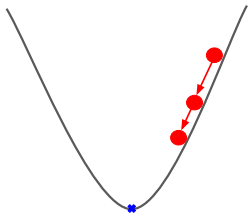
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



SGD setup

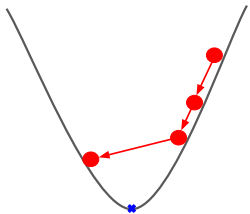
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



SGD setup

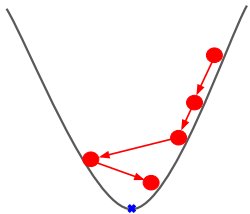
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

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

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



SGD setup

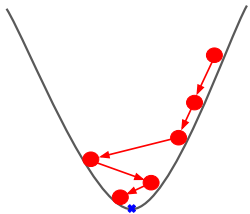
- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \;\; \theta^* := \arg\min_{\theta} \bar{\ell}(\theta).$$

- Stochastic Gradient Descent [Robbins and Monro, 1951]:

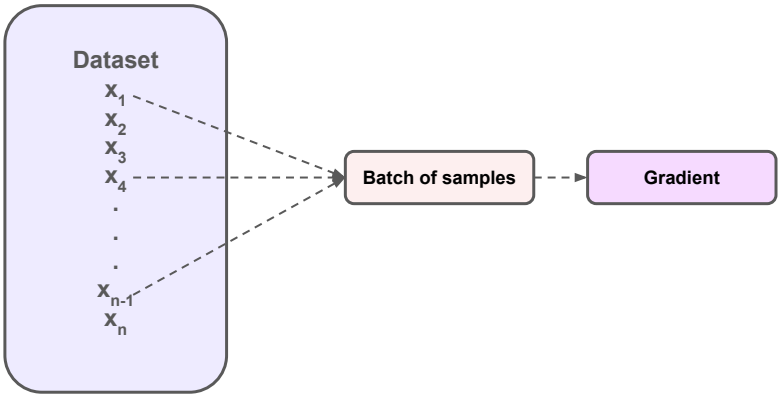
$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \xi_{n+1}),$$

with $\mathbb{E}[\nabla \ell(\theta, \xi)] = \nabla \bar{\ell}(\theta)$.



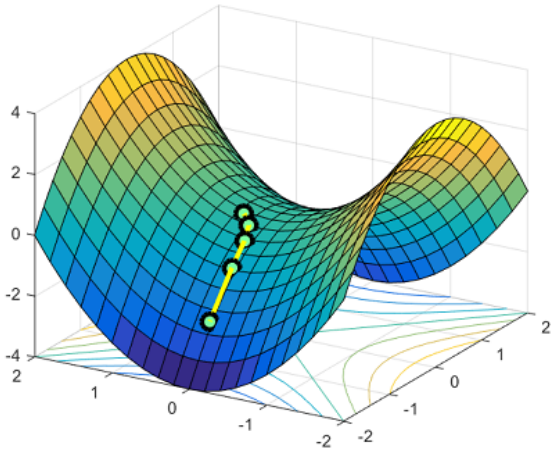
Benefits of SGD

- Computation efficiency: mini-batching [Bilmes et al., 1997].



Benefits of SGD

- Computation efficiency: mini-batching [Bilmes et al., 1997].
- Escaping Saddle points [Jin et al., 2017].



Background 2: Inference

- SGD, Inference, Infinite Variance.

With randomness, how to measure the accuracy of θ_n?

Uncertainty Quantification 1: Expectation Error

- Measurement:
  - Expectation error: $\mathbb{E}\|\theta_n - \theta^*\| \leq \varepsilon(n)$. E.g. $\varepsilon(n) = \frac{C}{\sqrt{n}}$.

Uncertainty Quantification 1: Expectation Error

- Measurement:
  - Expectation error: $\mathbb{E}\|\theta_n - \theta^*\| \leq \varepsilon(n)$. E.g. $\varepsilon(n) = \frac{C}{\sqrt{n}}$.
- Stopping criterion:

$$\text{Stop algorithm when } \varepsilon(n) \leq 0.01.$$

Uncertainty Quantification 1: Expectation Error

- Measurement:
  - Expectation error: $\mathbb{E}\|\theta_n - \theta^*\| \leq \varepsilon(n)$. E.g. $\varepsilon(n) = \frac{C}{\sqrt{n}}$.
- Stopping criterion:

$$\text{Stop algorithm when } \varepsilon(n) \leq 0.01.$$

- Issue: instance-dependent.

