Eliminating Sharp Minima with Truncated Heavy-tailed Noise

Xingyu Wang*, Sewoong Oh†, Chang-Han Rhee*

Northwestern University*, University of Washington†

DeepMath 2021

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**

Training Set

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**

Training Set

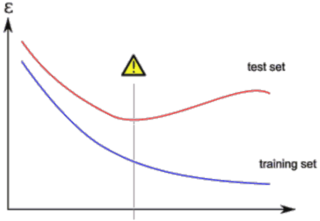
Test Set

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**

Training Set

Test Set



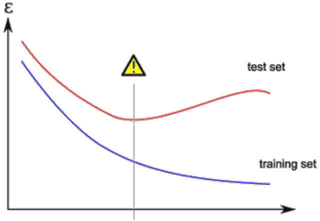
Training/Test Error

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)

Training Set

Test Set



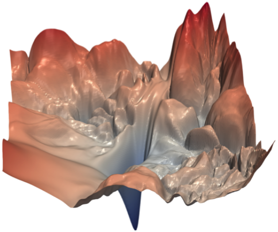
Training/Test Error

Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)
- **Nonconvex Landscape, Numerous Local Minima**

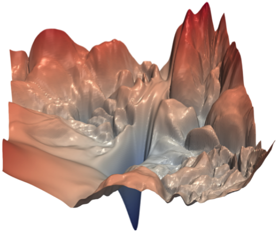
Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)
- **Nonconvex Landscape, Numerous Local Minima**



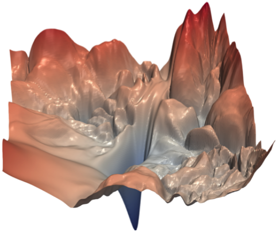
Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)
- **Empirical Observations:** Flat minima (as opposed to sharp minima) generalize better.



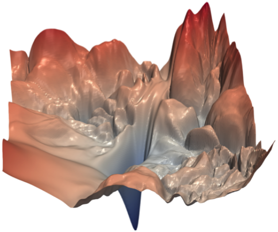
Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)
- **Empirical Observations:** Flat minima (as opposed to sharp minima) generalize better.
  - Among 40+ metrics, sharpness metrics predict generalization best. (Jiang et al., 2020)



Intro: Generalization Gap and Flat Minima

- **Generalization of DNN**
  - Generalization Mystery of Stochastic Gradient Descent (SGD)
- **Empirical Observations:** Flat minima (as opposed to sharp minima) generalize better.
  - Among 40+ metrics, sharpness metrics predict generalization best. (Jiang et al., 2020)



- **Q:** SGD prefers flat minima?

Intro: Heavy-tailed SGD Prefers Flat Minima

GD $$X_j = X_{j-1} - \eta \ \nabla f(X_{j-1})$$

Intro: Heavy-tailed SGD Prefers Flat Minima

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

Traditional Assumption: Light-tailed ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0,\ Z_j \in RV_{-\alpha}$ with $\alpha > 1$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)

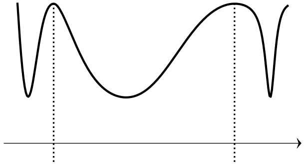
Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)



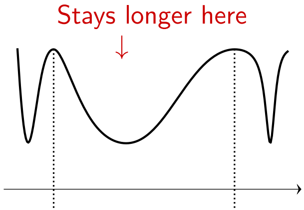
Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)



Intro: Heavy-tailed SGD Prefers Flat Minima

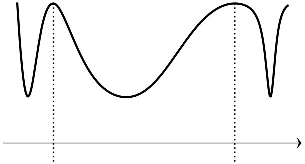
~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Noises:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x)$ resembles power law $x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)

Our Work: Complete Elimination of Sharp Minima



Intro: Truncated Heavy-tailed SGD

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Intro: Truncated Heavy-tailed SGD

Gradient Clipping

$$X_j = X_{j-1} - \underset{\substack{\uparrow}}{\varphi_b}\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Intro: Truncated Heavy-tailed SGD

Gradient Clipping
↓

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

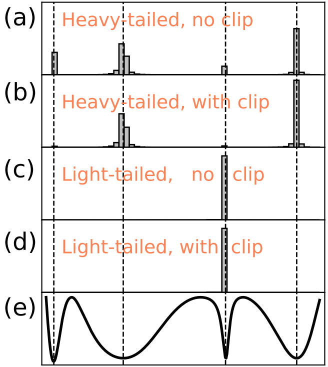
Q: How does truncated heavy-tailed noise help?

Intro: Truncated Heavy-tailed SGD

Gradient Clipping
↓

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Q: How does truncated heavy-tailed noise help?

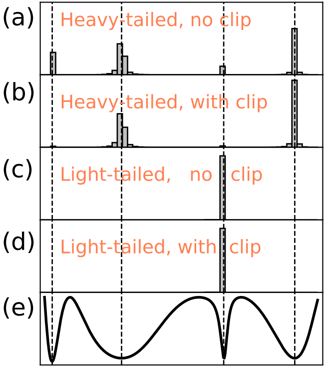


Intro: Truncated Heavy-tailed SGD

Gradient Clipping

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Q: Why does truncated heavy-tailed noise help?



Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

- Extreme Values are Very Rare
- Normal, Exponential, etc

Heavy-Tailed Distributions

- Extreme Values are Frequent
- Power Law, Weibull, etc

Instagram

Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

- Extreme Values are Very Rare
- Normal, Exponential, etc

Heavy-Tailed Distributions

- Extreme Values are Frequent
- Power Law, Weibull, etc

Instagram

Structural difference in the way systemwide rare events arise.

Rare Events depend on “Tail Behaviors”

Light-Tailed Distributions

- Extreme Values are Very Rare
- Normal, Exponential, etc

Systemwide rare events

arise because

EVERYTHING goes wrong.

(Conspiracy Principle)

Heavy-Tailed Distributions

- Extreme Values are Frequent
- Power Law, Weibull, etc

Structural difference in the way systemwide rare events arise.

Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

- Extreme Values are Very Rare
- Normal, Exponential, etc

Systemwide rare events

arise because

EVERYTHING goes wrong.

(Conspiracy Principle)

Heavy-Tailed Distributions

- Extreme Values are Frequent
- Power Law, Weibull, etc

Systemwide rare events

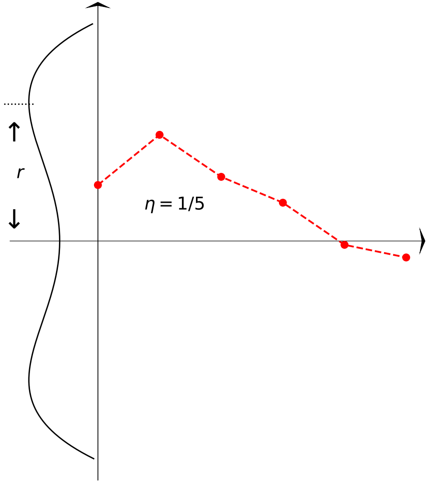
arise because of

A FEW Catastrophes.

(Catastrophe Principle)

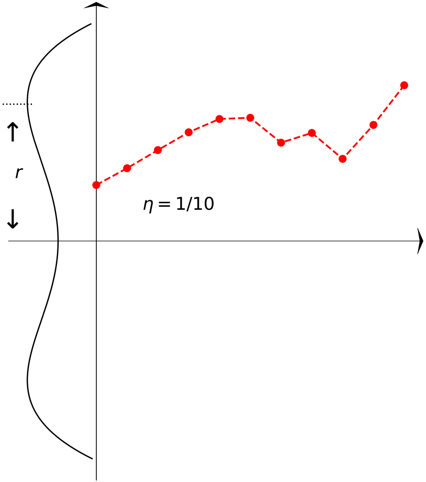
Structural difference in the way systemwide rare events arise.

Typical Behavior of SGD



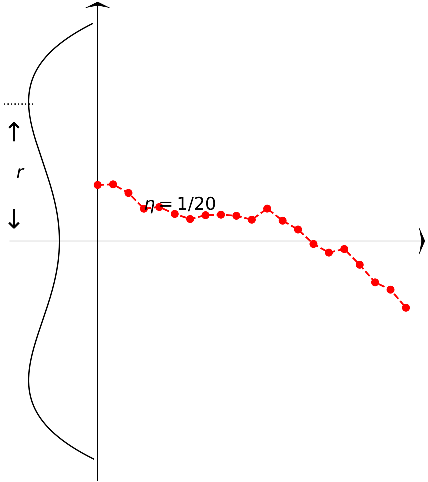
$$X_j^\eta = X_{j-1}^\eta - \eta\big(\nabla f(X_{j-1}^\eta) + Z_j\big)$$

Typical Behavior of SGD



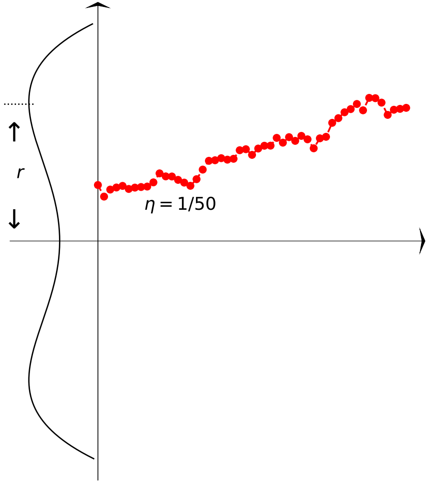
$$X_j^\eta = X_{j-1}^\eta - \eta\big(\nabla f(X_{j-1}^\eta) + Z_j\big)$$

Typical Behavior of SGD



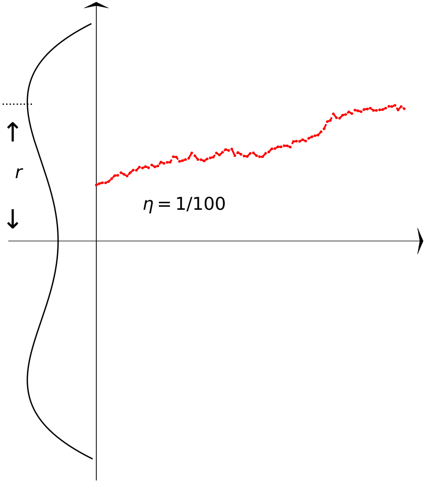
$$X_j^\eta = X_{j-1}^\eta - \eta\big(\nabla f(X_{j-1}^\eta) + Z_j\big)$$

Typical Behavior of SGD



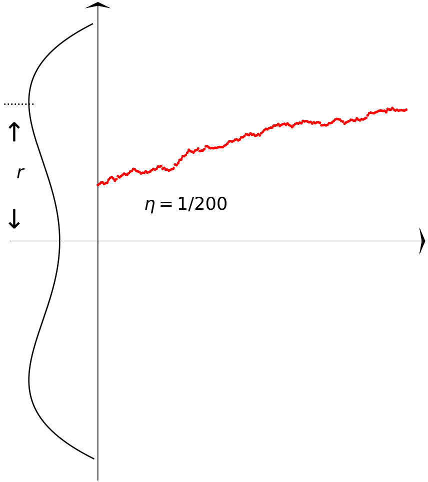
$$X_j^\eta = X_{j-1}^\eta - \eta\left(\nabla f(X_{j-1}^\eta) + Z_j\right)$$

Typical Behavior of SGD



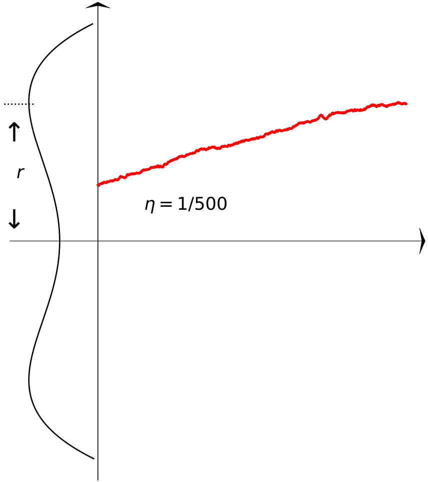
$$X_j^\eta = X_{j-1}^\eta - \eta\left(\nabla f(X_{j-1}^\eta) + Z_j\right)$$

Typical Behavior of SGD



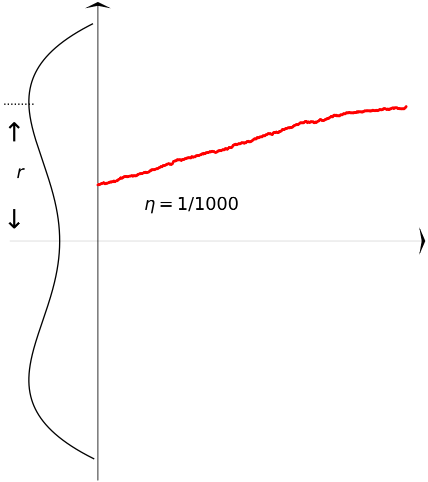
$$X_j^\eta = X_{j-1}^\eta - \eta\left(\nabla f(X_{j-1}^\eta) + Z_j\right)$$

Typical Behavior of SGD



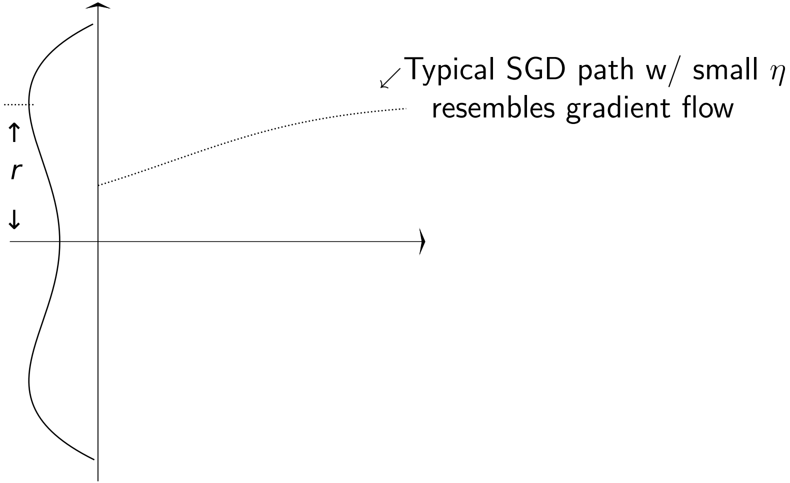
$$X_j^\eta = X_{j-1}^\eta - \eta\left(\nabla f(X_{j-1}^\eta) + Z_j\right)$$

