Limit theorems for stochastic gradient descent with infinite variance

Aleksandar Mijatović
Department of Statistics

Joint work with **Jose Blanchet** (Stanford) and **Wenhao Yang** (Peking)

Liverpool Discrete Mathematics Colloquium
University of Liverpool

9 September 2026

EPSRC
Engineering and Physical Sciences Research Council

EP/W006227/1 & EP/V009478/1

Structure of the talk

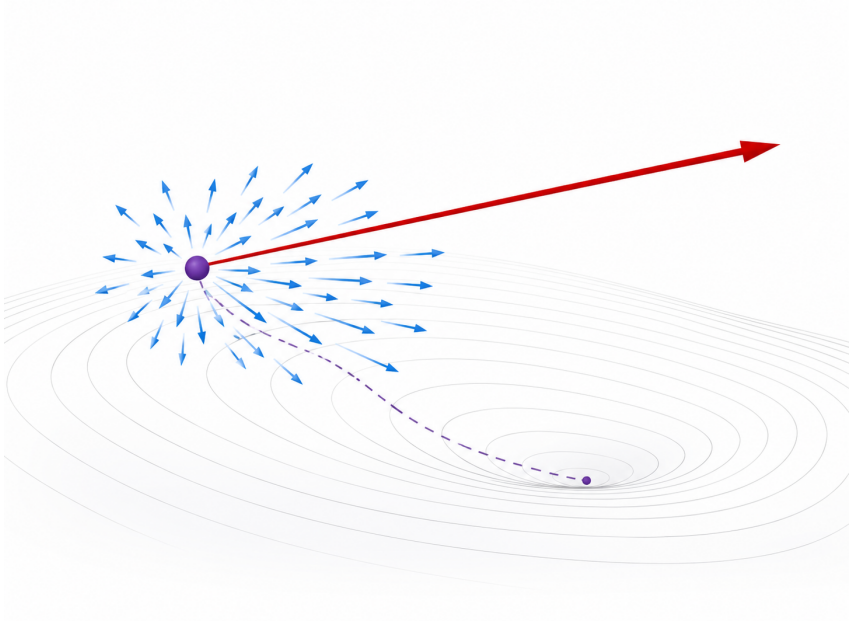
1. Introduction: why infinite variance changes SGD fluctuations
2. Main results: path and final-iterate limit theorems
3. Numerical examples: OLS and logistic regression
4. Ideas of Proof: point processes, time change, and tightness

A single extreme gradient can dominate an SGD step

Problem: find

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \ell(\theta) := \mathbb{E}_{\boldsymbol{\xi} \sim P}\big[\ell(\theta, \boldsymbol{\xi})\big],$$

where $\boldsymbol{\xi}$ random element and ℓ strongly convex



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

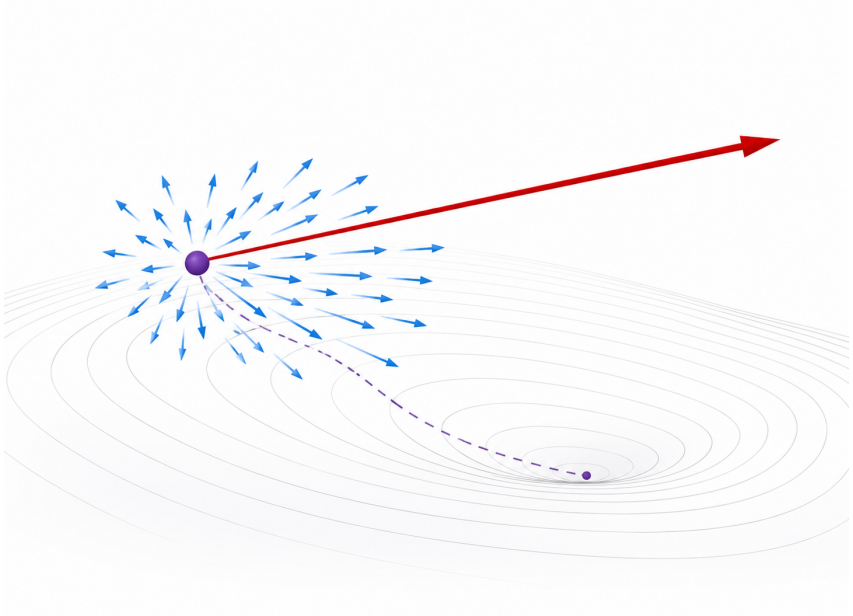
- One sample or mini-batch makes each step memory efficient.

A single extreme gradient can dominate an SGD step

Problem: find

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \ell(\theta) := \mathbb{E}_{\boldsymbol{\xi} \sim P}\big[\ell(\theta, \boldsymbol{\xi})\big],$$

where $\boldsymbol{\xi}$ random element and ℓ strongly convex



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

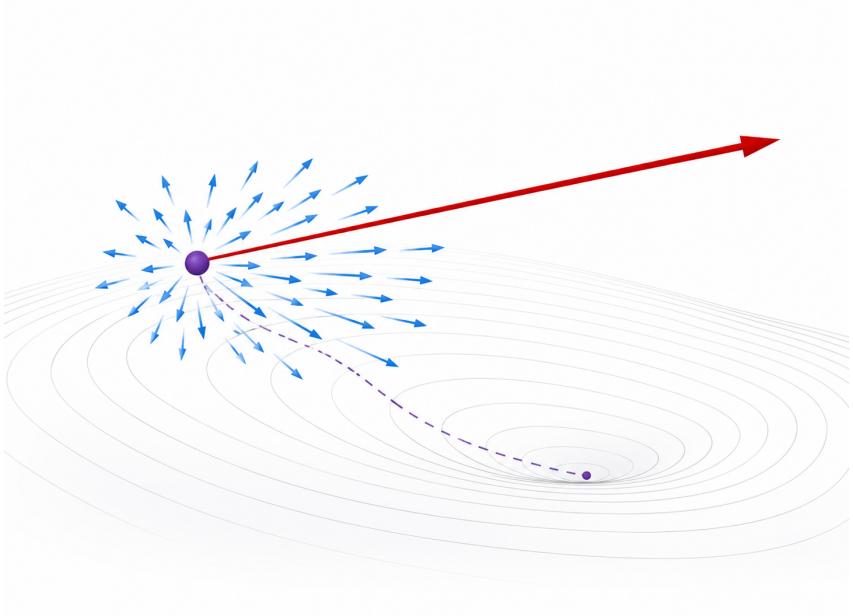
- One sample or mini-batch makes each step memory efficient.
- SGD good at escaping Saddle Points

A single extreme gradient can dominate an SGD step

Problem: find

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \ell(\theta) := \mathbb{E}_{\boldsymbol{\xi} \sim P}\big[\ell(\theta, \boldsymbol{\xi})\big],$$

where $\boldsymbol{\xi}$ random element and ℓ strongly convex



Stochastic approximation begins with Robbins&Monro [RM51].