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$
$$\mathbb{P}\left(-1.96 \leq N(0,1) \leq 1.96\right) = 0.95.$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$

$$\mathbb{P}\left(-1.96 \leq \varepsilon(n)^{-1}(\theta_n - \theta^*) \leq 1.96\right) \approx 0.95.$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$

$$\mathbb{P}\left(-1.96 \leq \varepsilon(n)^{-1}(\theta_n - \theta^*) \leq 1.96\right) \approx 0.95.$$

$$\mathsf{CI}_n = \left[\theta_n - 1.96\varepsilon(n), \theta_n + 1.96\varepsilon(n)\right].$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$

$$\mathbb{P}\left(-1.96 \leq \varepsilon(n)^{-1}(\theta_n - \theta^*) \leq 1.96\right) \approx 0.95.$$

$$\mathsf{CI}_n = \left[\theta_n - 1.96\varepsilon(n), \theta_n + 1.96\varepsilon(n)\right].$$

- Stopping criterion:

$$\text{Stop algorithm when } |\mathsf{CI}_n| \leq 0.01.$$

Uncertainty Quantification 2: Confidence Interval

- Confidence Interval $\mathsf{CI}_n = [a_n, b_n]$:

$$\mathbb{P}\left(\theta^* \in \mathsf{CI}_n\right) \approx 0.95.$$

- A principal approach [Chen et al., 2020, Lee et al., 2022]: Central Limit Theorem

$$\varepsilon(n)^{-1}(\theta_n - \theta^*) \xrightarrow{d} N(0,1).$$

$$\mathbb{P}\left(-1.96 \leq \varepsilon(n)^{-1}(\theta_n - \theta^*) \leq 1.96\right) \approx 0.95.$$

$$\mathsf{CI}_n = \left[\theta_n - 1.96\varepsilon(n), \theta_n + 1.96\varepsilon(n)\right].$$

- Stopping criterion:

$$\text{Stop algorithm when } |\mathsf{CI}_n| \leq 0.01.$$

- Requirement: $\mathbb{E}\|\nabla \ell(\theta, \xi)\|^2 < +\infty$.

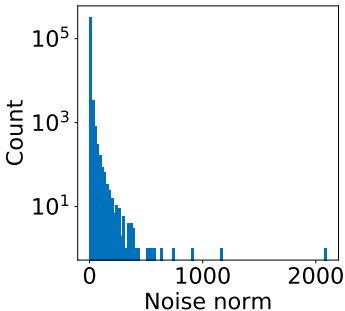
Background 3: Infinite Variance

- SGD, Inference, Infinite Variance.

In practice, is finite variance assumption $\mathbb{E}\|\nabla \ell(\theta, \xi)\|^2 < +\infty$ reasonable?

Empirical Observation

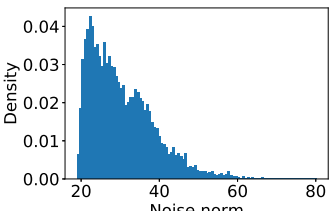
- The histogram of the norm of the gradient noises computed with AlexNet on Cifar10. [Simsekli et al., 2019].



Decaying rate $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > x) \sim x^{-1.1}$.
Implying $\mathbb{E}\|\nabla \ell(\theta, \xi)\|^2 = +\infty$.

Empirical Observation

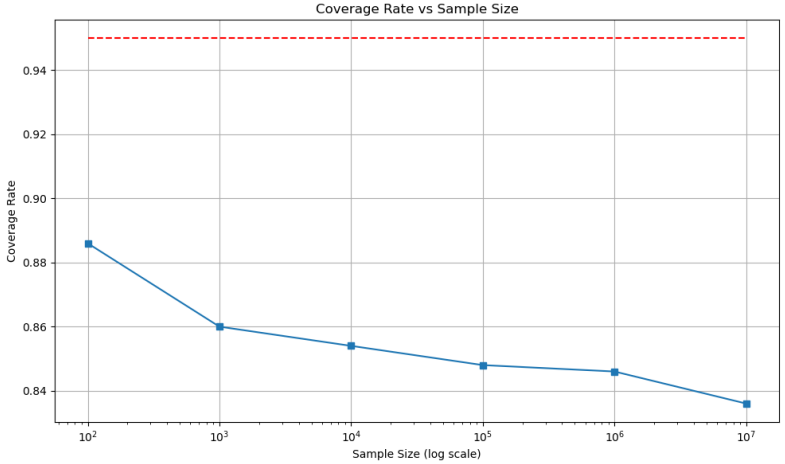
- The histogram of the norm of the gradient noises computed with Bert on Wiki. [Zhang et al., 2020].



Decaying rate $\mathbb{P}(\|\nabla \ell(\theta, \xi)\| > x) \sim x^{-1.08}$.
Implying $\mathbb{E}\|\nabla \ell(\theta, \xi)\|^2 = +\infty$.

Danger of Wrong Assumption

- Underlying model: $\nabla \ell(\theta, \xi)$ has infinite variance.
- 95% confidence interval CI_n for θ^* via finite variance approach.



Our Contributions

We need statistical inference methodologies for infinite variance SGD.

- Specifically:
  - Limit theorems for infinite variance SGD.

    Theorem [Informal]: $\eta_n^{\frac{1}{\alpha}-1} \cdot \widetilde{\mathcal{O}}(1) \cdot (\theta_n - \theta^*) \xrightarrow{d}$ Stable distribution.

  - Efficient inference approach to CI_n.
    - Techs: self-normalization + sub-sampling.
    - Model-agnostic: works for both finite variance and infinite variance.

SGD Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \text{ strongly convex, e.g. } \bar{\ell}(\theta) = \frac{1}{2}\|\theta\|^2.$$

- SGD:

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

SGD Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \text{ strongly convex, e.g. } \bar{\ell}(\theta) = \frac{1}{2}\|\theta\|^2.$$

- SGD:

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

- Assume $\nabla \ell(\theta, \xi)$ follows:

$$\mathbb{P}\left(\|\nabla \ell(\theta, \xi)\| > t\right) = \frac{\widetilde{\mathcal{O}}(1)}{t^{\alpha}}, \ \alpha \in (1, 2),$$

where $\widetilde{\mathcal{O}}(1)$ hide logarithm and constant factors.