Typical Behavior of SGD

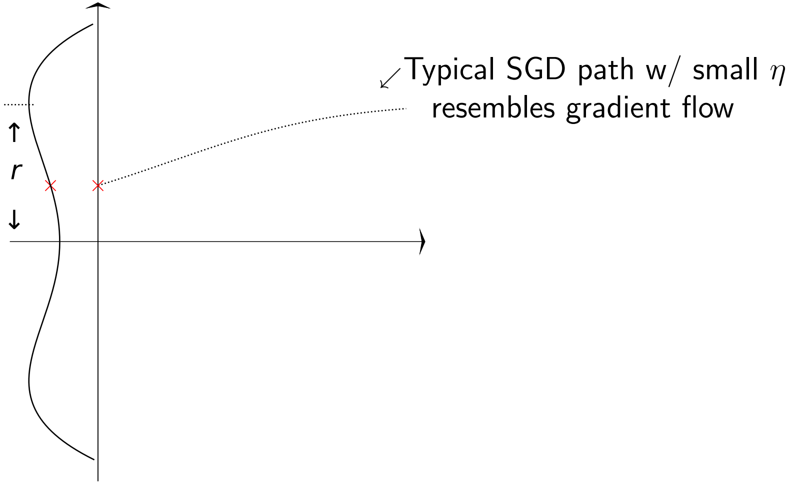


$$X_j^\eta = X_{j-1}^\eta - \eta\left(\nabla f(X_{j-1}^\eta) + Z_j\right)$$

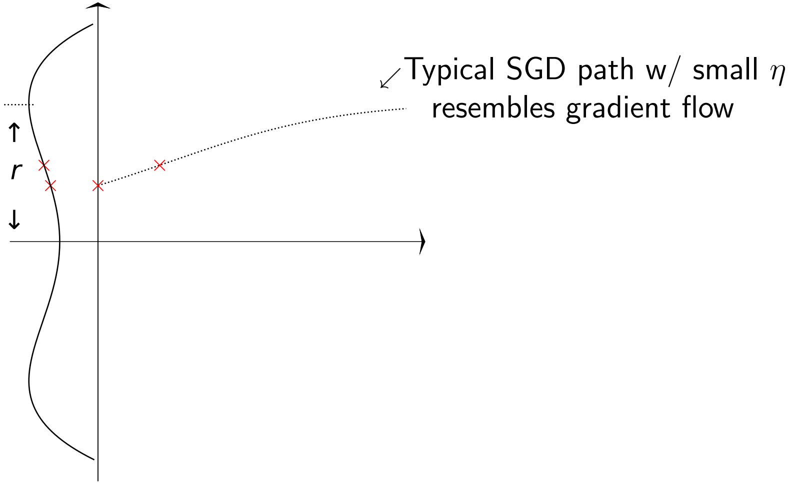
Typical Behavior of SGD



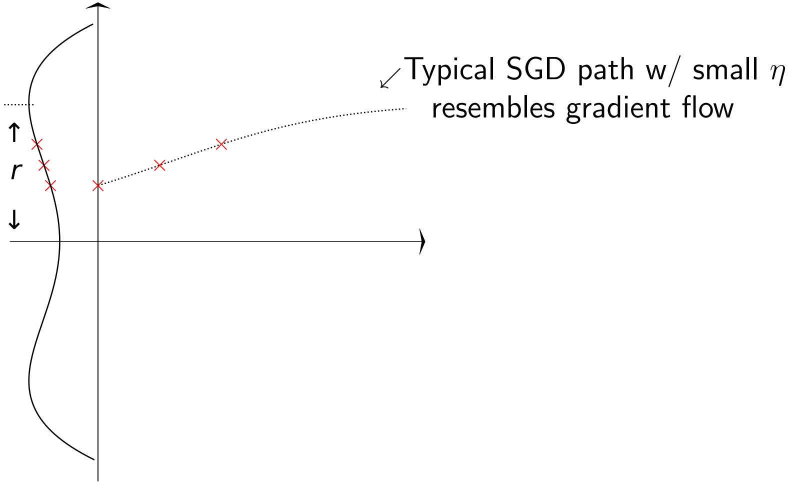
Typical Behavior of SGD



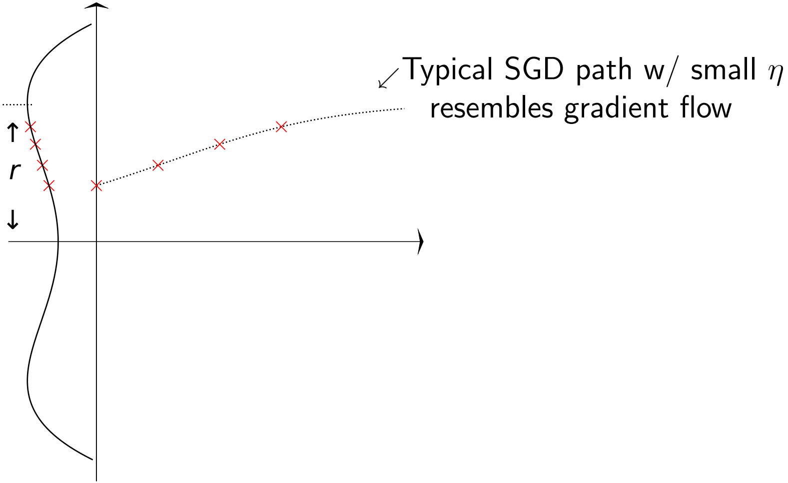
Typical Behavior of SGD



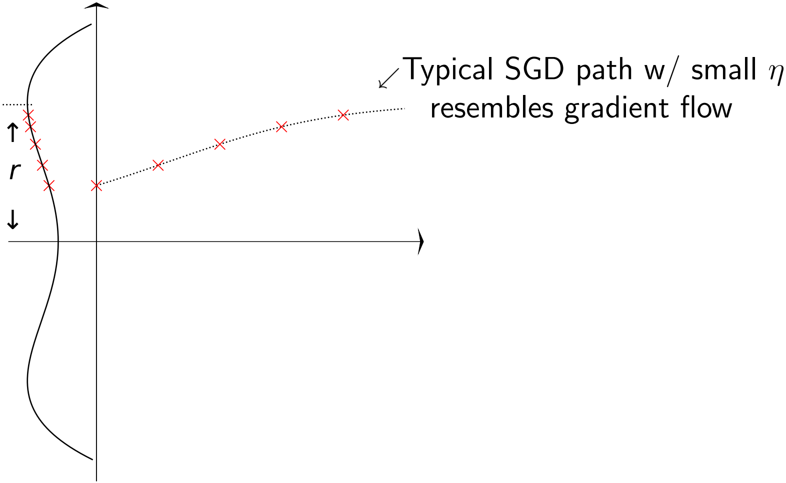
Typical Behavior of SGD



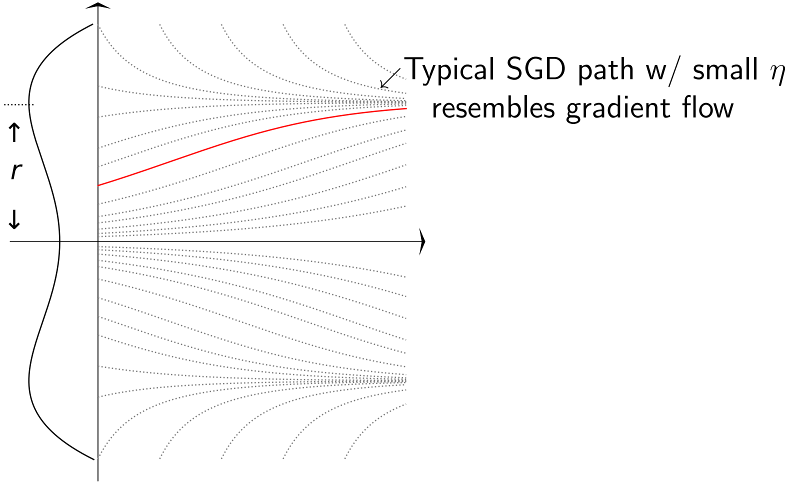
Typical Behavior of SGD



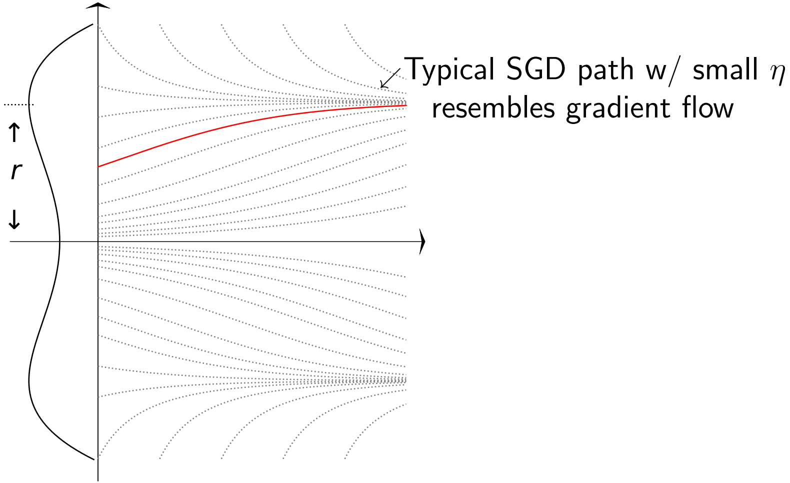
Typical Behavior of SGD



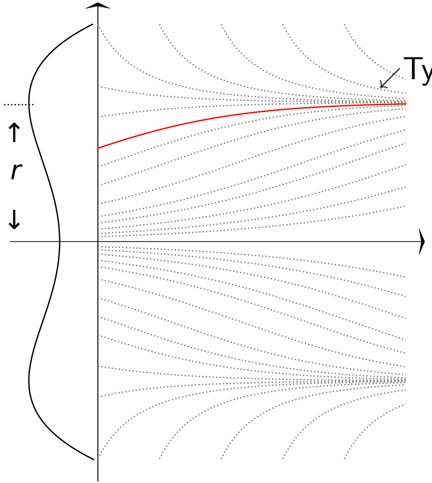
Typical Behavior of SGD



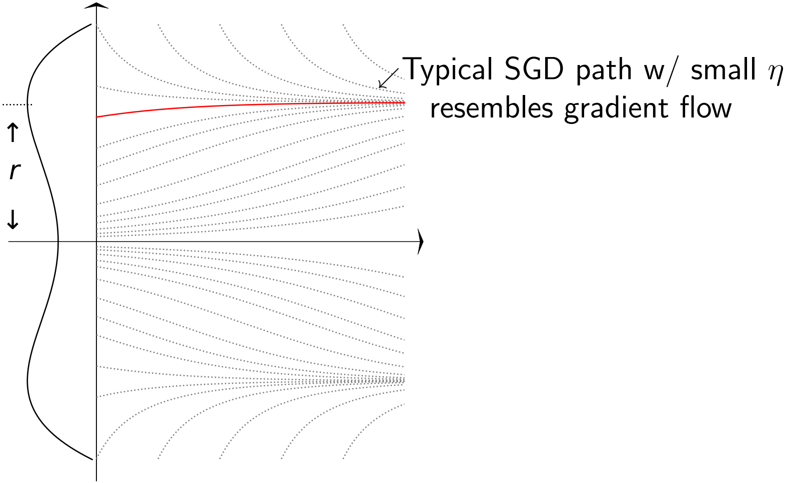
Typical Behavior of SGD



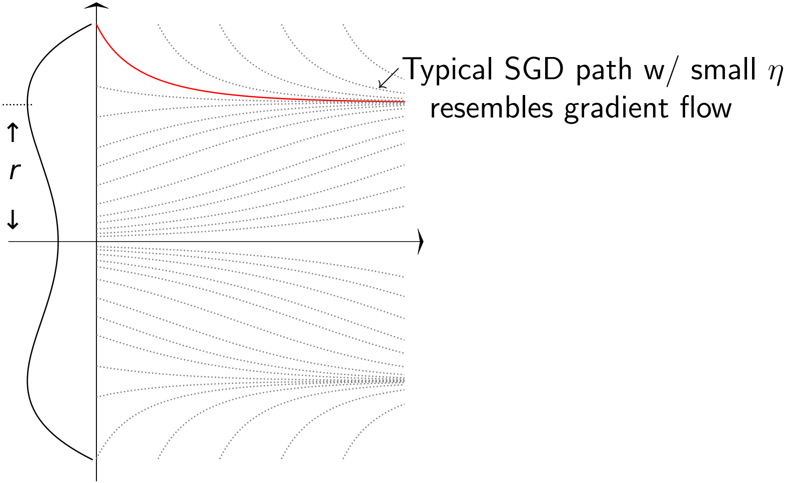
Typical Behavior of SGD



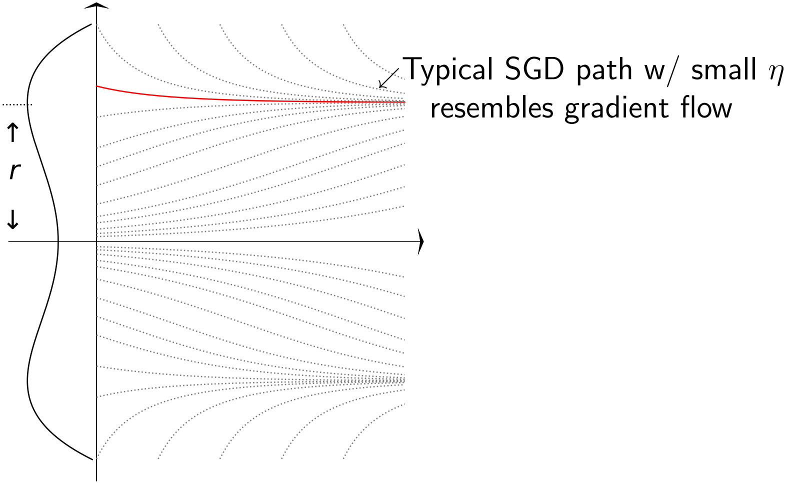
Typical Behavior of SGD



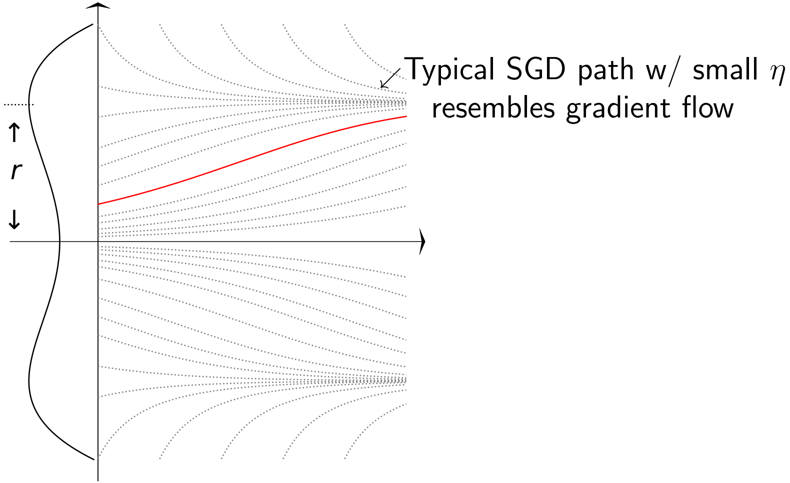
Typical Behavior of SGD



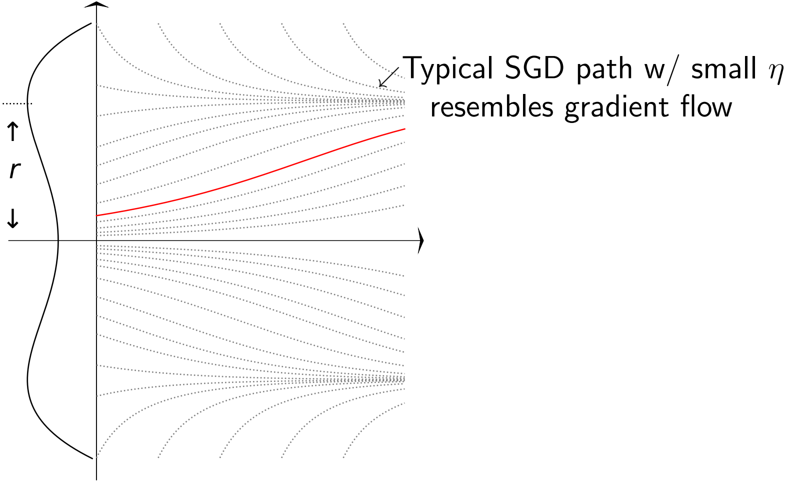
Typical Behavior of SGD



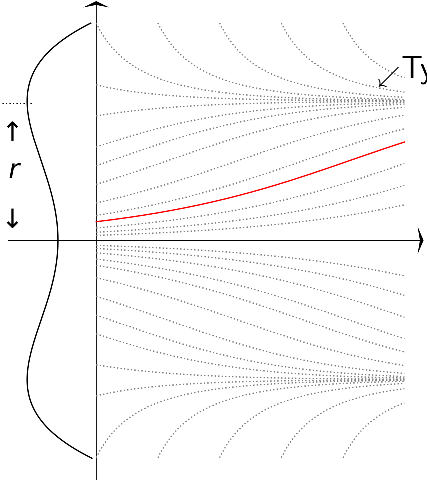
Typical Behavior of SGD



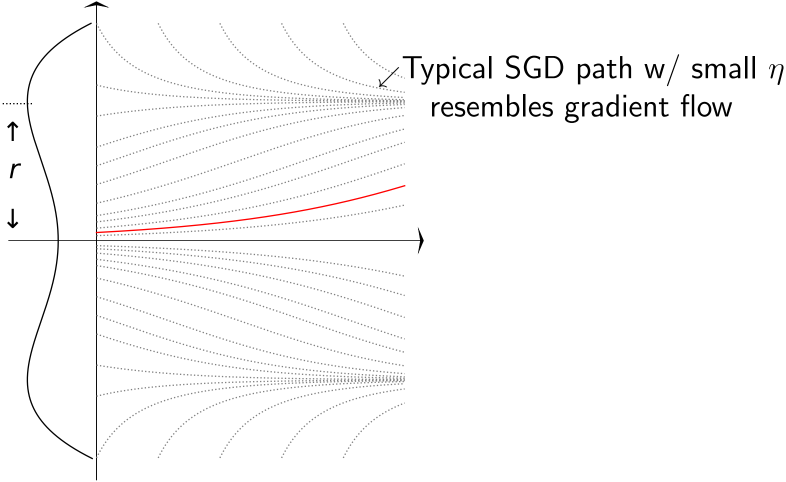
Typical Behavior of SGD



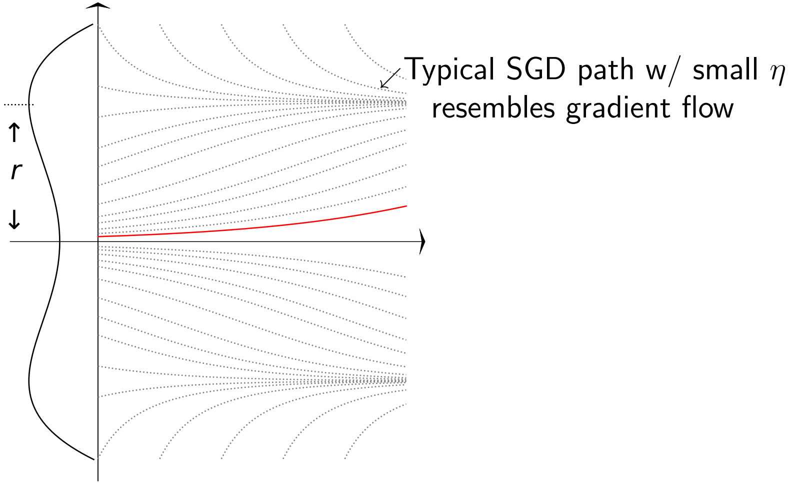