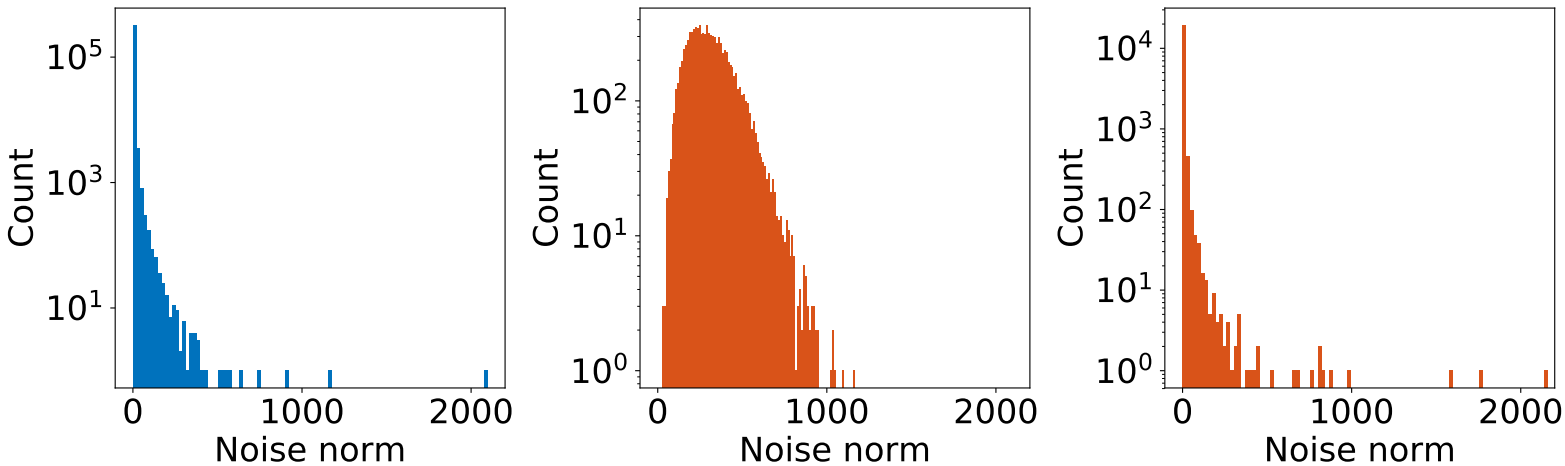
$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1})$$

- One sample or mini-batch makes each step memory efficient.
- SGD good at escaping Saddle Points
- Good generalisation: SGD finds flat minima

SGD with heavy tails for training data in compacts

$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1}), \quad \{\boldsymbol{\xi}_n\}_{n \geq 1}$ i.i.d. sequence, $\eta_n > 0$ learning rate

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

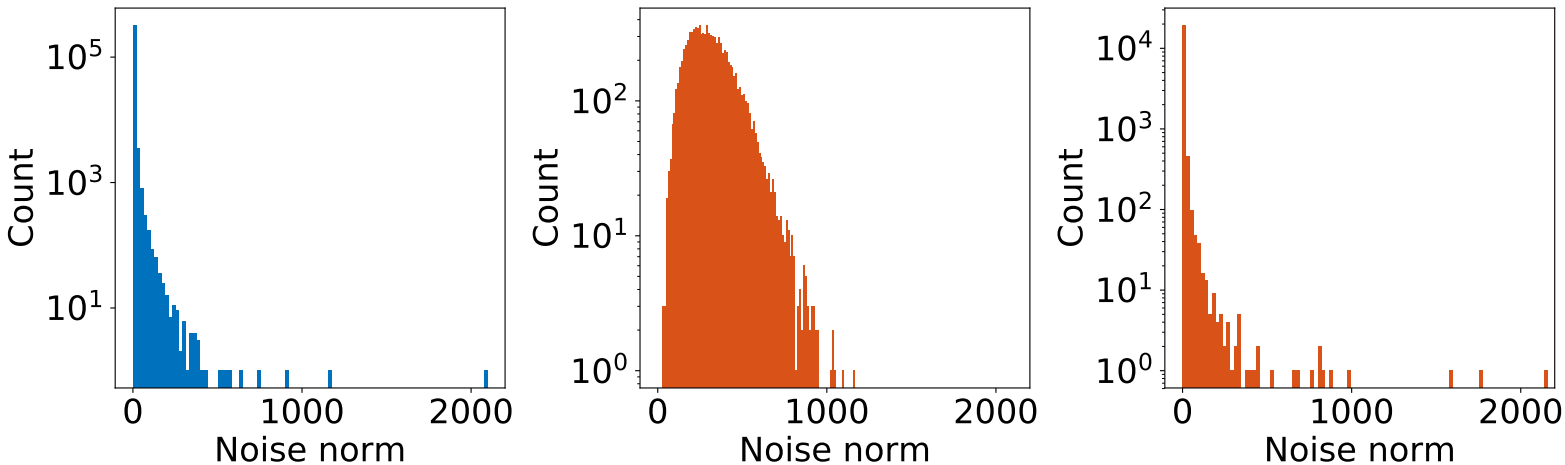


(a) Real (b) Gaussian (c) α-stable

SGD with heavy tails for training data in compacts

$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1}), \quad \{\boldsymbol{\xi}_n\}_{n \geq 1}$ i.i.d. sequence, $\eta_n > 0$ learning rate