SGD Setup

- Objective:

$$\min_{\theta} \bar{\ell}(\theta), \text{ strongly convex, e.g. } \bar{\ell}(\theta) = \frac{1}{2}\|\theta\|^2.$$

- SGD:

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

- Assume $\nabla \ell(\theta, \xi)$ follows:

$$\mathbb{P}\left(\|\nabla \ell(\theta, \xi)\| > t\right) = \frac{\widetilde{\mathcal{O}}(1)}{t^{\alpha}}, \ \alpha \in (1, 2),$$

where $\widetilde{\mathcal{O}}(1)$ hide logarithm and constant factors.

Goal: Asymptotic behavior of $\varepsilon(n)^{-1}(\theta_n - \theta^*)$.

Intuition: Stochastic Dynamical System

- Gradient descent to Gradient flow:

$$d\widetilde{\theta}_t = -\nabla \bar{\ell}(\widetilde{\theta}_t)\, dt.$$

Intuition: Stochastic Dynamical System

- Gradient descent to Gradient flow:

$$d\widetilde{\theta}_t = -\nabla \bar{\ell}(\widetilde{\theta}_t)\, dt.$$

- SGD to Stochastic gradient flow,

$$d\widetilde{\theta}_t = -\nabla \bar{\ell}(\widetilde{\theta}_t)\, dt + dL_t^{\alpha}.$$

Intuition: Stochastic Dynamical System

- Gradient descent to Gradient flow:

$$d\widetilde{\theta}_t = -\nabla \bar{\ell}(\widetilde{\theta}_t)\, dt.$$

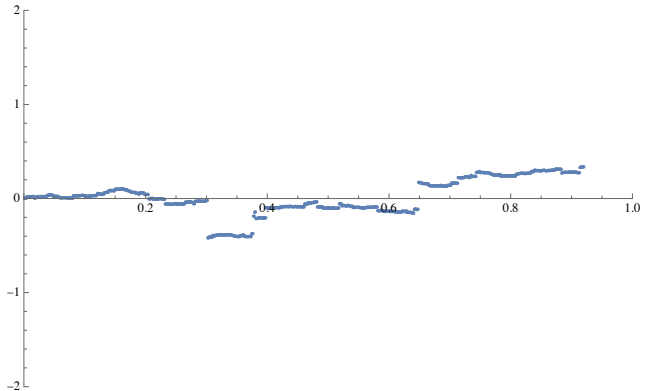
- SGD to Stochastic gradient flow,

$$d\widetilde{\theta}_t = -\nabla \bar{\ell}(\widetilde{\theta}_t)\, dt + dL_t^{\alpha}.$$

- L_t^{α} is an α-stable process.

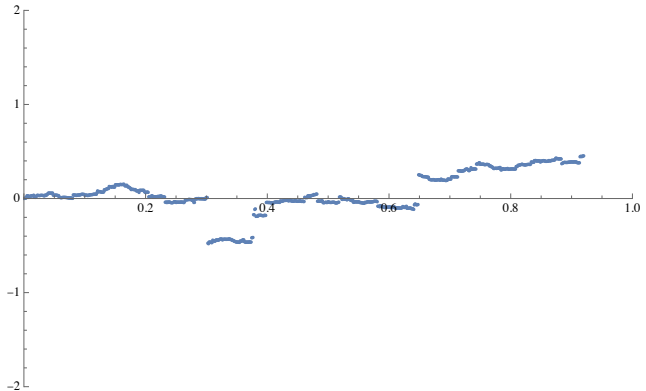
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



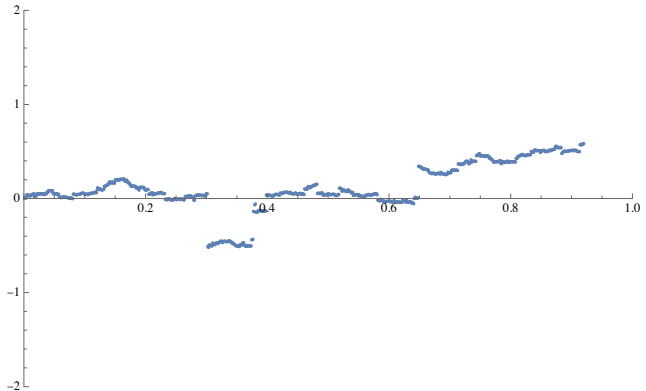
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



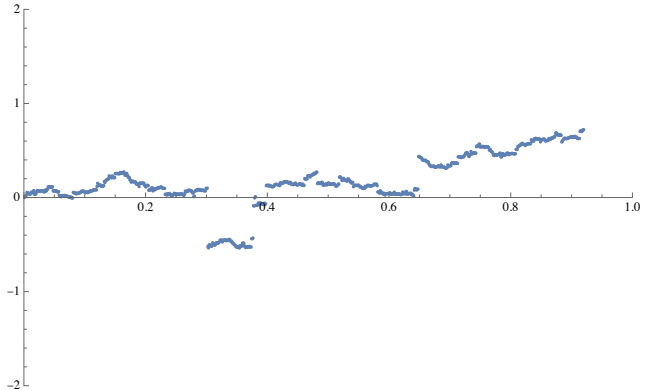
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