Typical Behavior of SGD



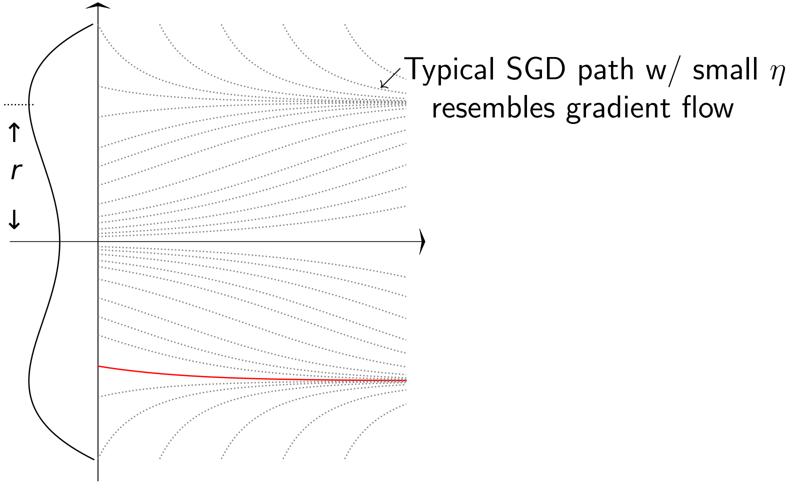
Typical Behavior of SGD



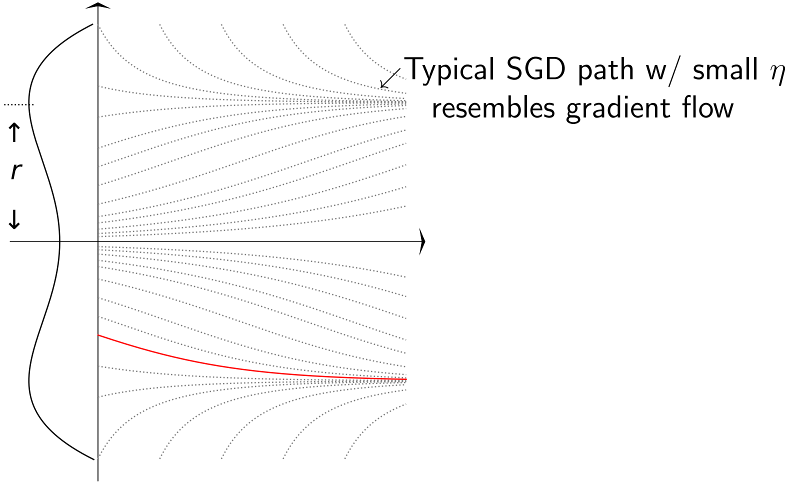
Typical Behavior of SGD



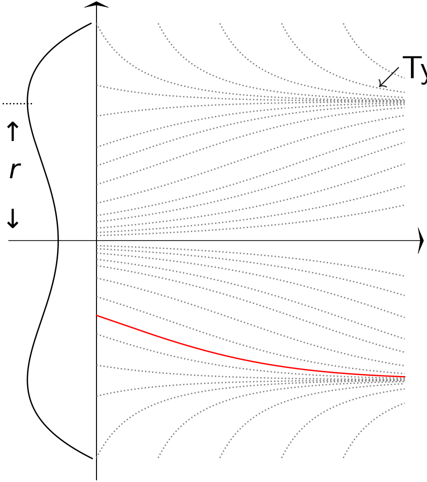
Typical Behavior of SGD



Typical Behavior of SGD

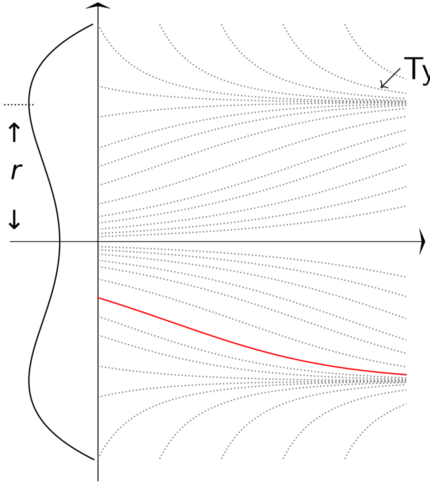


Typical Behavior of SGD



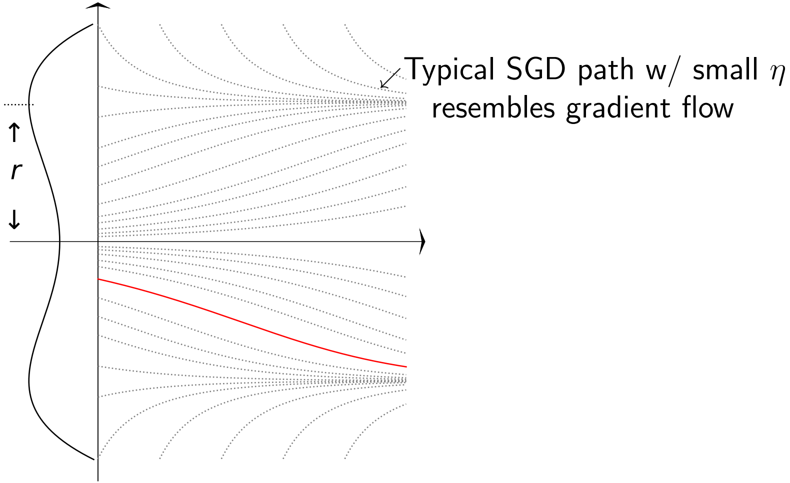
Typical SGD path w/ small η
resembles gradient flow

Typical Behavior of SGD

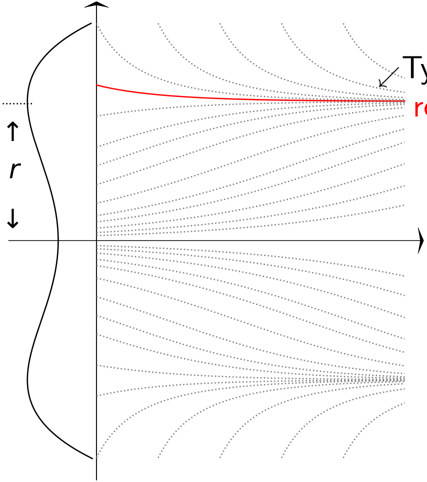


Typical SGD path w/ small η
resembles gradient flow

Typical Behavior of SGD



Typical Behavior of SGD



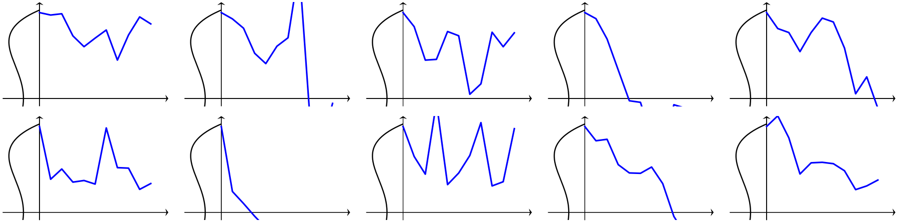
Typical Behavior of SGD

Typical Behavior of SGD

Trajectory of SGD X^η: $\eta = 1/10$ & noises are **light-tailed**

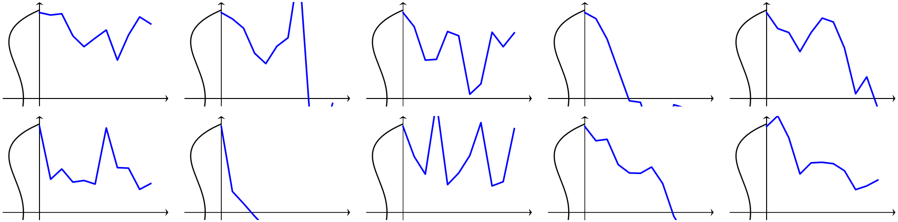
Typical Behavior of SGD

Trajectory of SGD X^{η}: $\eta = 1/10$ & noises are **light-tailed**



Typical Behavior of SGD

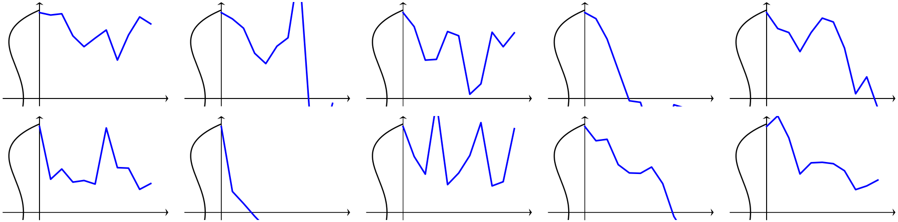
Trajectory of SGD X^η: $\eta = 1/10$ & noises are **light-tailed**



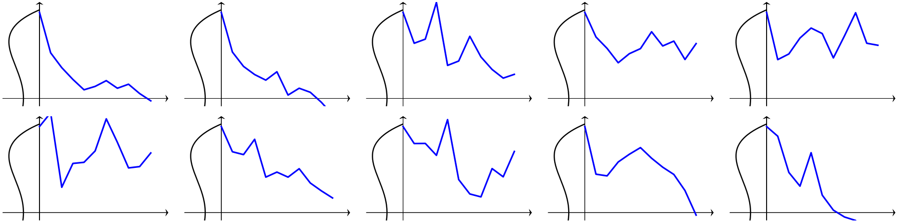
Trajectory of SGD X^η: $\eta = 1/10$ & noises are **heavy-tailed**