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



(d) Real (e) Gaussian (f) α-stable

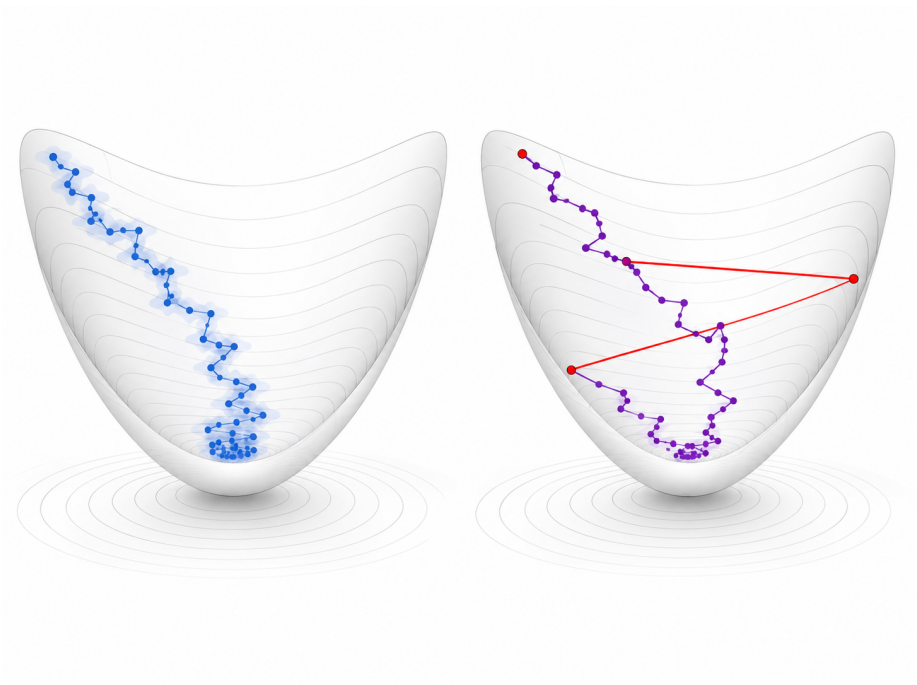
- Intrinsic heaviness [SSG19]:
  - Widths and depths for NN: heaviness depends on the landscape.
  - Minibatch size: heavy noise even for large batch size.
- Extrinsic heaviness [WOR22]:
  - Escape sharp local minimum.

Infinite variance changes both scaling and limit geometry

$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1}), \quad \{\boldsymbol{\xi}_n\}_{n \geq 1} \text{ i.i.d. sequence, learning rate } \eta_n \downarrow 0$$

Infinite variance changes both scaling and limit geometry

$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1}), \quad \{\boldsymbol{\xi}_n\}_{n \geq 1} \text{ i.i.d. sequence, learning rate } \eta_n \downarrow 0$$



Finite variance of $\nabla \ell(\theta, \boldsymbol{\xi})$:

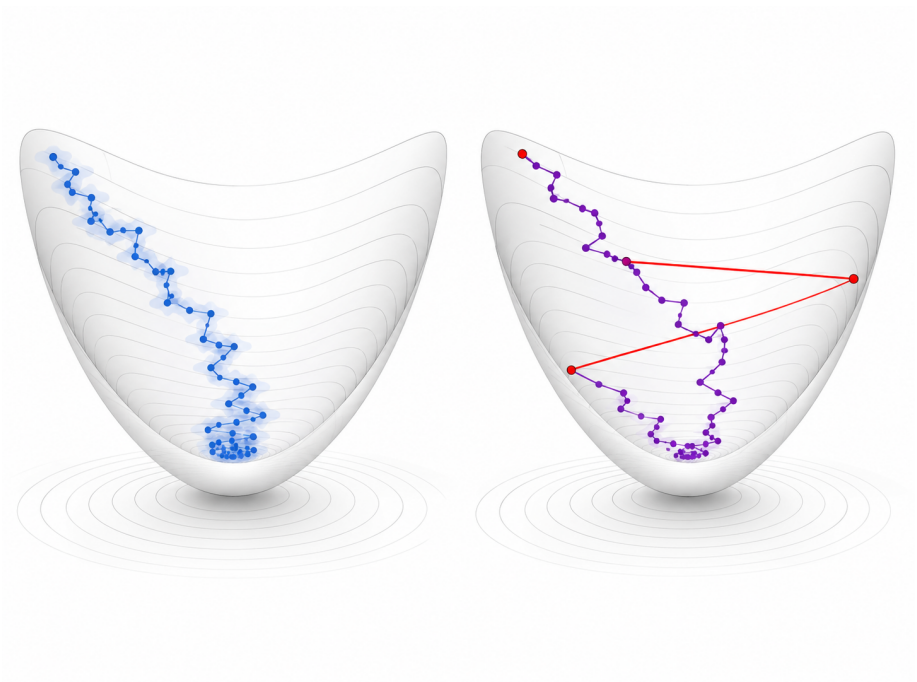
$$\eta_n^{-1/2}(\theta_n - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma)$$

Infinite variance of $\nabla \ell(\theta, \boldsymbol{\xi})$:

$$\eta_n^{1/\alpha - 1} b_1(\eta_n^{-1})(\theta_n - \theta^*) \Rightarrow Z_\infty$$

Infinite variance changes both scaling and limit geometry

$$\theta_{n+1} = \theta_n - \eta_n \nabla \ell(\theta_n, \boldsymbol{\xi}_{n+1}), \quad \{\boldsymbol{\xi}_n\}_{n \geq 1} \text{ i.i.d. sequence, learning rate } \eta_n \downarrow 0$$



Finite variance of $\nabla \ell(\theta, \boldsymbol{\xi})$:

$$\eta_n^{-1/2}(\theta_n - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma)$$

Infinite variance of $\nabla \ell(\theta, \boldsymbol{\xi})$:

$$\eta_n^{1/\alpha - 1} b_1(\eta_n^{-1})(\theta_n - \theta^*) \Rightarrow Z_\infty$$

The OU mechanism survives, but Brownian motion is replaced by a Lévy driver.

Classical Gaussian theory: [Pel98, PJ92].