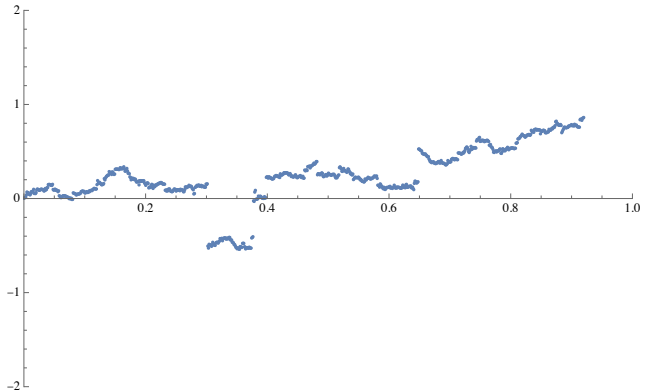
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



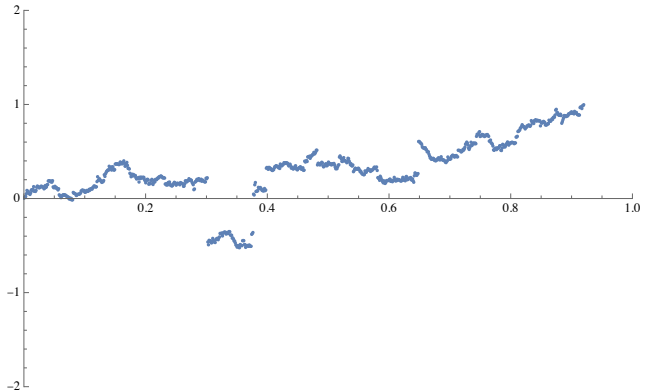
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



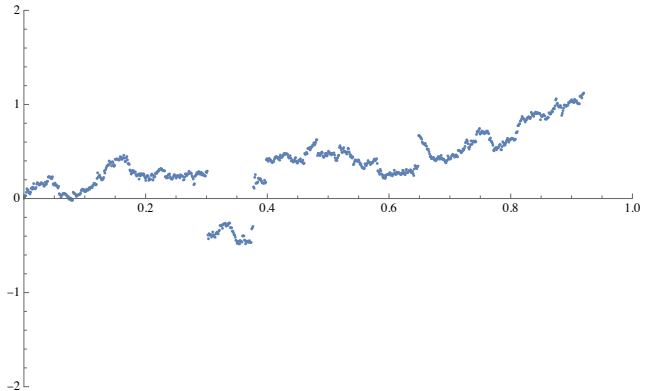
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



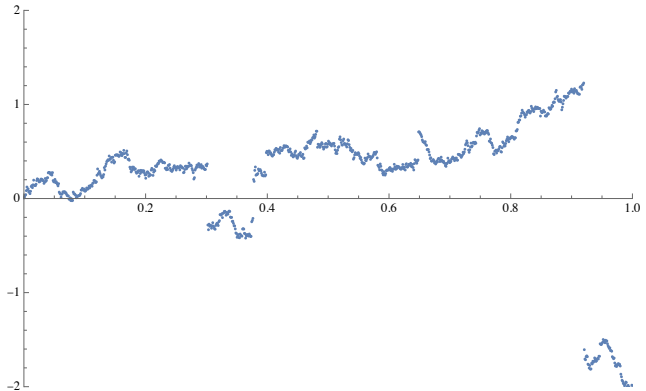
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



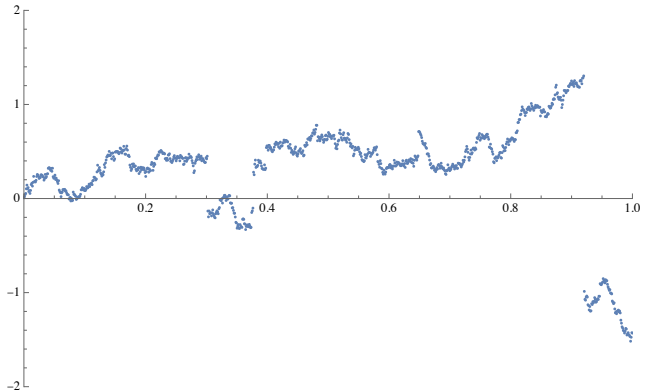
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



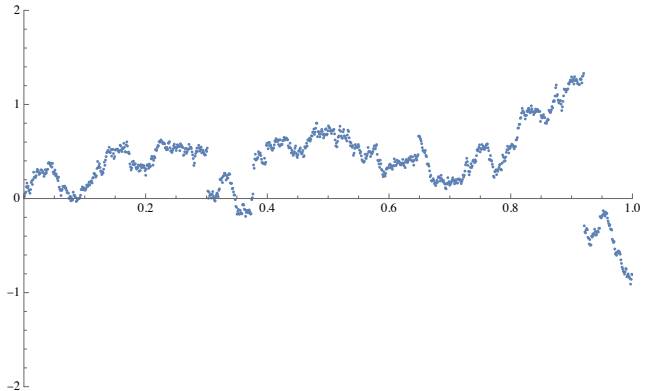
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