Typical Behavior of SGD

Trajectory of SGD X^{η}: $\eta = 1/10$ & noises are **light-tailed**

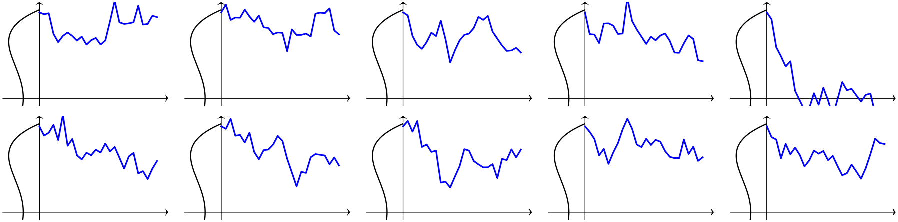


Trajectory of SGD X^{η}: $\eta = 1/10$ & noises are **heavy-tailed**

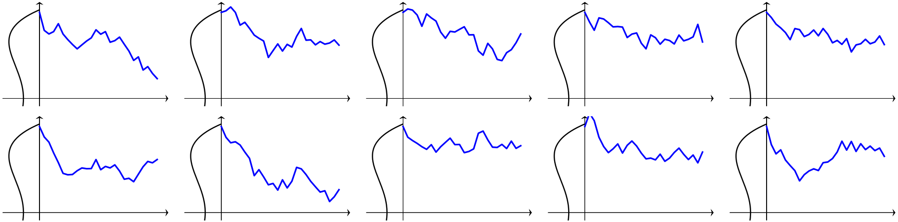


Typical Behavior of SGD

Trajectory of SGD X^η: $\eta = 1/25$ & noises are **light-tailed**

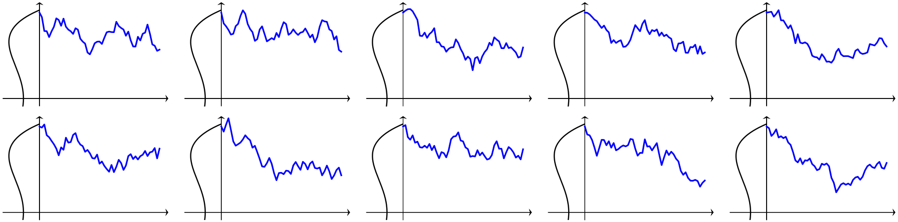


Trajectory of SGD X^η: $\eta = 1/25$ & noises are **heavy-tailed**

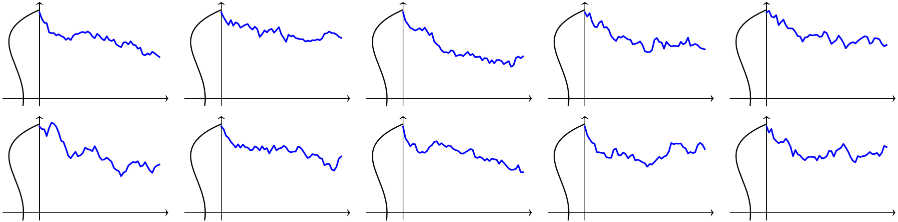


Typical Behavior of SGD

Trajectory of SGD X^η: $\eta = 1/50$ & noises are **light-tailed**

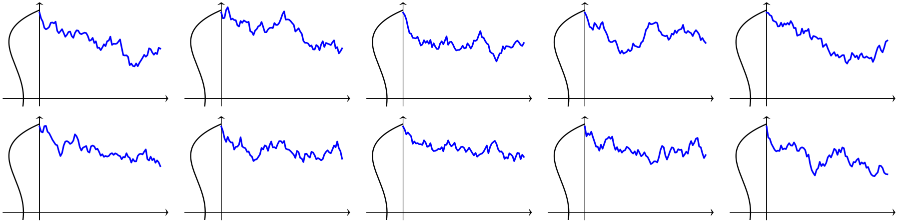


Trajectory of SGD X^η: $\eta = 1/50$ & noises are **heavy-tailed**

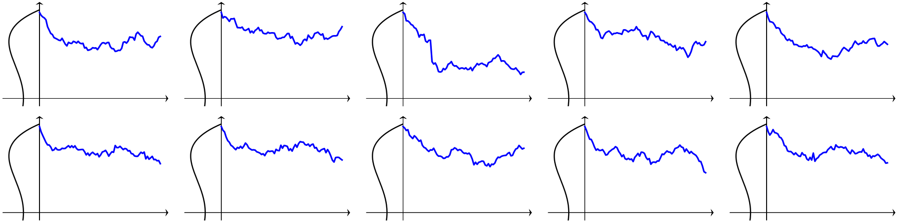


Typical Behavior of SGD

Trajectory of SGD X^η: $\eta = 1/75$ & noises are **light-tailed**

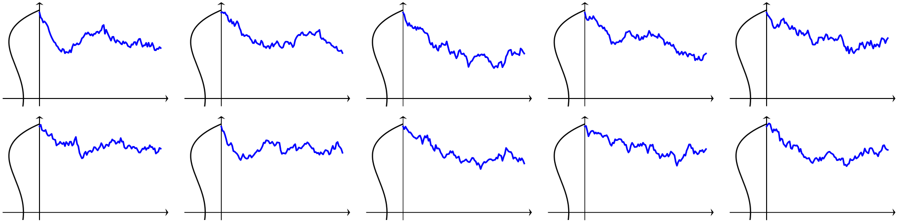


Trajectory of SGD X^η: $\eta = 1/75$ & noises are **heavy-tailed**

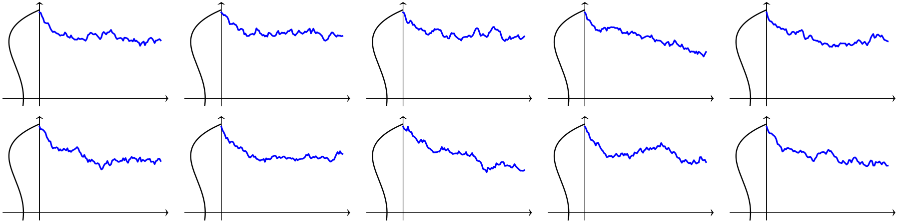


Typical Behavior of SGD

Trajectory of SGD X^{η}: $\eta = 1/100$ & noises are **light-tailed**

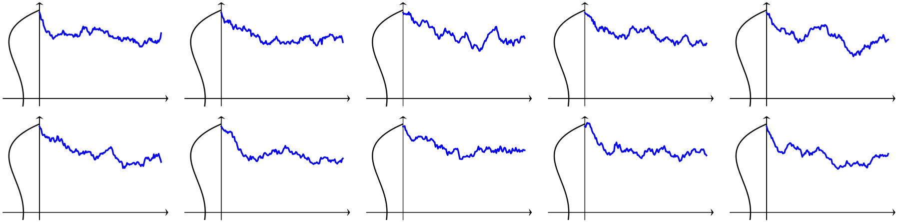


Trajectory of SGD X^{η}: $\eta = 1/100$ & noises are **heavy-tailed**

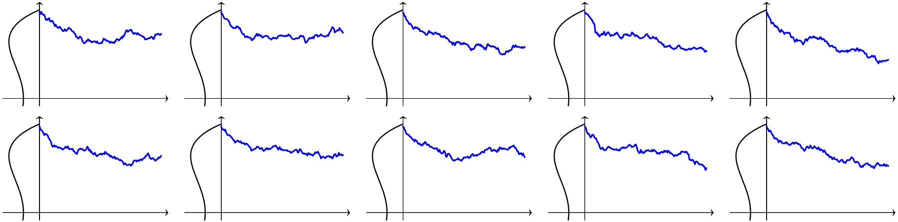


Typical Behavior of SGD

Trajectory of SGD X^η: $\eta = 1/150$ & noises are **light-tailed**

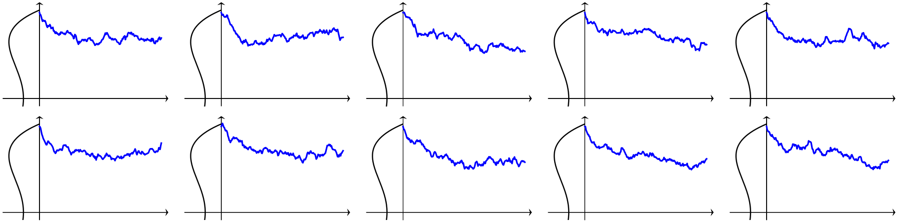


Trajectory of SGD X^η: $\eta = 1/150$ & noises are **heavy-tailed**

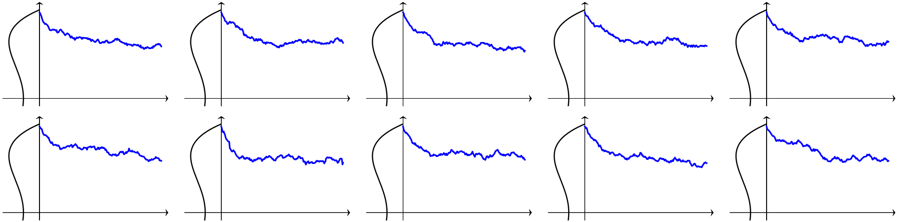


Typical Behavior of SGD

Trajectory of SGD X^{η}: $\eta = 1/200$ & noises are **light-tailed**



Trajectory of SGD X^{η}: $\eta = 1/200$ & noises are **heavy-tailed**



How does SGD escape local minima?

Catastrophe Principle in Heavy-tailed SGD

(Su, Wang, Rhee, 2021+) For "rare event" A,

Catastrophe Principle in Heavy-tailed SGD

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(X^\eta \in A) \to 0$ as $\eta \downarrow 0$)

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

- $\mathbb{P}(X^\eta \in A) \approx \eta^{(\alpha-1)I^*(A)}$

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

- $\mathbb{P}(X^\eta \in A) \approx \eta^{(\alpha-1)l^*(A)}$
- Conditioned on $\{X^\eta \in A\}$, X^η resembles piece-wise gradient flow with $l^*(A)$ jumps

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

- $\mathbb{P}(X^\eta \in A) \approx \eta^{(\alpha-1)I^*(A)}$

Typical Behavior

- Conditioned on $\{X^\eta \in A\}$, X^η resembles piece-wise $\boxed{\text{gradient flow}}$ with $I^*(A)$ jumps

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

- $\mathbb{P}(X^\eta \in A) \approx \eta^{(\alpha-1)I^*(A)}$

Typical Behavior — Catastrophes

- Conditioned on $\{X^\eta \in A\}$, X^η resembles piece-wise $\boxed{\text{gradient flow}}$ with $I^*(A)$ $\boxed{\text{jumps}}$

Catastrophe Principle in Heavy-tailed SGD

SGD path

(Su, Wang, Rhee, 2021+) For "rare event" A, (i.e. $\mathbb{P}(\boxed{X^\eta} \in A) \to 0$ as $\eta \downarrow 0$)

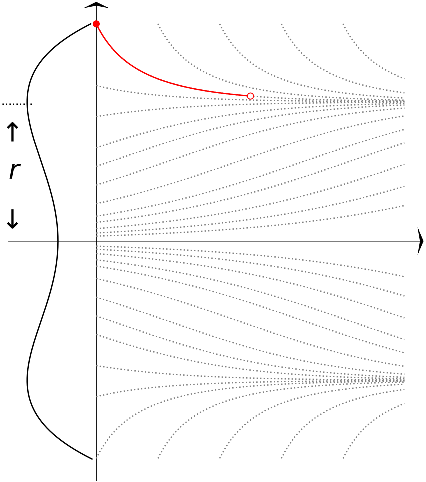
- $\mathbb{P}(X^\eta \in A) \approx \eta^{(\alpha-1)I^*(A)}$

  Typical Behavior ↘ ... ↙ Catastrophes

- Conditioned on $\{X^\eta \in A\}$, X^η resembles piece-wise $\boxed{\text{gradient flow}}$ with $I^*(A)$ $\boxed{\text{jumps}}$
- $I^*(A)$:Min # of jumps (catastrophes) to cause event A

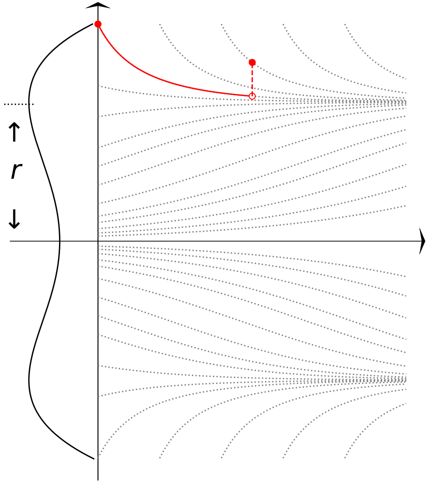
Catastrophe Principle Dictates SGD's Escape Route

This way?



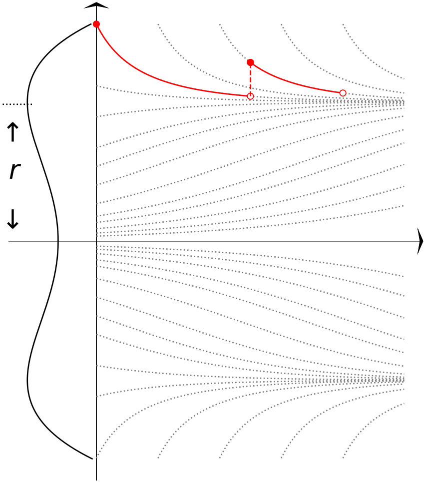
Catastrophe Principle Dictates SGD's Escape Route

This way?



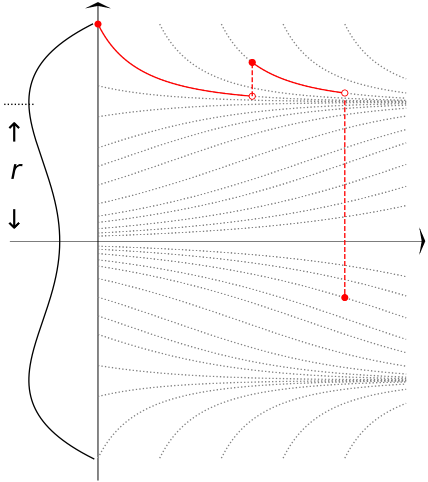
Catastrophe Principle Dictates SGD's Escape Route

This way?



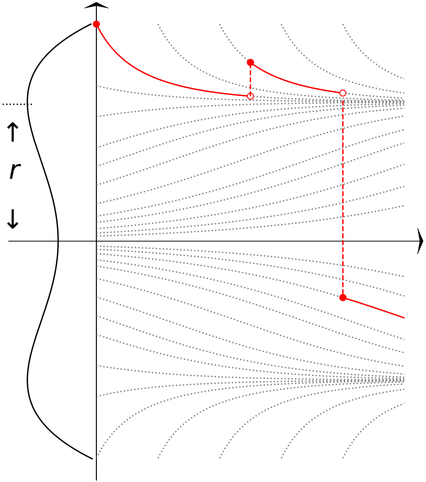
Catastrophe Principle Dictates SGD's Escape Route

This way?



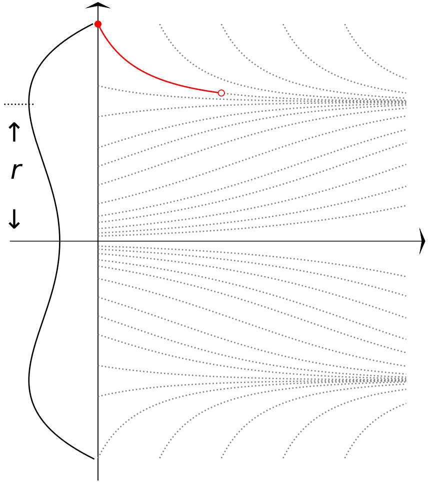
Catastrophe Principle Dictates SGD's Escape Route

This way?



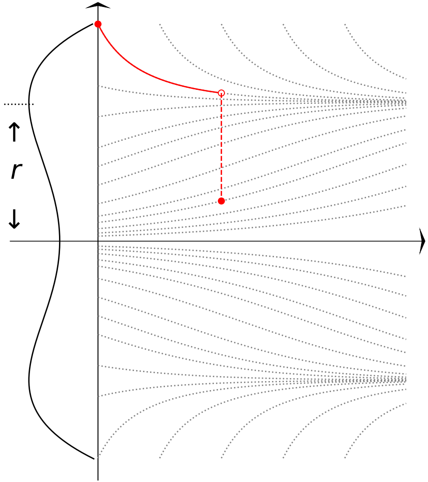
Catastrophe Principle Dictates SGD's Escape Route

This way?



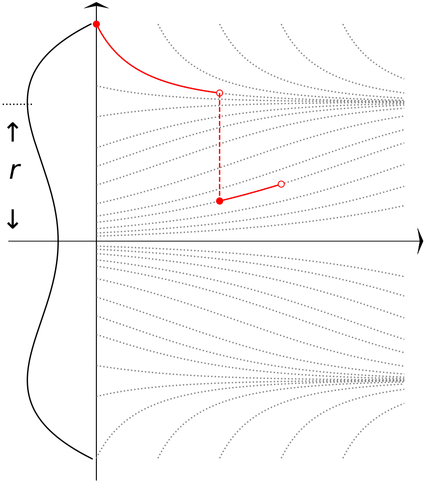
Catastrophe Principle Dictates SGD's Escape Route

This way?