Classical final-iterate limits

Finite variance: Gaussian limit [Pel98]

$$\eta_n = cn^{-\rho}, \quad 0 < \rho < 1, \qquad H := \nabla^2 \ell(\theta^*) \succ 0,$$
$$\mathbb{E}\left[\|\nabla \ell(\theta^*, \boldsymbol{\xi})\|^2\right] < \infty, \qquad \Gamma := \mathrm{Cov}(\nabla \ell(\theta^*, \boldsymbol{\xi})),$$
$$\eta_n^{-1/2}(\theta_n - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma), \qquad H\Sigma + \Sigma H = \Gamma,$$
$$\mathrm{d}Z_t = -HZ_t\,\mathrm{d}t + \Gamma^{1/2}\mathrm{d}B_t, \qquad Z_\infty \sim \mathcal{N}(0, \Sigma).$$

Classical final-iterate limits

Finite variance: Gaussian limit [Pel98]

$$\eta_n = cn^{-\rho}, \quad 0 < \rho < 1, \qquad H := \nabla^2 \ell(\theta^*) \succ 0,$$
$$\mathbb{E}\left[\|\nabla \ell(\theta^*, \boldsymbol{\xi})\|^2\right] < \infty, \qquad \Gamma := \mathrm{Cov}(\nabla \ell(\theta^*, \boldsymbol{\xi})),$$
$$\eta_n^{-1/2}(\theta_n - \theta^*) \;\Rightarrow\; \mathcal{N}(0, \Sigma), \qquad H\Sigma + \Sigma H = \Gamma,$$
$$\mathrm{d}Z_t = -HZ_t\,\mathrm{d}t + \Gamma^{1/2}\mathrm{d}B_t, \qquad Z_\infty \sim \mathcal{N}(0, \Sigma).$$

Infinite variance: Krasulina (1969) [Kra69]

$$d = 1, \qquad \eta_n = 1/n, \qquad \ell''(\theta^*) > 1 - 1/\alpha,$$
$$\nabla \ell(\theta, \boldsymbol{\xi}_n) - \nabla \ell(\theta) = \zeta_n, \qquad (\zeta_n)_{n \geq 1} \text{ i.i.d.},$$
$$n^{-1/\alpha} \sum_{k=1}^{n} \zeta_k \;\Rightarrow\; S_\alpha, \qquad 1 < \alpha < 2,$$
$$n^{1-1/\alpha}(\theta_n - \theta^*) \;\Rightarrow\; \widetilde{S}_\alpha, \qquad \widetilde{S}_\alpha \text{ is } \alpha\text{-stable}.$$

Multidimensional regular variation [BDM02]: the right language for extreme gradients

For some $\alpha \in (1,2)$, we assume

$$\mathbb{P}(\|\nabla \ell(\theta, \boldsymbol{\xi})\| > x) = x^{-\alpha} b_0(x, \theta),$$

$$x\mathbb{P}\left(\frac{b_1(x)\nabla \ell(\theta, \boldsymbol{\xi})}{x^{1/\alpha}} \in \cdot \right) \xrightarrow{v} \nu(\theta, \cdot) \quad \text{as } x \to \infty$$

- b_0, b_1 allow **slowly varying** corrections; ν records jump directions.
- Gradient noise can be asymmetric, multidimensional, and parameter dependent.
- Earlier theory did not provide both path and final-iterate limits at this level.

Background: [BDM02, Res07]; evidence and earlier results: [SSG19, GSZ21, Kra69, WGZ+21].

Two step-size regimes answer two distinct questions

	Constant step η	Decaying step η_n
Question	Whole path around gradient flow?	Final iterate around θ^*?
Limit	Time-inhomogeneous OU	Stationary Lévy-driven OU
Topology	J_1 in $\mathcal{D}[0,\infty)$	Weak convergence in $\mathbb{R}^d$
Use	Trajectory approximation	Asymptotic inference

Both limits use the same tail measure ν, viewed along different clocks.

The assumptions allow rich noise while stabilizing the drift

- Multivariate regular variation holds uniformly on compact θ-sets:

$$\mathbb{P}(\|\nabla \ell(\theta, \boldsymbol{\xi})\| > x) = x^{-\alpha} b_0(x, \theta),$$

$$x\mathbb{P}\left(x^{-1/\alpha} b_1(x) \nabla \ell(\theta, \boldsymbol{\xi}) \in \cdot\right) \xrightarrow{v} \nu(\theta, \cdot), \quad \text{uniformly for } \theta \in K \Subset \mathbb{R}^d.$$

The assumptions allow rich noise while stabilizing the drift

- Multivariate regular variation holds uniformly on compact θ-sets:

$$\mathbb{P}(\|\nabla \ell(\theta, \boldsymbol{\xi})\| > x) = x^{-\alpha} b_0(x, \theta),$$

$$x\mathbb{P}\left(x^{-1/\alpha} b_1(x) \nabla \ell(\theta, \boldsymbol{\xi}) \in \cdot\right) \xrightarrow{v} \nu(\theta, \cdot), \quad \text{uniformly for } \theta \in K \Subset \mathbb{R}^d.$$

- The stochastic gradient is Lipschitz in θ almost surely:

$$\|\nabla \ell(\theta, \boldsymbol{\xi}) - \nabla \ell(\theta', \boldsymbol{\xi})\| \leq C_{\mathrm{Lip}} \|\theta - \theta'\| \qquad \text{a.s.}$$

The assumptions allow rich noise while stabilizing the drift

- Multivariate regular variation holds uniformly on compact θ-sets:

$$\mathbb{P}(\|\nabla \ell(\theta, \boldsymbol{\xi})\| > x) = x^{-\alpha} b_0(x, \theta),$$

$$x\mathbb{P}\left(x^{-1/\alpha} b_1(x) \nabla \ell(\theta, \boldsymbol{\xi}) \in \cdot\right) \xrightarrow{v} \nu(\theta, \cdot), \quad \text{uniformly for } \theta \in K \Subset \mathbb{R}^d.$$

- The stochastic gradient is Lipschitz in θ almost surely:

$$\|\nabla \ell(\theta, \boldsymbol{\xi}) - \nabla \ell(\theta', \boldsymbol{\xi})\| \leq C_{\mathrm{Lip}} \|\theta - \theta'\| \qquad \text{a.s.}$$

- The population loss is smooth and strongly convex: for $H(\theta) = \nabla^2 \ell(\theta)$,

$$\sup_{\theta \in \mathbb{R}^d} \|H(\theta)\| < \infty, \qquad \inf_{\substack{\theta \in \mathbb{R}^d \\ \|u\|_p = 1}} u^{\mathsf{T}} H(\theta) u^{\langle p-1 \rangle} > 0, \quad p \in (1, \alpha),$$

  where $u^{\langle p-1 \rangle} = (\mathrm{sgn}(u_i)|u_i|^{p-1})_{i=1}^d$.