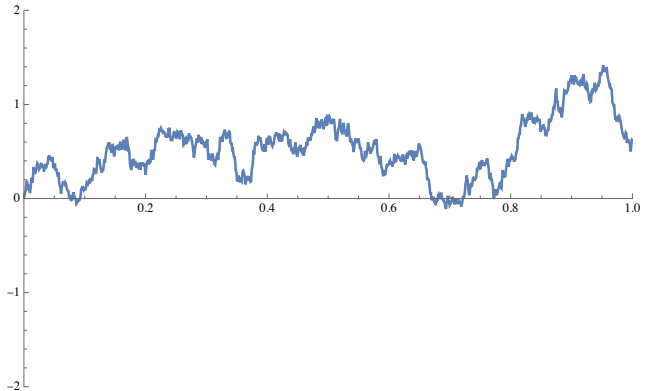
Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



Stable process

https://demonstrations.wolfram.com/StableLevyProcess/



Our Result

Theorem [BMY, 2024]

For $\eta_n = c \cdot n^{-\rho}$ ($\rho \in (\alpha^{-1}, 1]$), we have:

$$\underbrace{\eta_n^{\frac{1}{\alpha}-1}}_{\text{dominate}} \cdot \underbrace{\widetilde{\mathcal{O}}(1)}_{\text{logrithm}} \cdot (\theta_n - \theta^*) \xrightarrow{d} Z_{\text{Final}},$$

where Z_{Final} satisfies:

$$Z_{\text{Final}} \overset{d}{=} \int_0^{+\infty} \exp\left[\left(-\nabla^2 \ell(\theta^*) + \mathbb{1}(\rho = 1)\frac{1-\alpha^{-1}}{c}\right) t\right] dL_t^\alpha.$$

L_t^α is an α-stable process.

- Tool: Generalized central limit theorem [Gnedenko and Kolmogorov, 1968].
- Prior: 1-dim and $\widetilde{\mathcal{O}}(1) = \text{const.}$ [Krasulina, 1969]

Our Result

Theorem [BMY, 2024]

For $\eta_n = c \cdot n^{-\rho}$ ($\rho \in (\alpha^{-1}, 1]$), we have:

$$\underbrace{\eta_n^{\frac{1}{\alpha}-1}}_{\text{dominate}} \cdot \underbrace{\widetilde{\mathcal{O}}(1)}_{\text{logrithm}} \cdot (\theta_n - \theta^*) \xrightarrow{d} Z_{\text{Final}},$$

where Z_{Final} satisfies:

$$Z_{\text{Final}} \overset{d}{=} \int_0^{+\infty} \exp\left[\left(-\nabla^2 \ell(\theta^*) + \mathbb{1}(\rho = 1)\frac{1-\alpha^{-1}}{c}\right) t\right] dL_t^\alpha.$$

L_t^α is an α-stable process.

- $\alpha = 2$ recovers finite variance case.
- Fastest rate is achieved when $\rho = 1$. But $c > \frac{1-\alpha^{-1}}{\sigma_{\min}(\nabla^2 \ell(\theta^*))}$.
- In practice, determining c is difficult...

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, Prep]

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

$$\underbrace{n^{1-\frac{1}{\alpha}}}_{\text{dominate}} \cdot \widetilde{\mathcal{O}}(1) \cdot \left(\bar{\theta}_n - \theta^*\right) \xrightarrow{d} Z_{\text{Polyak}} = \nabla^2 \ell(\theta^*)^{-1} L_1^{\alpha}.$$

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, Prep]

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

$$\underbrace{n^{1-\frac{1}{\alpha}}}_{\text{dominate}} \cdot \widetilde{\mathcal{O}}(1) \cdot \left(\bar{\theta}_n - \theta^*\right) \xrightarrow{d} Z_{\text{Polyak}} = \nabla^2 \ell(\theta^*)^{-1} L_1^{\alpha}.$$

- Benefit 1: no tuning on c.

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, Prep]

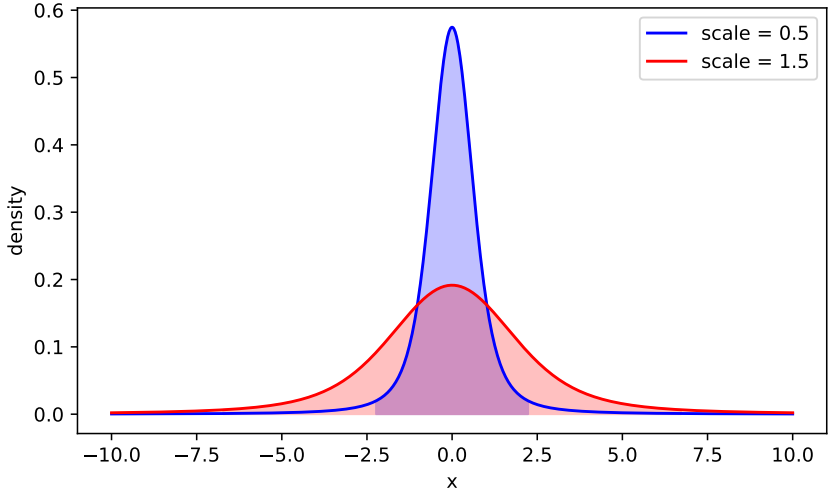
When $\eta_n \propto n^{-\rho}$ and $\rho \in (\alpha^{-1}, 1)$:

$$\underbrace{n^{1-\frac{1}{\alpha}}}_{\text{dominate}} \cdot \widetilde{\mathcal{O}}(1) \cdot \left(\bar{\theta}_n - \theta^*\right) \xrightarrow{d} Z_{\text{Polyak}} = \nabla^2 \ell(\theta^*)^{-1} L_1^{\alpha}.$$