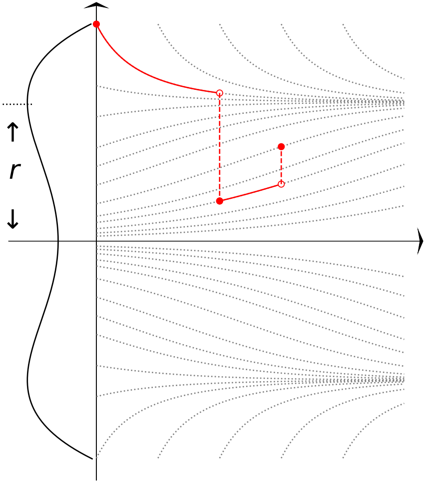
Catastrophe Principle Dictates SGD's Escape Route

This way?



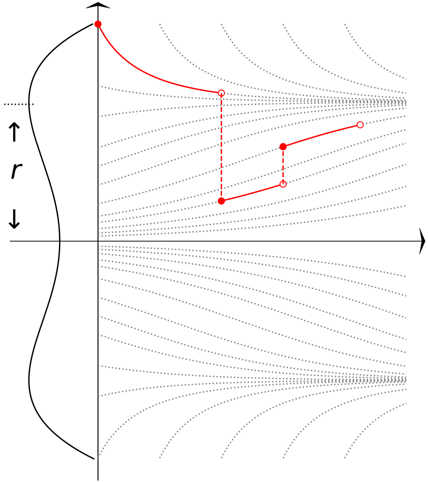
Catastrophe Principle Dictates SGD's Escape Route

This way?



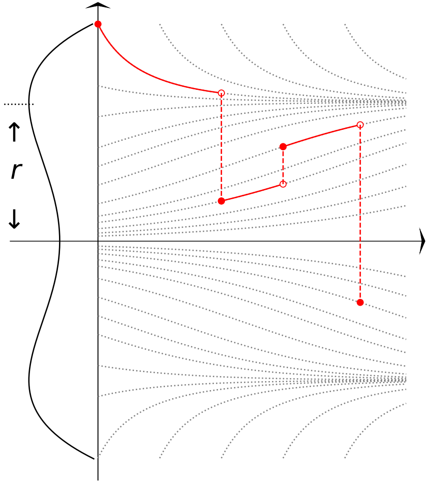
Catastrophe Principle Dictates SGD's Escape Route

This way?



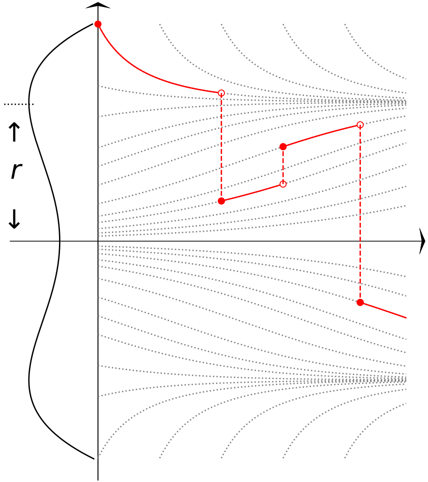
Catastrophe Principle Dictates SGD's Escape Route

This way?



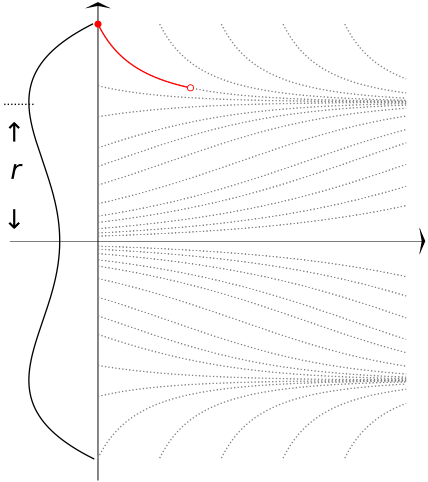
Catastrophe Principle Dictates SGD's Escape Route

This way?



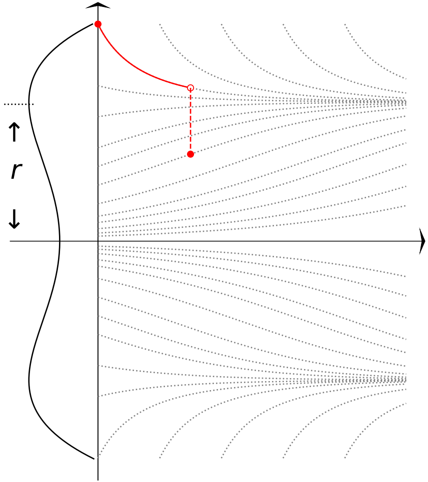
Catastrophe Principle Dictates SGD's Escape Route

This way?



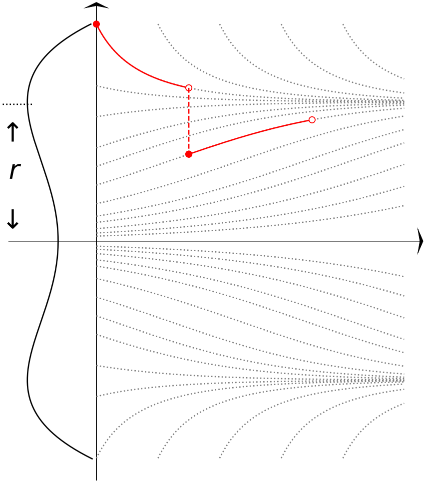
Catastrophe Principle Dictates SGD's Escape Route

This way?



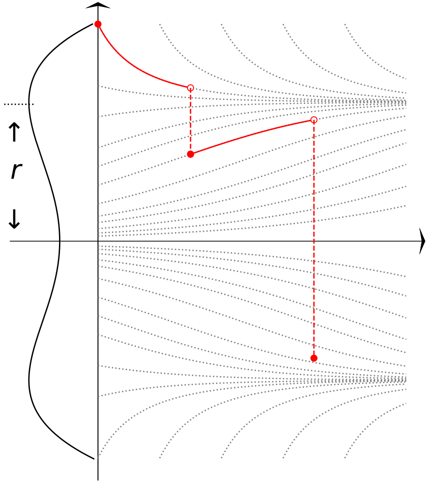
Catastrophe Principle Dictates SGD's Escape Route

This way?



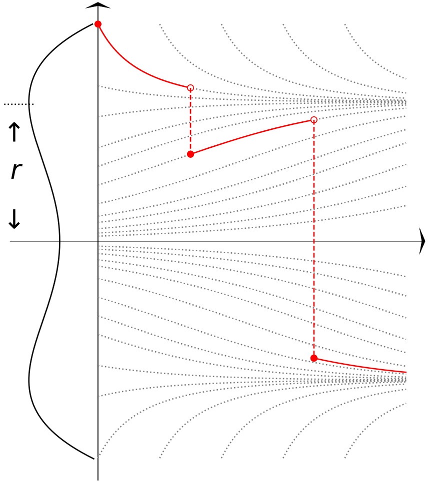
Catastrophe Principle Dictates SGD's Escape Route

This way?



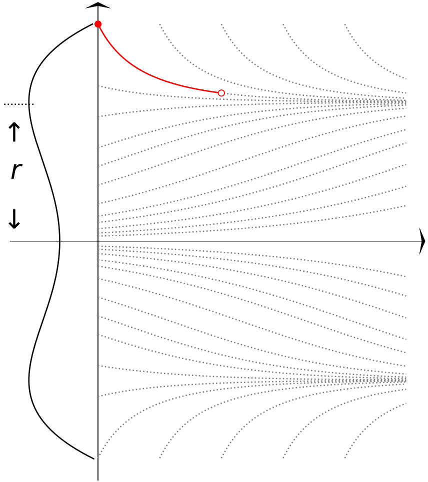
Catastrophe Principle Dictates SGD's Escape Route

This way?



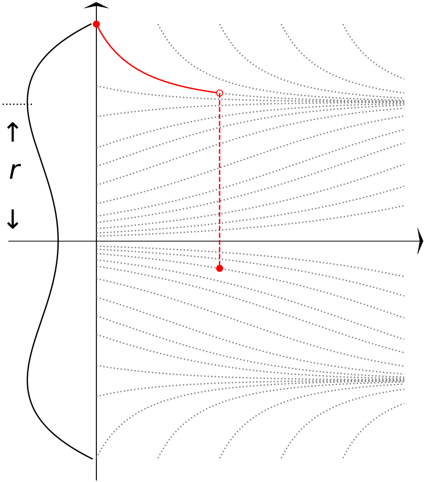
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



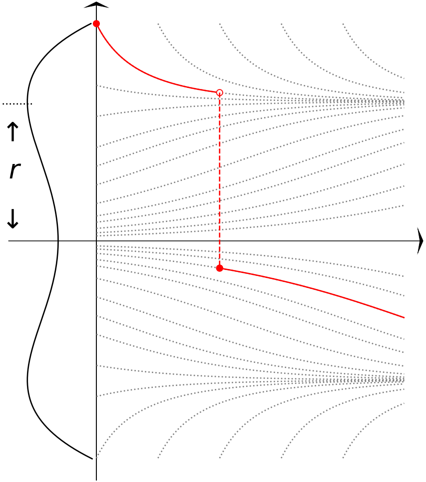
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



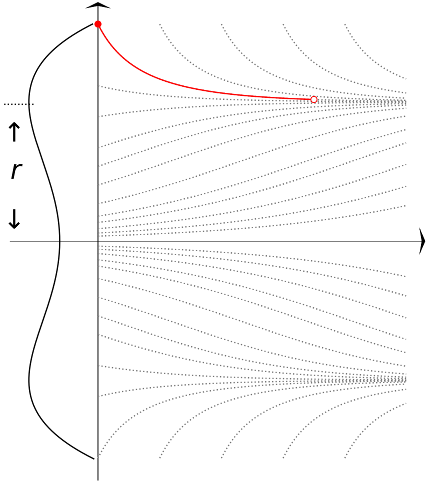
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



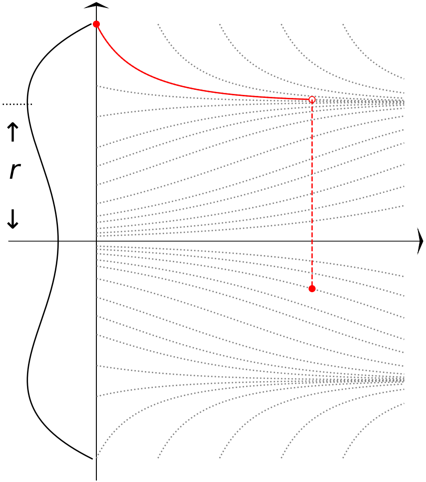
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



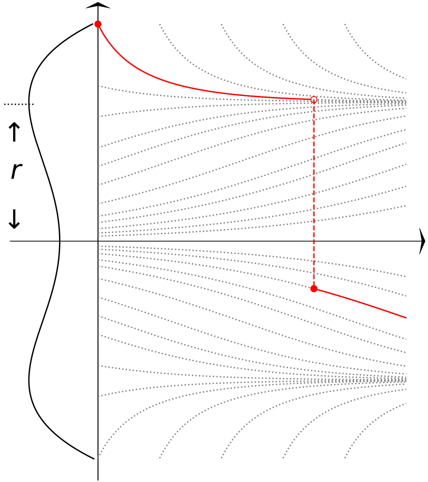
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



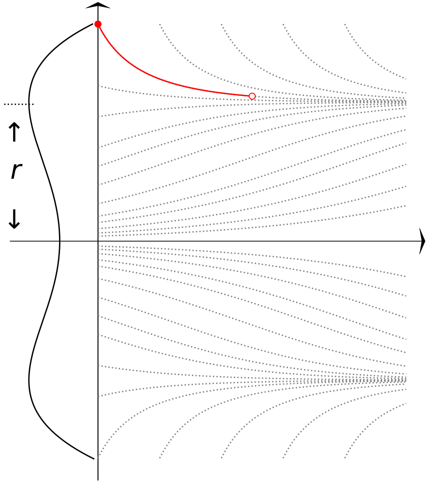
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



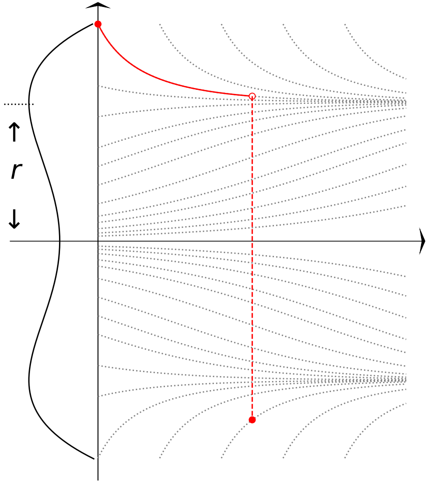
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



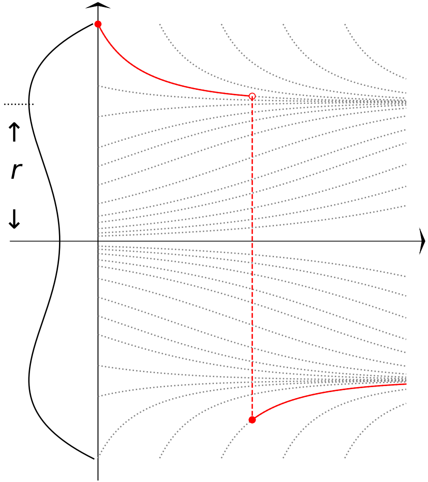
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump



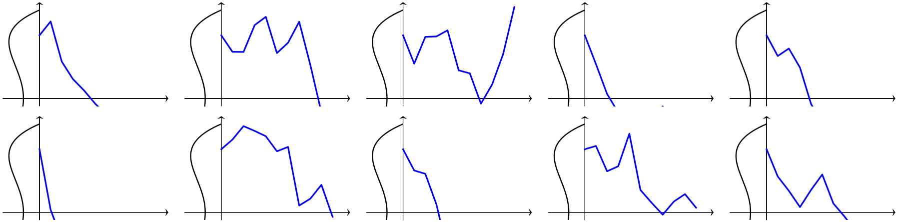
Catastrophe Principle Dictates SGD's Escape Route

Most likely path under heavy-tailed noises: with $l^* = 1$ jump

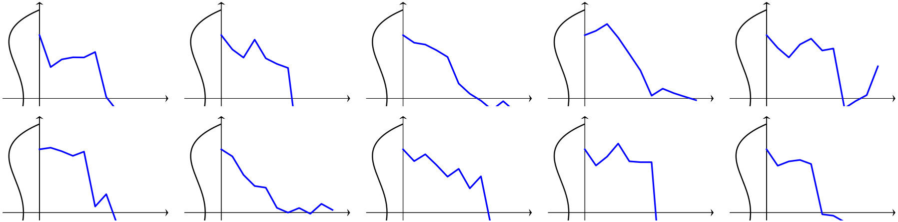


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^{η} conditional on exit: **light-tailed** noises with $\eta = 1/10$