The assumptions allow rich noise while stabilizing the drift

- Multivariate regular variation holds uniformly on compact θ-sets:

$$\mathbb{P}(\|\nabla \ell(\theta, \boldsymbol{\xi})\| > x) = x^{-\alpha} b_0(x, \theta),$$

$$x\mathbb{P}\left(x^{-1/\alpha} b_1(x) \nabla \ell(\theta, \boldsymbol{\xi}) \in \cdot\right) \xrightarrow{v} \nu(\theta, \cdot), \quad \text{uniformly for } \theta \in K \Subset \mathbb{R}^d.$$

- The stochastic gradient is Lipschitz in θ almost surely:

$$\|\nabla \ell(\theta, \boldsymbol{\xi}) - \nabla \ell(\theta', \boldsymbol{\xi})\| \leq C_{\mathrm{Lip}} \|\theta - \theta'\| \qquad \text{a.s.}$$

- The population loss is smooth and strongly convex: for $H(\theta) = \nabla^2 \ell(\theta)$,

$$\sup_{\theta \in \mathbb{R}^d} \|H(\theta)\| < \infty, \qquad \inf_{\substack{\theta \in \mathbb{R}^d \\ \|u\|_p = 1}} u^{\mathsf{T}} H(\theta) u^{\langle p-1 \rangle} > 0, \quad p \in (1, \alpha),$$

  where $u^{\langle p-1 \rangle} = (\mathrm{sgn}(u_i)|u_i|^{p-1})_{i=1}^d$.

- A qth-order Taylor remainder is available for some $q \in (1, \alpha)$:

$$\left\|\nabla \ell(\theta_1) - \nabla \ell(\theta_2) - \nabla^2 \ell(\theta_2)(\theta_1 - \theta_2)\right\| \leq K \|\theta_1 - \theta_2\|^q.$$

No symmetry, coordinate independence, or exact Pareto form is required.

Theorem 1: constant-step paths converge to a jump process

Let $\bar{\theta}$ solve the gradient flow ODE

$$\dot{\bar{\theta}} = -\nabla \ell(\bar{\theta})$$

and let θ^η interpolate constant-step SGD.

Theorem 1 (constant step: functional limit). Under the stated assumptions, for $1 < \alpha < 2$ and as step size $\eta \to 0$ we have

$$\eta^{1/\alpha - 1} b_1(\eta^{-1})\big(\theta^\eta(\cdot) - \bar{\theta}(\cdot)\big) \Rightarrow Z(\cdot) \quad \text{in } (\mathcal{D}[0,\infty), J_1),$$

where

$$\mathrm{d}Z_t = -\nabla^2 \ell(\bar{\theta}(t)) Z_t \,\mathrm{d}t + \mathrm{d}L_t, \qquad \nu_t(A) = \int_0^t \nu(\bar{\theta}(s), A)\,\mathrm{d}s.$$

Source: Theorem 3.1 of Blanchet–M–Yang [BMY26]. Theorem 1 does not need strong convexity of ℓ!.

Theorem 1 preserves where and when each extreme occurs

The tail law of $\nabla \ell(\theta, \boldsymbol{\xi})$ may change as θ moves.

- L has independent, but generally non-stationary, increments.
- Jump measure of L at time s is $\nu(\bar{\theta}(s), \mathrm{d}z)\,\mathrm{d}s$.
- The Hessian transports and damps every jump:

$$Z_t = \Phi(t)\left(Z_0 + \int_0^t \Phi(s)^{-1}\mathrm{d}L_s\right), \qquad \dot{\Phi}(t) = -\nabla^2 \ell(\bar{\theta}(t))\Phi(t).$$

Theorem 1 preserves where and when each extreme occurs

The tail law of $\nabla \ell(\theta, \boldsymbol{\xi})$ may change as θ moves.

- L has independent, but generally non-stationary, increments.
- Jump measure of L at time s is $\nu(\bar{\theta}(s), \mathrm{d}z)\,\mathrm{d}s$.
- The Hessian transports and damps every jump:

$$Z_t = \Phi(t)\left(Z_0 + \int_0^t \Phi(s)^{-1}\mathrm{d}L_s\right), \qquad \dot{\Phi}(t) = -\nabla^2 \ell(\bar{\theta}(t))\Phi(t).$$

The result approximates rare excursions along the whole SGD trajectory.

Theorem 2: decaying-step iterates converge to stationarity

Let $\eta_n = cn^{-\rho}$, $\rho \in (0, 1]$ and $H = \nabla^2 \ell(\theta^*)$.

Theorem 2 (decaying step: final-iterate limit). Under the stated assumptions,

$$\eta_n^{1/\alpha - 1} b_1(\eta_n^{-1})(\theta_n - \theta^*) \Rightarrow Z_\infty,$$

where Z_∞ is the unique stationary law of

$$\mathrm{d}Z_t = -\nabla^2 \ell(\theta^*) Z_t \, \mathrm{d}t + \mathrm{d}L_t, \qquad Z_\infty \overset{d}{=} \int_0^\infty e^{-\nabla^2 \ell(\theta^*) t} \, \mathrm{d}L_t,$$

and L has Lévy measure $\nu(\theta^*, \cdot)$.

Source: Theorem 4.1 of Blanchet–M–Yang [BMY26].

The stationary law separates tail geometry from loss geometry

The Lévy measure of the limit $Z_\infty \overset{d}{=} \int_0^\infty e^{-\nabla^2 \ell(\theta^*)t}\,\mathrm{d}L_t$ is explicit:

$$\widetilde{\nu}(A) = \int_0^\infty \nu\Big(\theta^*, \{x : e^{-\nabla^2 \ell(\theta^*)t} x \in A\}\Big)\,\mathrm{d}t.$$

Incoming extremes described by ν; contraction&rotation by $e^{-\nabla^2 \ell(\theta^*)t}$.

The stationary law separates tail geometry from loss geometry

The Lévy measure of the limit $Z_\infty \stackrel{d}{=} \int_0^\infty e^{-\nabla^2\ell(\theta^*)t}\,\mathrm{d}L_t$ is explicit:

$$\widetilde{\nu}(A) = \int_0^\infty \nu\Big(\theta^*, \{x : e^{-\nabla^2\ell(\theta^*)t}x \in A\}\Big)\,\mathrm{d}t.$$

Incoming extremes described by ν; contraction&rotation by $e^{-\nabla^2\ell(\theta^*)t}$.
For the borderline step $\eta_n = c/n$, replace $\nabla^2\ell(\theta^*)$ by

$$\nabla^2\ell(\theta^*) - \frac{1-1/\alpha}{c}\mathrm{Id}, \qquad c > \frac{1-1/\alpha}{\sigma_{\min}(\nabla^2\ell(\theta^*))}.$$

Unlike for $\rho < 1$, the constant c changes the limiting law.