- Benefit 1: no tuning on c.
- Benefit 2: $Z_{\text{Polyak}} \overset{d}{\preceq} Z_{\text{Final}}$. (Corollary [BGY, Prep])

Polyak-Averaging SGD: Benefit 2

- Benefit 2: $Z_{\text{Polyak}} \overset{d}{\preceq} Z_{\text{Final}}$. (Corollary [BGY, Prep])



Takeaways

- SGD with infinite variance [BMY, BGY]:
  - Distributions converge to non-degenerate but complex distributions.
    - Final iterate rate $\eta_n^{1-\frac{1}{\alpha}}$.
    - Polyak averaging rate $\left(\frac{1}{n}\right)^{1-\frac{1}{\alpha}}$.
  - Polyak averaging is still better in infinite variance case!
    - No tuning learning rate.
    - Tighter confidence size.

Confidence Interval

$$n^{1-\frac{1}{\alpha}} \cdot \widetilde{\mathcal{O}}(1) \cdot (\bar{\theta}_n - \theta^*) \xrightarrow{d} Z_{\text{Polyak}}$$

Confidence Interval

$$n^{1-\frac{1}{\alpha}} \cdot \widetilde{\mathcal{O}}(1) \cdot (\bar{\theta}_n - \theta^*) \xrightarrow{d} Z_{\text{Polyak}}$$

- Estimate α.
- Estimate $\widetilde{\mathcal{O}}(1)$.
- Estimate quantiles of Z_{Polyak}.

Can we do it more efficiently?

Step 1: Canceling Out Nuisance Parameters

- Core idea: "self-normalization".
- Example from i.i.d. mean estimation [Logan et al., 1973]:

$$\varepsilon(n)^{-1}\left(\bar{X}_n - \mathbb{E}[X_1]\right) \xrightarrow{d} S.$$

Step 1: Canceling Out Nuisance Parameters

- Core idea: "self-normalization".
- Example from i.i.d. mean estimation [Logan et al., 1973]:

$$\varepsilon(n)^{-1}\left(\bar{X}_n - \mathbb{E}[X_1]\right) \xrightarrow{d} S.$$

$$\varepsilon(n)^{-1}\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n^2}} \xrightarrow{d} W.$$

Step 1: Canceling Out Nuisance Parameters

- Core idea: "self-normalization".
- Example from i.i.d. mean estimation [Logan et al., 1973]:

$$\varepsilon(n)^{-1}\left(\bar{X}_n - \mathbb{E}[X_1]\right) \xrightarrow{d} S.$$

$$t\text{-statistic: } \frac{\sqrt{n}\left(\bar{X}_n - \mathbb{E}[X_1]\right)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n}}} \xrightarrow{d} \frac{S}{W}.$$

Self-normalization for SGD

Theorem [BGY, Prep]

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

$$\left(n^{1-\frac{1}{\alpha}} \cdot \underbrace{\widetilde{\mathcal{O}}(1)}_{\text{same}} \cdot (\bar{\theta}_n - \theta^*), n^{\frac{1}{2}-\frac{1}{\alpha}} \cdot \underbrace{\widetilde{\mathcal{O}}(1)}_{\text{same}} \cdot \sigma_n \right) \xrightarrow{d} (Z_{\text{Polyak}}, W).$$

- Self-normalization:

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\text{Trace}(\sigma_n^2)}} \xrightarrow{d} \frac{\|Z_{\text{Polyak}}\|_\infty}{\sqrt{\text{Trace}(WW^\top)}} := \textbf{SN}.$$

- Benefits:
  - No estimation on α and $\widetilde{\mathcal{O}}(1)$.
  - Working for finite variance. **SN** is different.

Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \mathbf{SN}.$$

- If **SN** is known:

Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \mathbf{SN}.$$

- If **SN** is known:

$$\mathbb{P}\left(\mathbf{SN} \leq q\right) = 0.95.$$

Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \textbf{SN}.$$

- If **SN** is known:

$$\mathbb{P}\left(\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \leq q\right) \approx 0.95.$$

Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \mathbf{SN}.$$

- If **SN** is known:

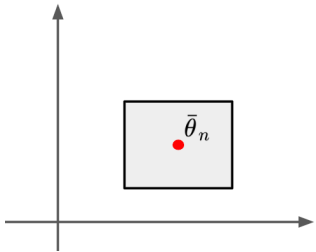
$$\mathbb{P}\left(\text{For each } i \in [d],\ \theta^{*,(i)} \in \left[\bar{\theta}_n^{(i)} - q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}, \bar{\theta}_n^{(i)} + q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}\right]\right) \approx 0.95.$$

Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \mathbf{SN}.$$

- If **SN** is known:

$$\mathbb{P}\left(\text{For each } i \in [d],\ \theta^{*,(i)} \in \left[\bar{\theta}_n^{(i)} - q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}, \bar{\theta}_n^{(i)} + q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}\right]\right) \approx 0.95.$$



Confidence Region

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \mathbf{SN}.$$

- If **SN** is known:

$$\mathbb{P}\left(\text{For each } i \in [d],\ \theta^{*,(i)} \in \left[\bar{\theta}_n^{(i)} - q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}, \bar{\theta}_n^{(i)} + q\sqrt{\frac{\mathsf{Trace}(\sigma_n^2)}{n}}\right]\right) \approx 0.95.$$

- Issue: q is unknown.

Step 2: Simulating the **SN**

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \textbf{SN}.$$

- Simulate **SN**:
  - Ideally, generate i.i.d. samples $T_1, \cdots, T_k \overset{d}{=} \textbf{SN}$.

Step 2: Simulating the **SN**

$$\frac{\sqrt{n}\left\|\bar{\theta}_n - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_n^2)}} \xrightarrow{d} \textbf{SN}.$$

- Simulate **SN**:
  - Ideally, generate i.i.d. samples $T_1, \cdots, T_k \overset{d}{=} \textbf{SN}$.
  - Sub-sampling [Romano and Wolf, 1999]: generate approximations $\widehat{T}_1, \cdots, \widehat{T}_k$ for $T_1, \cdots, T_k$:

$$\widehat{T}_i = \frac{\sqrt{n}\left\|\bar{\theta}_{n,i} - \theta^*\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_{n,i}^2)}}.$$

Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

Sub-sampling for SGD

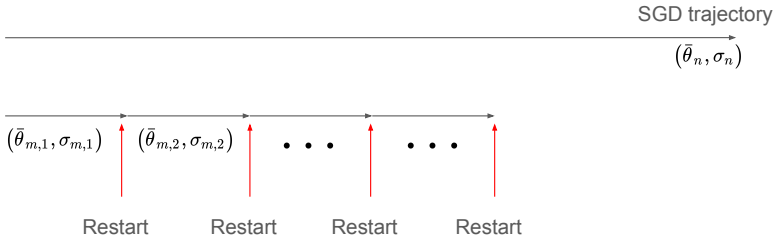
- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.

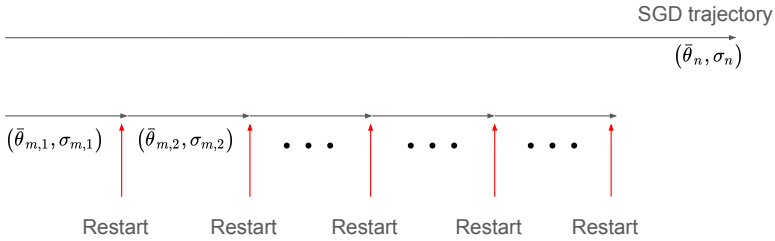
Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.



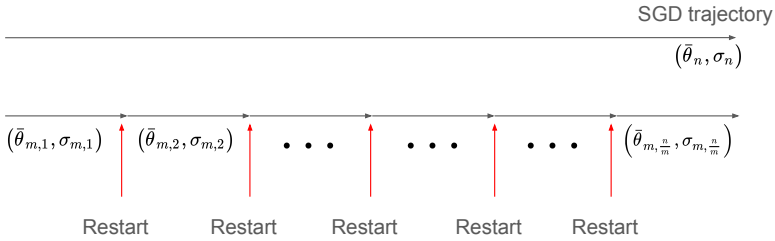
Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.



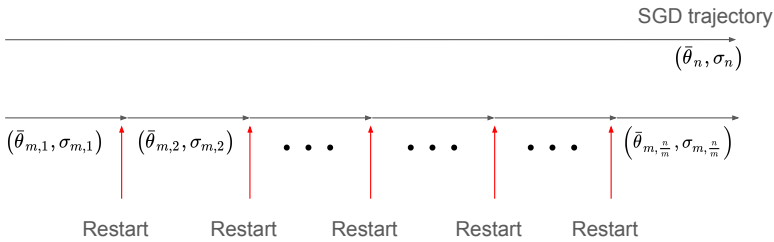
Sub-sampling for SGD

- Sub-sampling size $m = n^r$ with $r \in (0, 1)$.



Sub-sampling for SGD

- Main trajectory: $\left(\bar{\theta}_n, \sigma_n\right)$.
- Sub-sampling trajectory: $\left(\bar{\theta}_{m,1}, \sigma_{m,1}\right), \ldots, \left(\bar{\theta}_{m,\frac{n}{m}}, \sigma_{m,\frac{n}{m}}\right)$.



Sub-sampling for SGD

- Main trajectory: $\left(\bar{\theta}_n, \sigma_n\right)$.
- Sub-sampling trajectory: $\left(\bar{\theta}_{m,1}, \sigma_{m,1}\right), \ldots, \left(\bar{\theta}_{m,\frac{n}{m}}, \sigma_{m,\frac{n}{m}}\right)$.