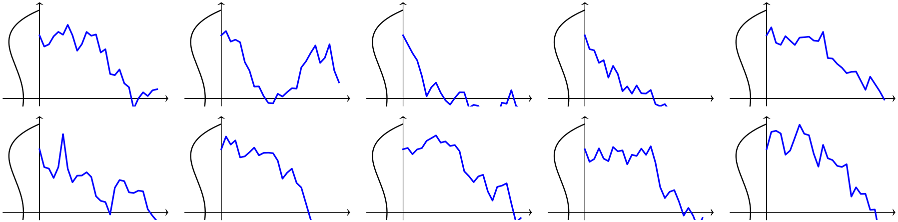


Trajectory of SGD X^{η} conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

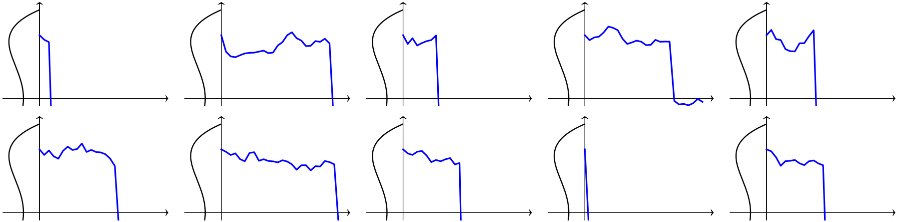


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/25$

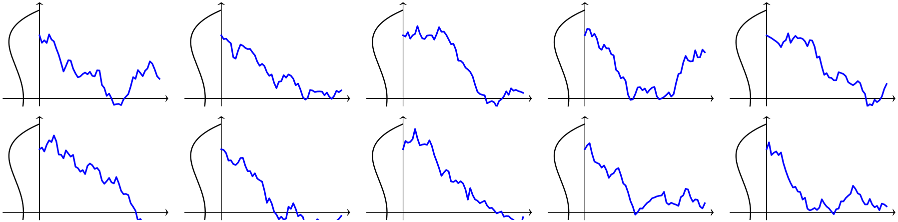


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

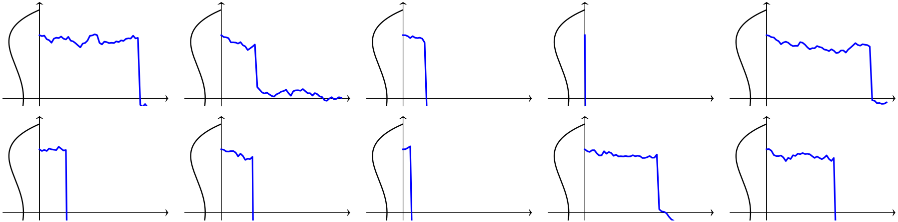


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/50$

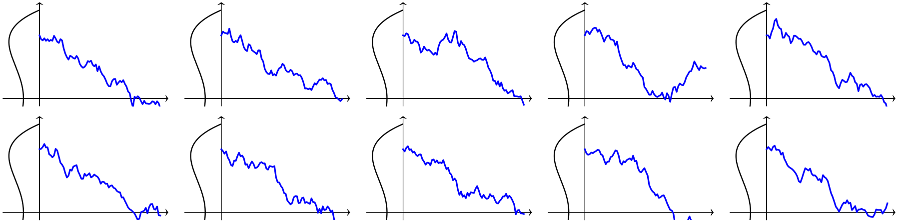


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/50$

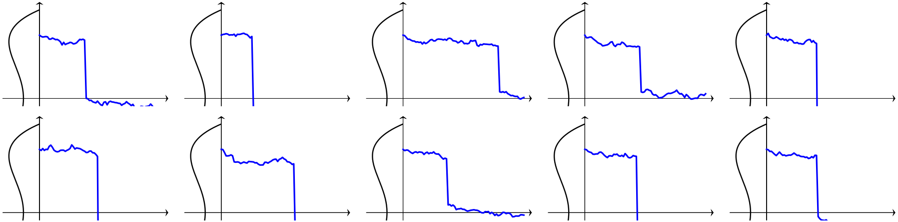


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/75$

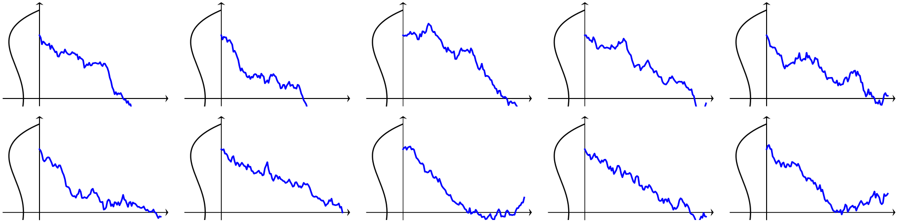


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

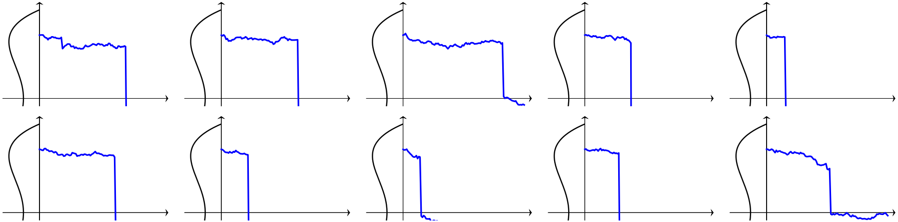


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/100$

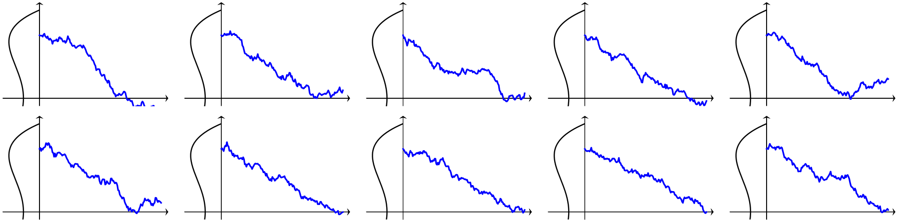


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

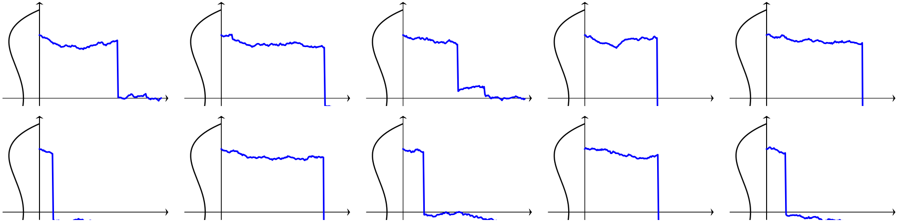


Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/150$

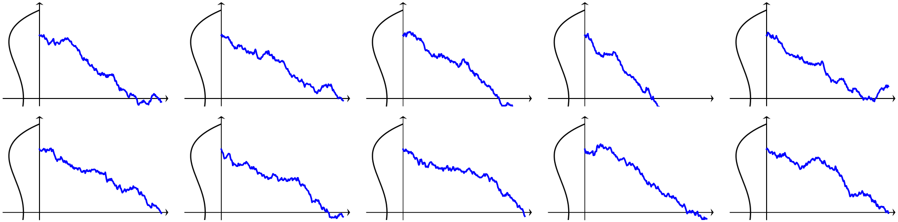


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

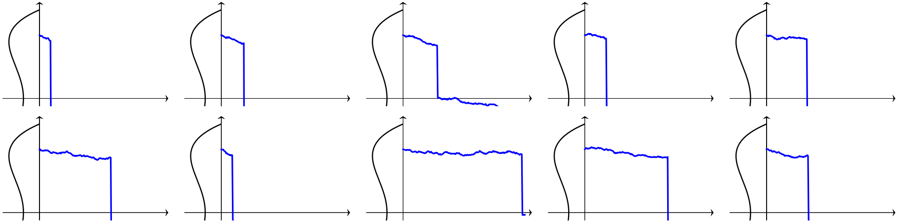


Catastrophe Principle Dictates SGD's Escape Route

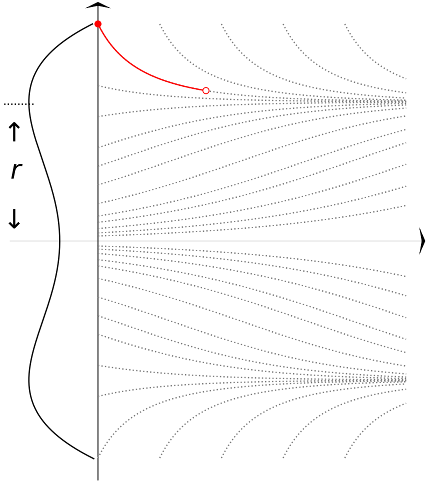
Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/200$



Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



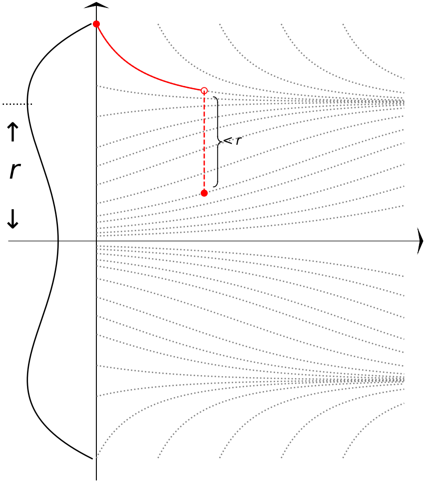
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

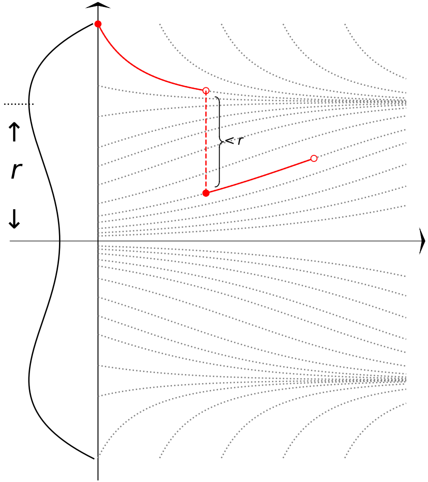
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

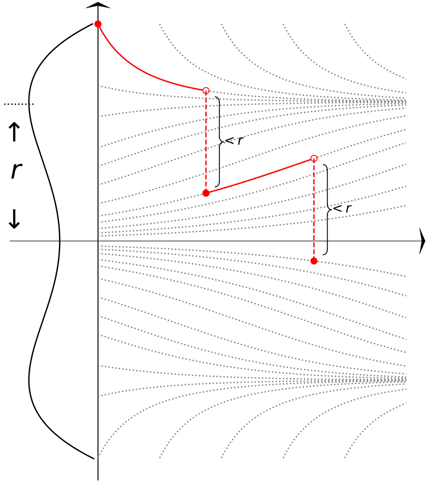
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^{\eta} = X_{j-1}^{\eta} + \varphi_b\big(-\eta\nabla f(X_{j-1}^{\eta}) + \eta Z_j\big),\ b \in (r/2, r)$$

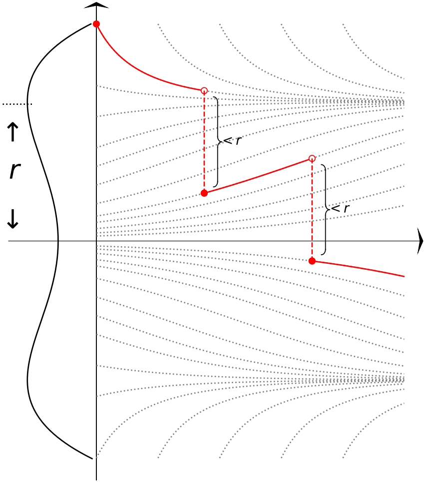
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

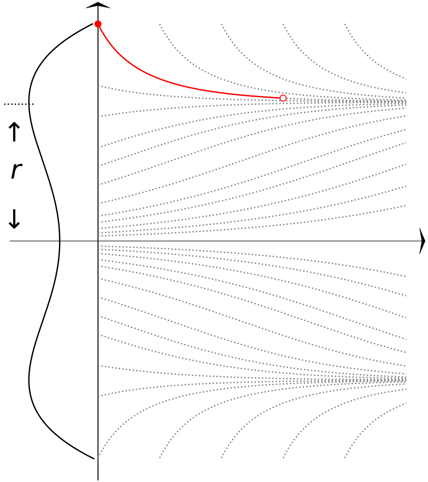
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

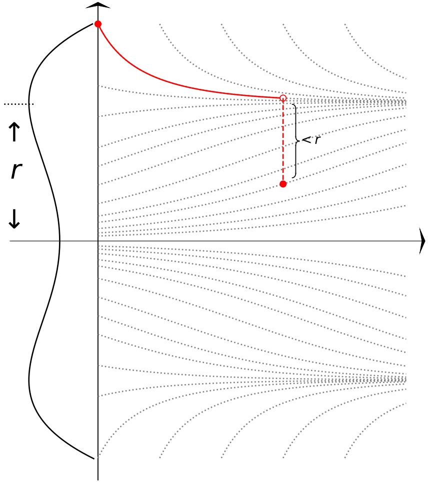
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

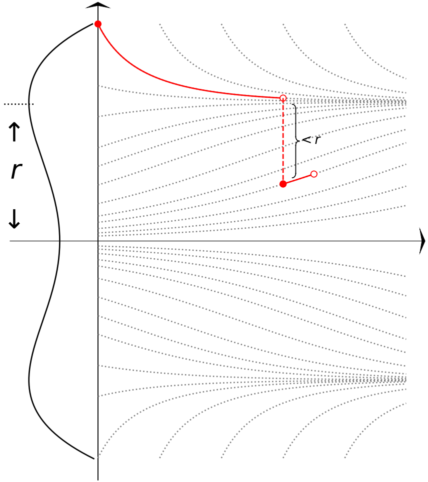
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

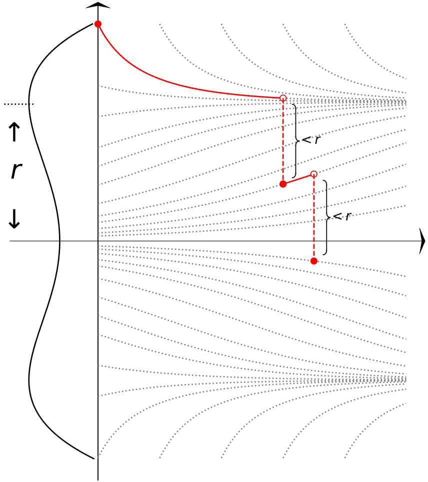
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta \nabla f(X_{j-1}^\eta) + \eta Z_j\big), \; b \in (r/2, r)$$

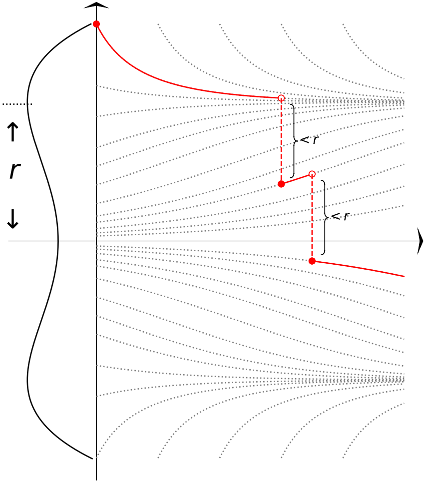
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