Stationary Lévy-driven OU laws: [LM05].

OLS inherits heavy tails directly from the regression errors

Consider

$$y = x^\top \theta^* + \varepsilon, \qquad \nabla \ell(\theta, x, \varepsilon) = x x^\top (\theta - \theta^*) - x\varepsilon.$$

If x is bounded and $\mathbb{P}(|\varepsilon| > t) \sim t^{-\alpha}$, then:

- $x\varepsilon$ makes the gradient α-regularly varying;
- extreme-gradient directions are determined by $\pm x/\|x\|$;
- the tail law does not depend on the current θ.

Hence Theorems 1 and 2 apply with a fixed angular tail measure.

Logistic regression can switch tail regimes in optimisation

Logistic regression model: $y_i \in \{-1, 1\}$ binary response, x_i α-regular varying random vector, $\theta^* \in \mathbb{R}^d$ true param:

$$\mathbb{P}(y_i = 1 | x_i, \theta^*) = \frac{1}{1 + \exp(-x_i^\top \theta^*)}, \quad \mathbb{P}(y_i = -1 | x_i, \theta^*) = \frac{\exp(-x_i^\top \theta^*)}{1 + \exp(-x_i^\top \theta^*)}$$

Loss $\ell(\theta, x, y) = \ln\left(1 + \exp\left(-y\theta^\top x\right)\right) + \frac{\lambda}{2}\|\theta\|^2$ with gradient

$$\nabla \ell(\theta, x, y) = -\frac{y e^{-y\theta^\top x}}{1 + e^{-y\theta^\top x}} x + \lambda\theta.$$

- tail directions depend on the current margin $y\theta^\top x$ and hence on θ;
- at $\theta = \theta^* \neq 0$, the surviving tail directions lie on

$$\mathbb{S}^{d-1} \cap (\theta^*)^\perp \simeq \mathbb{S}^{d-2}.$$

- in $d = 1$ and $\theta\theta^* > 0$, heavy tails vanish locally.

Logistic regression can switch tail regimes in optimisation

Logistic regression model: $y_i \in \{-1, 1\}$ binary response, x_i α-regular varying random vector, $\theta^* \in \mathbb{R}^d$ true param:

$$\mathbb{P}(y_i = 1 | x_i, \theta^*) = \frac{1}{1 + \exp(-x_i^\top \theta^*)}, \quad \mathbb{P}(y_i = -1 | x_i, \theta^*) = \frac{\exp(-x_i^\top \theta^*)}{1 + \exp(-x_i^\top \theta^*)}$$

Loss $\ell(\theta, x, y) = \ln\left(1 + \exp\left(-y\theta^\top x\right)\right) + \frac{\lambda}{2}\|\theta\|^2$ with gradient

$$\nabla \ell(\theta, x, y) = -\frac{y e^{-y\theta^\top x}}{1 + e^{-y\theta^\top x}} x + \lambda\theta.$$

- tail directions depend on the current margin $y\theta^\top x$ and hence on θ;
- at $\theta = \theta^* \neq 0$, the surviving tail directions lie on

$$\mathbb{S}^{d-1} \cap (\theta^*)^\perp \simeq \mathbb{S}^{d-2}.$$

- in $d = 1$ and $\theta\theta^* > 0$, heavy tails vanish locally.

The local scaling then returns to $\eta_n^{-1/2}$ and the limit is Gaussian.

The limit theorems replace a covariance by a full tail geometry

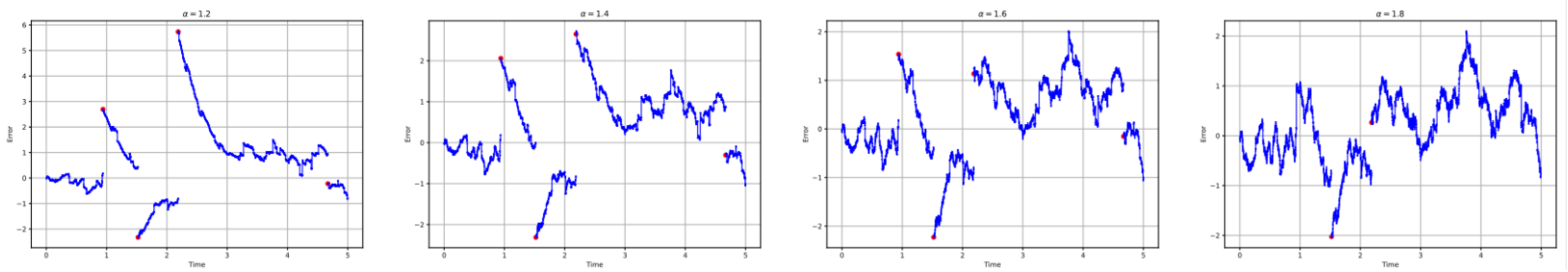
	Finite variance	Infinite variance
Scaling	$\eta_n^{-1/2}$	$\eta_n^{1/\alpha-1} b_1(\eta_n^{-1})$
Noise limit	Brownian motion	Additive / Lévy process
Final law	Gaussian	Infinitely divisible
Geometry	covariance Γ	Lévy measure $\nu(\theta, \cdot)$

- Correct normalization prevents misleading Gaussian uncertainty quantification.
- The explicit Z_∞ supports asymptotic confidence regions.
- Parameter dependence reveals tail-regime changes invisible to covariance methods.