Theorem [BGY, Prep]

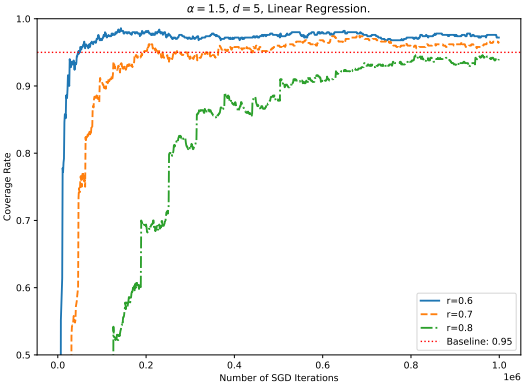
For self-normalized statistics ($k = 1, 2, \cdots, \frac{n}{m}$) and $m = n^r$ with $r \in (0, 1)$:

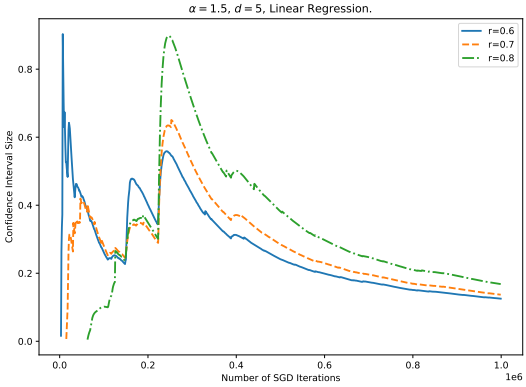
$$\widehat{T}_k := \frac{\sqrt{m}\left\|\bar{\theta}_{m,k} - \bar{\theta}_n\right\|_\infty}{\sqrt{\mathsf{Trace}(\sigma_{m,k})}},$$

they satisfy (for any $x \in \mathbb{R}$):

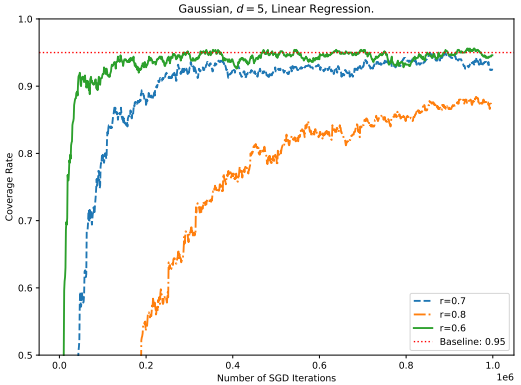
$$\frac{1}{n/m}\sum_{k=1}^{n/m} \mathbb{1}\left(\widehat{T}_k \le x\right) \xrightarrow{p} \mathbb{P}\left(\mathbf{SN} \le x\right).$$

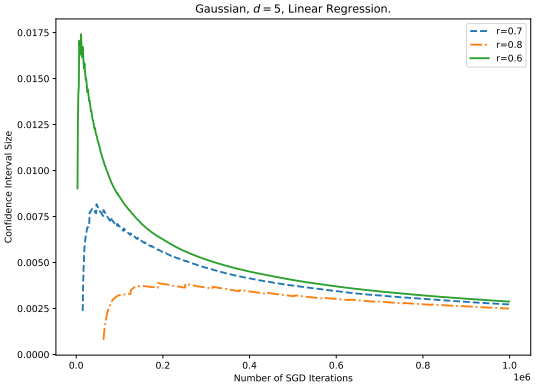
Simulation: Linear Regression with Infinite Variance





Simulation: Linear Regression with Finite Variance





- Optimal choice of $m = n^r$: trade-off between
  - r small, quantiles accurate, θ_m or $\bar{\theta}_m$ inaccurate.
  - r large, θ_m or $\bar{\theta}_m$ accurate, quantiles inaccurate.

Takeaways

- Statistical inference for infinite variance SGD [BGY]:
  - Leverage self-normalization and sub-sampling.

$$\mathsf{CI}_n = \left[\bar{\theta}_n - q\sqrt{\frac{\sigma_n^2}{n}}, \bar{\theta}_n + q\sqrt{\frac{\sigma_n^2}{n}}\right].$$

  - Model-agnostic.
- Open directions:
  - Optimal sub-sample size.
  - Momentum SGD, Adam, ...
  - Non-convex.

Jeff Bilmes, Krste Asanovic, Chee-Whye Chin, and Jim Demmel. Using phipac to speed error back-propagation learning. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 4153–4156. IEEE, 1997.

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

B. V. Gnedenko and A. N. Kolmogorov. *Limit distributions for sums of independent random variables*. Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills., Ont., 1968.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pages 1724–1732. PMLR, 2017.

Tatiana Pavlovna Krasulina. On stochastic approximation processes with infinite variance. *Theory of Probability & Its Applications*, 14(3):522–526, 1969.

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.

Benjamin F Logan, CL Mallows, SO Rice, and Larry A Shepp. Limit distributions of self-normalized sums. *The Annals of Probability*, 1(5):788–809, 1973.

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

Joseph P Romano and Michael Wolf. Subsampling inference for the mean in the heavy-tailed case. *Metrika*, 50:55–69, 1999.

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.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.