A quadratic experiment isolates the stable-noise mechanism

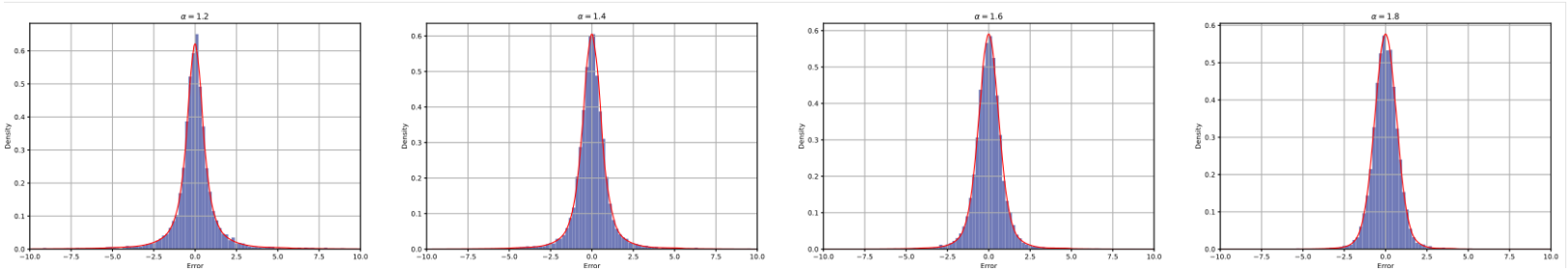
Problem: $\min_{\theta\in\mathbb{R}^2} \frac{1}{2}\theta^\top A\theta + \theta^\top b$. Take $A = \mathrm{diag}(2,1)$, $b = (1,1)^\mathsf{T}$ and

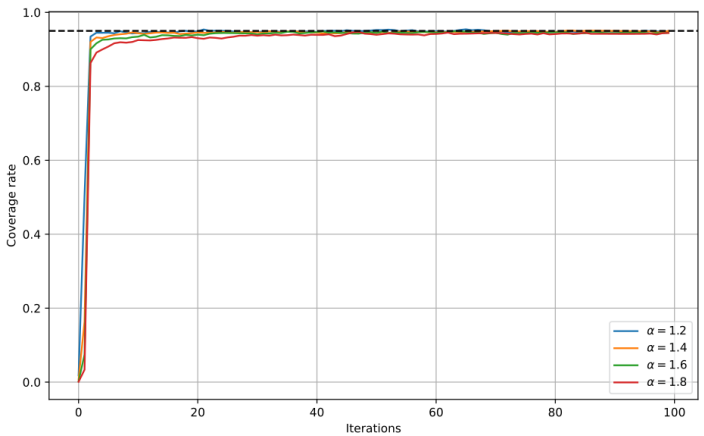
$$\theta_{i+1} = \theta_i - \eta(A\theta_i + b + Z_i), \qquad Z_i \text{ i.i.d. standard } \alpha\text{-stable.}$$



Constant $\eta = 10^{-3}$: red points mark scaled increments above 1; jumps dominate for small α and soften as $\alpha \uparrow 2$.

The stationary law matches both densities and coverage





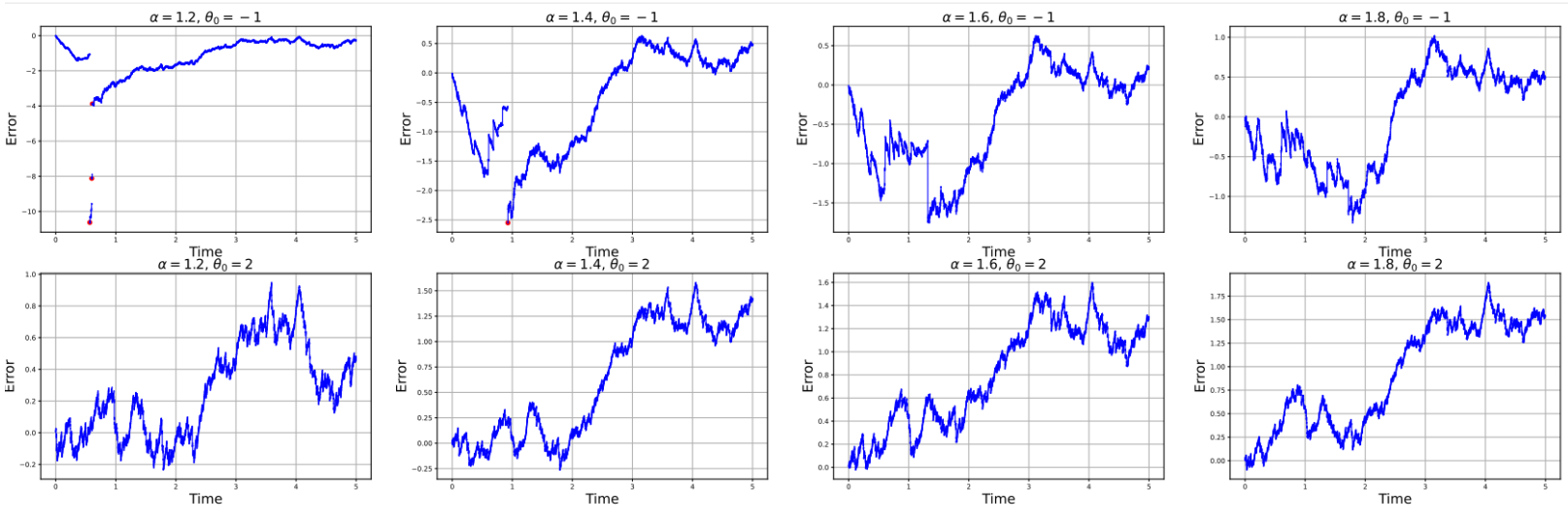
- Learning rate $\eta_n = n^{-0.6}$.
- 10^4 runs of SGD.
- Red curves above: stationary OU densities.
- Coverage approaches the 95% target.

Coverage (oracle confidence interval):

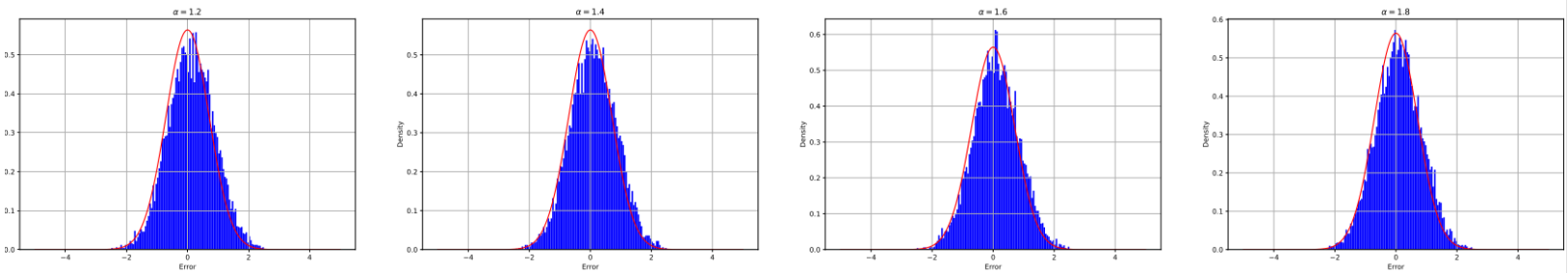
$$Y_n^{(r)} := \eta_n^{1/\alpha - 1} b_1(\eta_n^{-1})(\theta_n^{(r)} - \theta^*), \qquad \mathbb{P}(|Z_\infty| \le q_\alpha) = 0.95,$$

$$\widehat{C}_n := \frac{1}{10^4} \sum_{r=1}^{10^4} \mathbf{1}_{\{|Y_n^{(r)}| \le q_\alpha\}} \longrightarrow 0.95.$$

Logistic regression confirms the predicted phase change



Top: starting at $\theta_0 = -1$, jump fluctuations shrink after the path crosses 0; bottom: starting at $\theta_0 = 2$, the path remains light-tailed.



Near $\theta^* = 1$, the $\eta_n^{-1/2}$-scaled histograms match Gaussian densities.

Theorem 1 follows from extreme-gradient point processes

For constant η, decompose the scaled SGD–GD error into a linear recursion and a Taylor remainder.

1. Regular variation yields convergence of extreme-gradient point measures.

Theorem 1 follows from extreme-gradient point processes

For constant η, decompose the scaled SGD–GD error into a linear recursion and a Taylor remainder.

1. Regular variation yields convergence of extreme-gradient point measures.
2. Their partial sums converge in J_1 to the additive process L.

Theorem 1 follows from extreme-gradient point processes

For constant η, decompose the scaled SGD–GD error into a linear recursion and a Taylor remainder.

1. Regular variation yields convergence of extreme-gradient point measures.
2. Their partial sums converge in J_1 to the additive process L.
3. The solution map transports jumps through $\nabla^2 \ell(\bar{\theta}(t))$.

Theorem 1 follows from extreme-gradient point processes

For constant η, decompose the scaled SGD–GD error into a linear recursion and a Taylor remainder.

1. Regular variation yields convergence of extreme-gradient point measures.
2. Their partial sums converge in J_1 to the additive process L.
3. The solution map transports jumps through $\nabla^2 \ell(\bar{\theta}(t))$.
4. Lipschitz and qth-order bounds make the remainder negligible.

The extremes survive the scaling as the jumps of Z.

For the J_1 framework, see Whitt [Whi02].

Theorem 2 follows from time change and stationary tightness

Embed the decaying-step recursion with the clock $s_n = \sum_{k \le n} \eta_k$.

1. Far along the path, the drift freezes at $H = \nabla^2 \ell(\theta^*)$.

Theorem 2 follows from time change and stationary tightness

Embed the decaying-step recursion with the clock $s_n = \sum_{k \le n} \eta_k$.

1. Far along the path, the drift freezes at $H = \nabla^2 \ell(\theta^*)$.
2. The embedded process approaches a time-homogeneous Lévy-driven OU process.

Theorem 2 follows from time change and stationary tightness

Embed the decaying-step recursion with the clock $s_n = \sum_{k \le n} \eta_k$.

1. Far along the path, the drift freezes at $H = \nabla^2 \ell(\theta^*)$.
2. The embedded process approaches a time-homogeneous Lévy-driven OU process.
3. Tightness is proved first for symmetric regularly varying noise.

Theorem 2 follows from time change and stationary tightness

Embed the decaying-step recursion with the clock $s_n = \sum_{k \le n} \eta_k$.

1. Far along the path, the drift freezes at $H = \nabla^2 \ell(\theta^*)$.
2. The embedded process approaches a time-homogeneous Lévy-driven OU process.
3. Tightness is proved first for symmetric regularly varying noise.
4. General asymmetric noise is decomposed into symmetrizable components.

Theorem 2 follows from time change and stationary tightness

Embed the decaying-step recursion with the clock $s_n = \sum_{k \le n} \eta_k$.

1. Far along the path, the drift freezes at $H = \nabla^2 \ell(\theta^*)$.
2. The embedded process approaches a time-homogeneous Lévy-driven OU process.
3. Tightness is proved first for symmetric regularly varying noise.
4. General asymmetric noise is decomposed into symmetrizable components.
5. Letting the embedded time grow gives $Z_\infty = \int_0^\infty e^{-Ht} \mathrm{d}L_t$.

The key difficulty is tightness without a finite αth moment.

References

[BDM02] Bojan Basrak, Richard A. Davis, and Thomas Mikosch.
A characterization of multivariate regular variation.
The Annals of Applied Probability, 12:908–920, 2002.

[BMY26] Jose Blanchet, Aleksandar Mijatović, and Wenhao Yang.
Limit theorems for stochastic gradient descent with infinite variance.
The Annals of Applied Probability, 2026.

[GSZ21] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu.
The heavy-tail phenomenon in SGD.
In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 3964–3975, 2021.

[Kra69] T. P. Krasulina.
On stochastic approximation processes with infinite variance.
Theory of Probability and Its Applications, 14:522–526, 1969.

[LM05] Alexander Lindner and Ross Maller.
Lévy integrals and the stationarity of generalised Ornstein–Uhlenbeck processes.
Stochastic Processes and their Applications, 115:1701–1722, 2005.

[Pel98] Mariane Pelletier.
Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing.
The Annals of Applied Probability, 8:10–44, 1998.

[PJ92] Boris T. Polyak and Anatoli B. Juditsky.
Acceleration of stochastic approximation by averaging.
SIAM Journal on Control and Optimization, 30:838–855, 1992.

[Res07] Sidney I. Resnick.
Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.
Springer, New York, 2007.

[RM51] Herbert Robbins and Sutton Monro.
A stochastic approximation method.
The Annals of Mathematical Statistics, 22:400–407, 1951.

[SSG19] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban.
A tail-index analysis of stochastic gradient noise in deep neural networks.
In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 5827–5837, 2019.

[WGZ+21] Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A. Erdogdu.
Convergence rates of stochastic gradient descent under infinite noise variance.
In *Advances in Neural Information Processing Systems*, volume 34, pages 18866–18877, 2021.

[Whi02] Ward Whitt.
Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues.
Springer, New York, 2002.

[WOR22] Xingyu Wang, Sewoong Oh, and Chang-Han Rhee.
Eliminating sharp minima from sgd with truncated heavy-tailed noise.
In *International Conference on Learning Representations (ICLR)*, 2022.