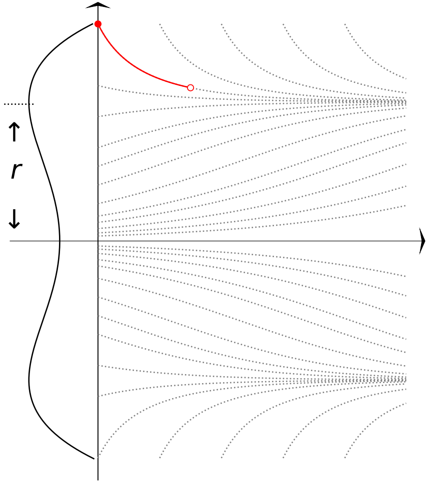
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

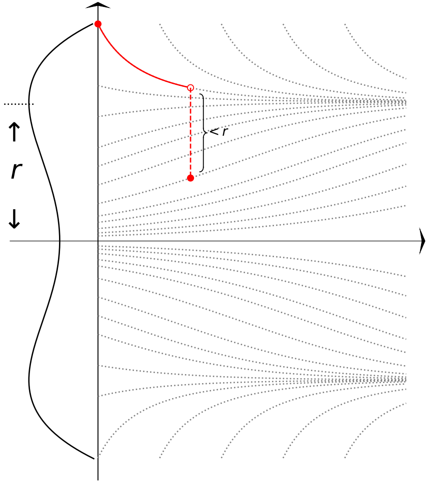
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

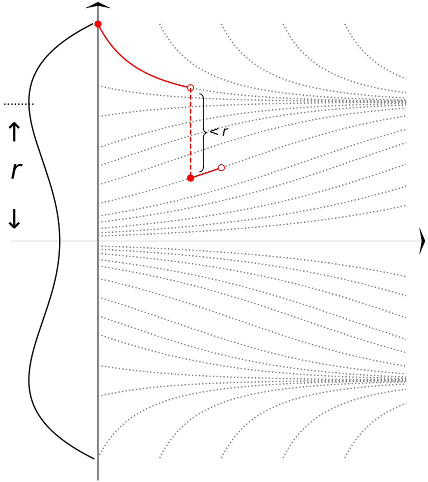
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

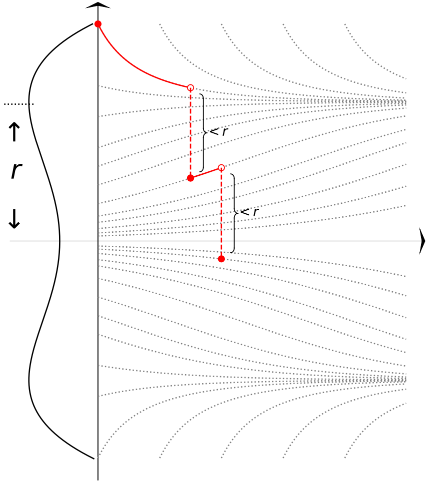
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

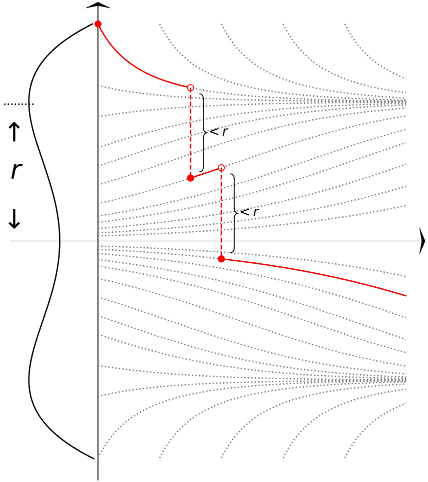
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

SGD's Escaping Route under Gradient Clipping

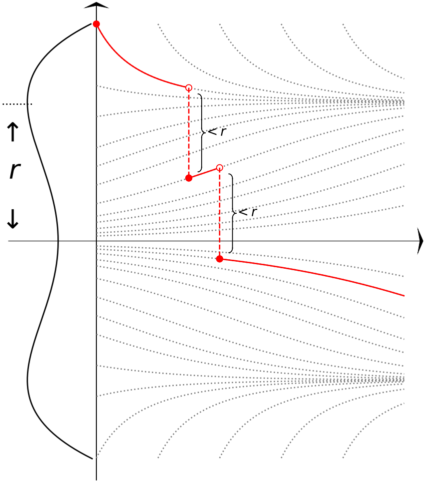


Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big), \; b \in (r/2, r)$$

SGD's Escaping Route under Gradient Clipping

Most likely path under heavy-tailed noises: with $l^* = 2$ jumps

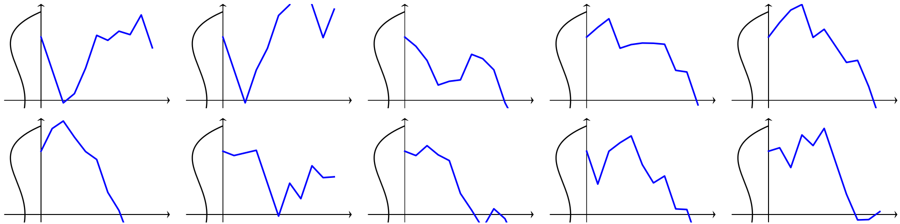


Clipping threshold

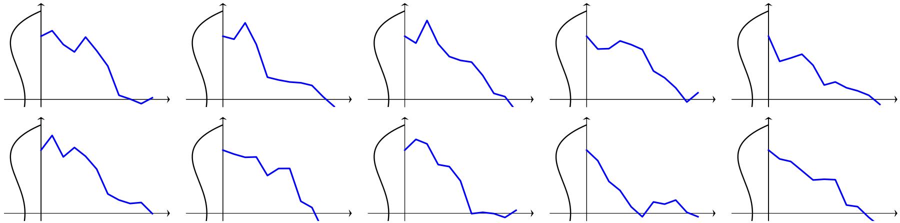
$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/2, r)$$

SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/10$

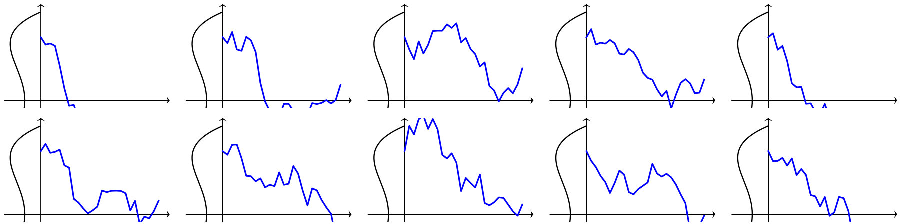


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

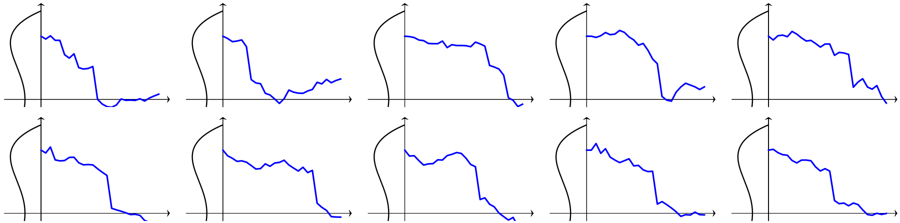


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/25$

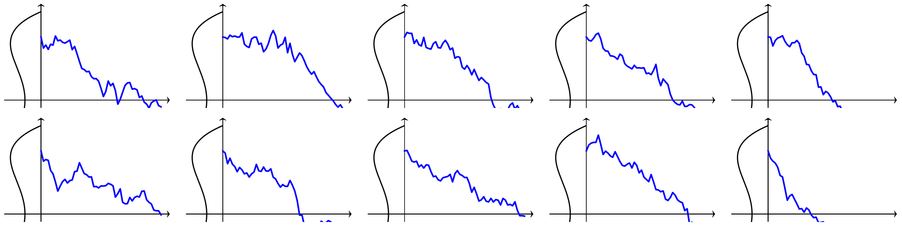


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

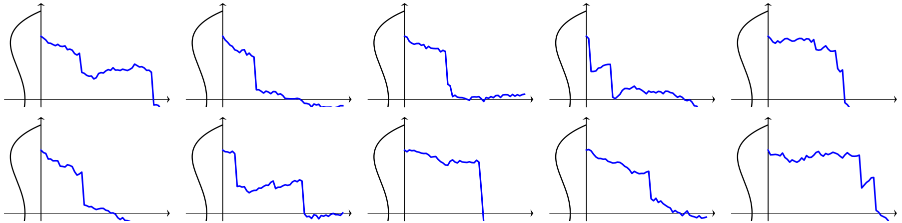


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/50$

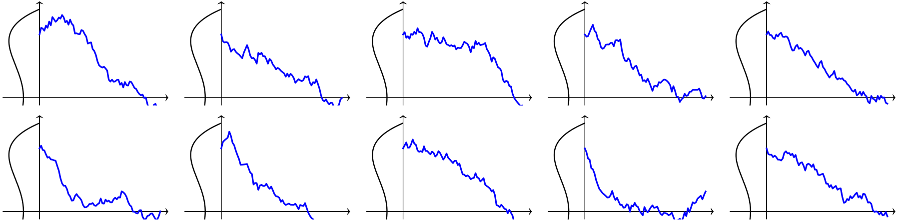


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

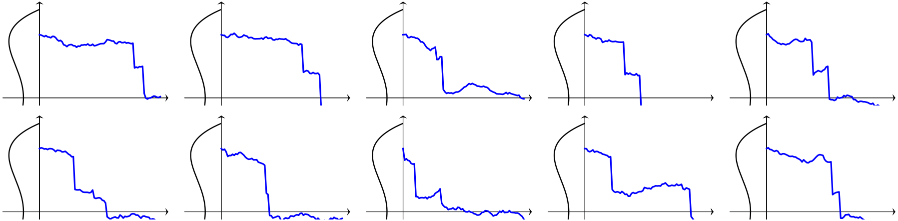


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/75$

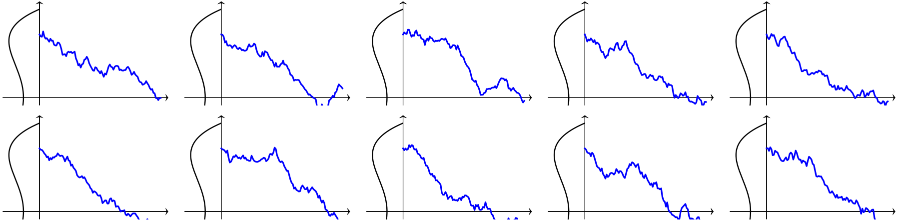


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

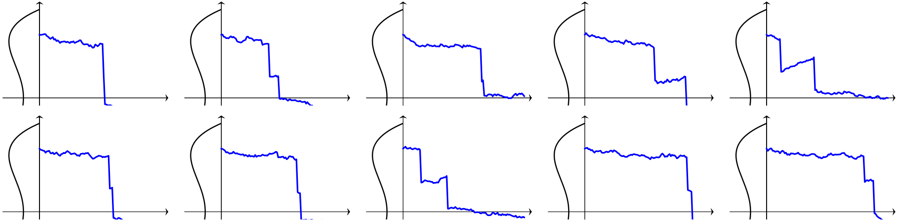


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/100$

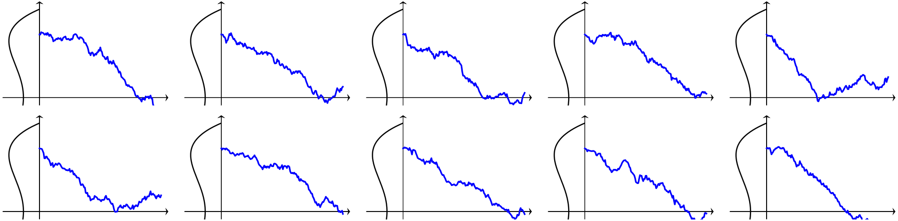


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

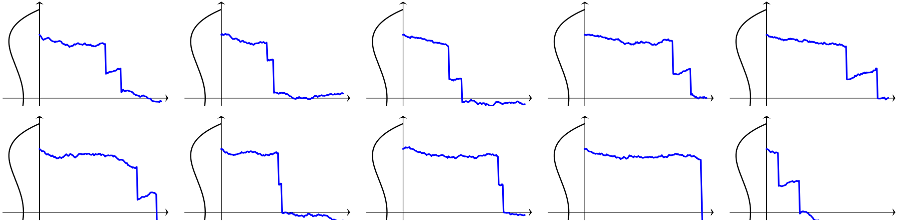


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/150$

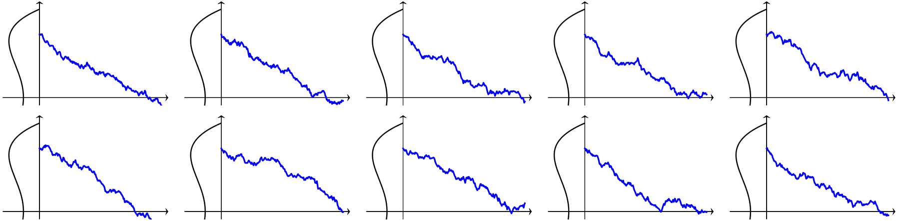


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

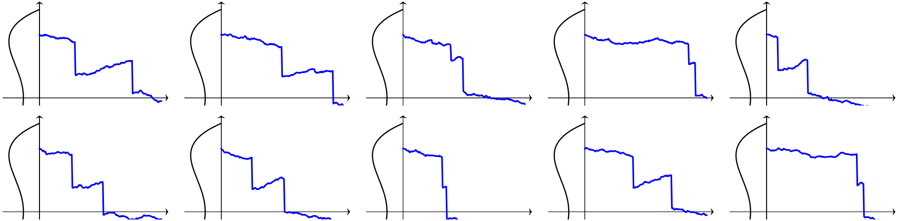


SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^{η} conditional on exit: **light-tailed** noises with $\eta = 1/200$

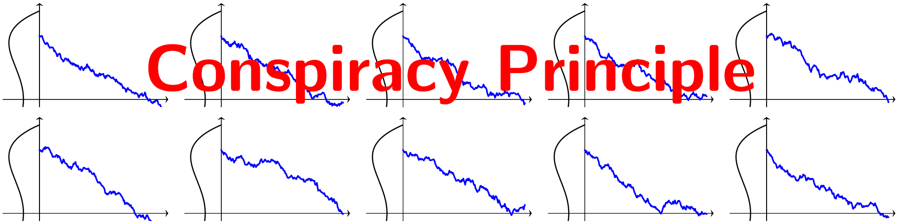


Trajectory of SGD X^{η} conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



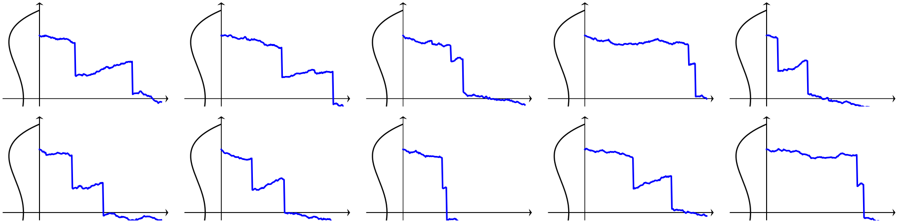
SGD's Escaping Route under Gradient Clipping

Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/200$



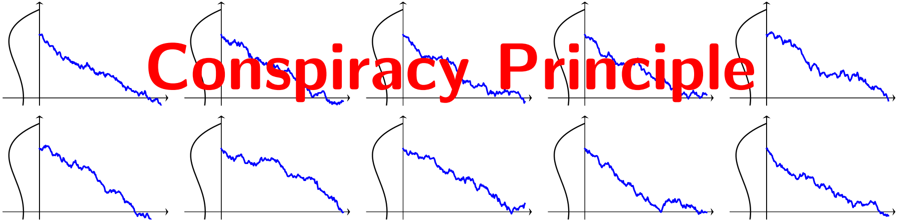
Conspiracy Principle

Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$

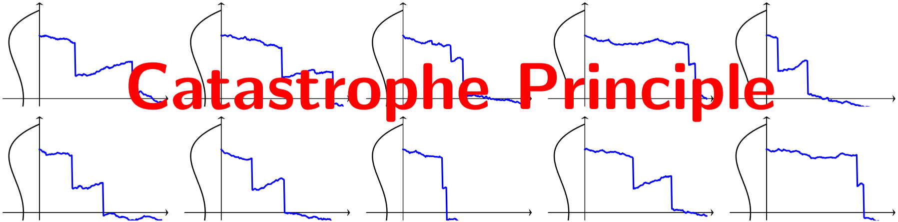


SGD's Escaping Route under Gradient Clipping

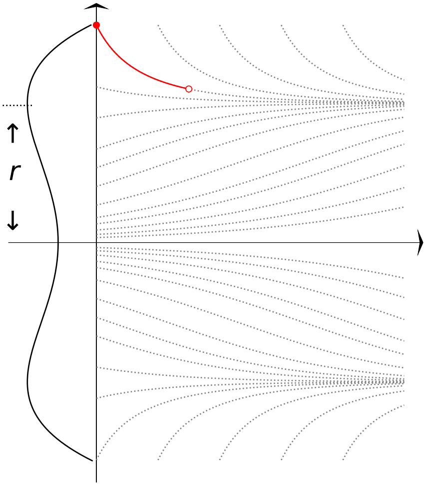
Trajectory of SGD X^η conditional on exit: **light-tailed** noises with $\eta = 1/200$



Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



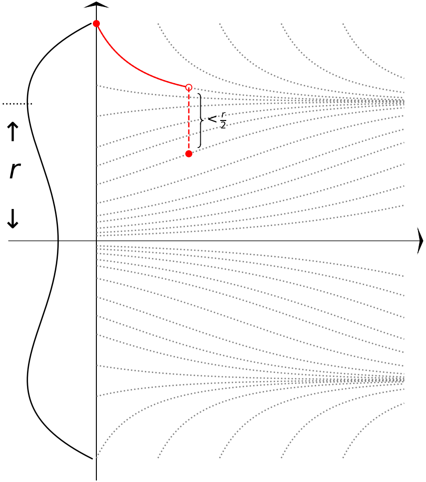
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/3, r/2)$$

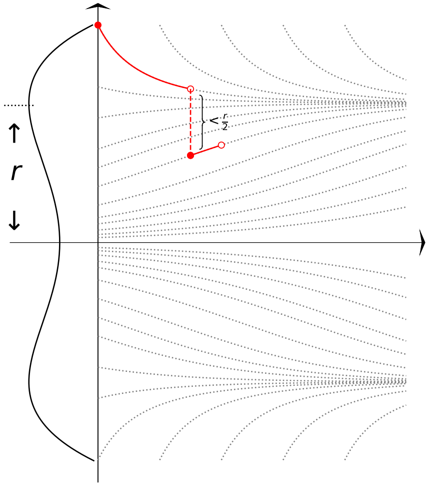
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/3, r/2)$$

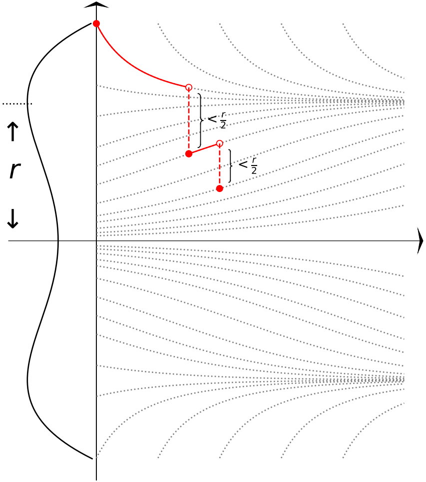
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big), \; b \in (r/3, r/2)$$

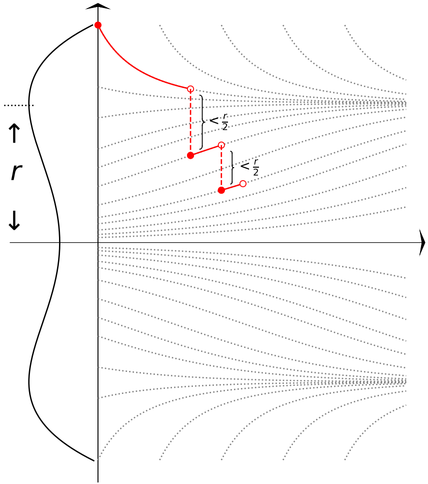
SGD's Escaping Route under Gradient Clipping



$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/3, r/2)$$

Clipping threshold

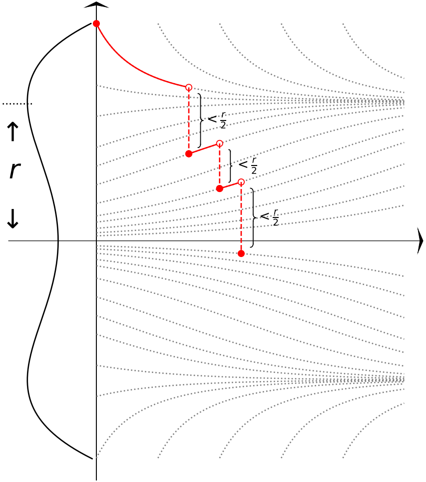
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/3, r/2)$$

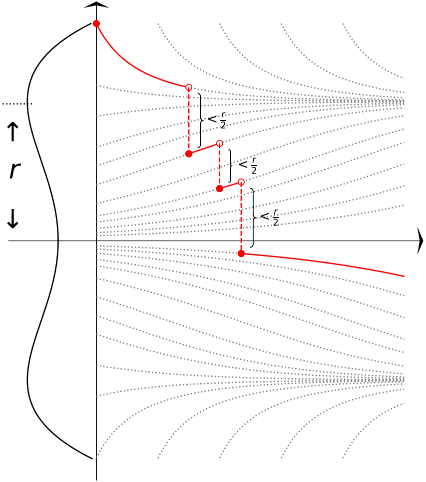
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/3, r/2)$$

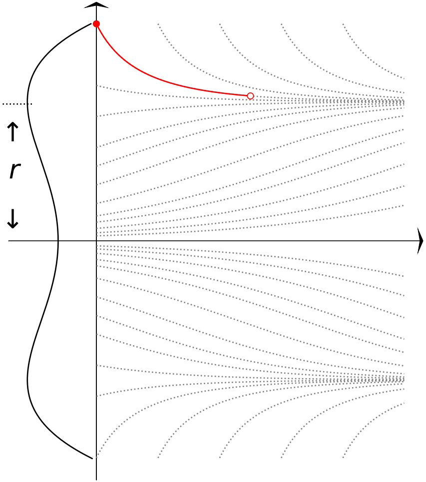
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta \nabla f(X_{j-1}^\eta) + \eta Z_j\big), \; b \in (r/3, r/2)$$

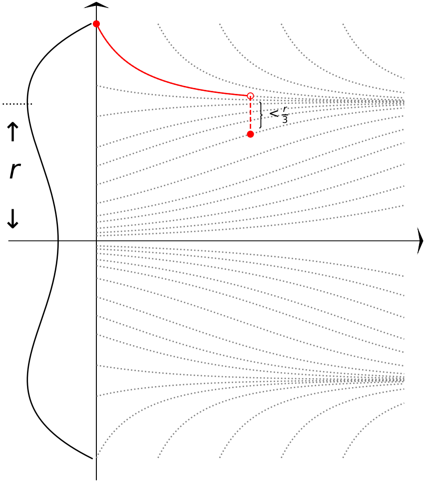
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta \nabla f(X_{j-1}^\eta) + \eta Z_j \big),\ b \in (r/4, r/3)$$

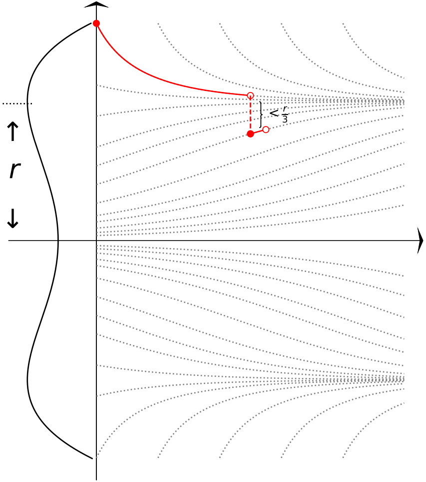
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/4, r/3)$$

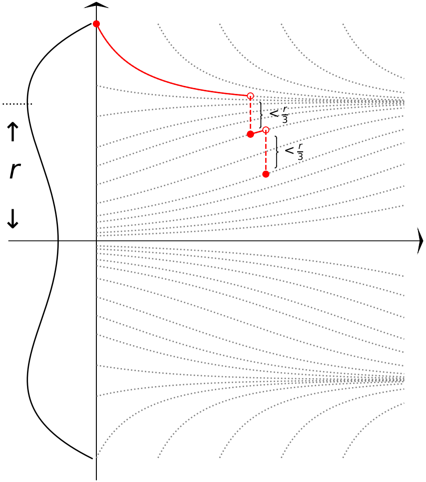
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/4, r/3)$$

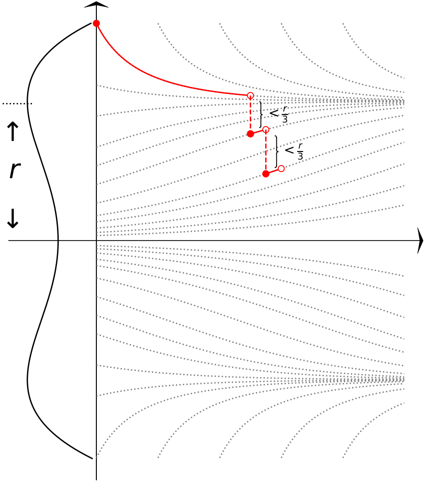
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta \nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/4, r/3)$$

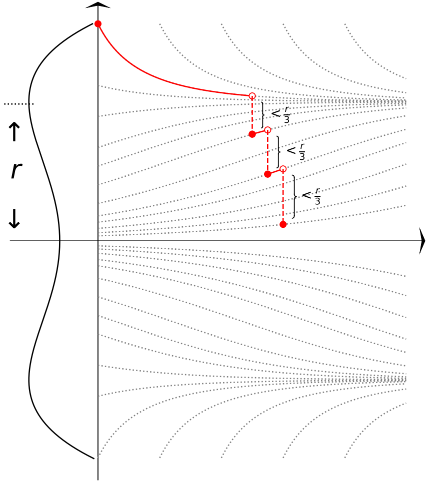
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/4, r/3)$$

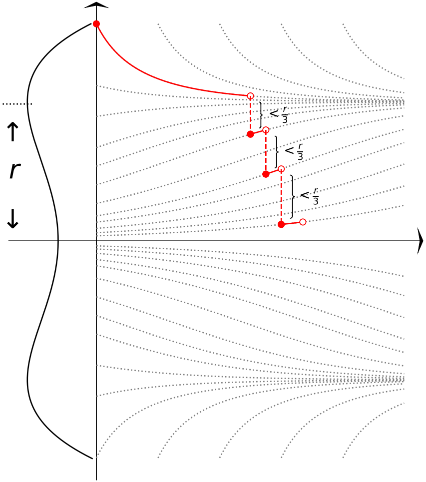
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^{\eta} = X_{j-1}^{\eta} + \varphi_b\big(-\eta \nabla f(X_{j-1}^{\eta}) + \eta Z_j \big),\ b \in (r/4, r/3)$$

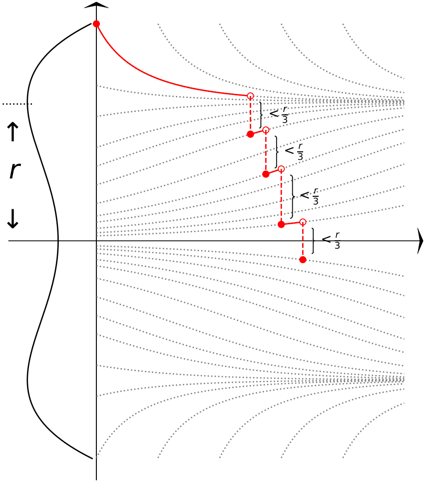
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta\nabla f(X_{j-1}^\eta) + \eta Z_j\big),\; b \in (r/4, r/3)$$

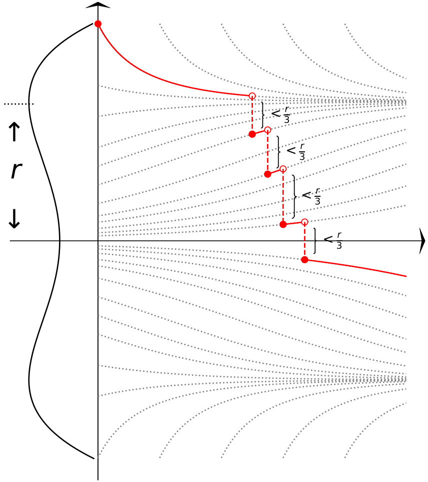
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^\eta = X_{j-1}^\eta + \varphi_b\big(-\eta \nabla f(X_{j-1}^\eta) + \eta Z_j\big),\ b \in (r/4, r/3)$$

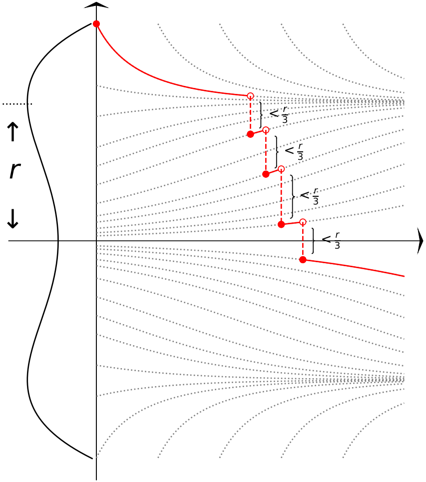
SGD's Escaping Route under Gradient Clipping



Clipping threshold

$$X_j^{\eta} = X_{j-1}^{\eta} + \varphi_b\big(-\eta\nabla f(X_{j-1}^{\eta}) + \eta Z_j\big),\ b \in (r/4, r/3)$$

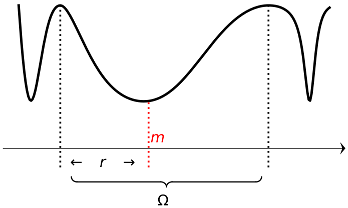
SGD's Escaping Route under Gradient Clipping



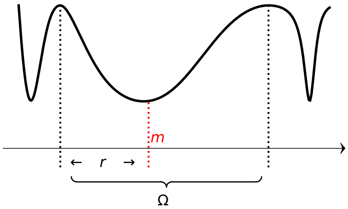
Clipping threshold

(Min # of jumps for escape) $l^* = \lceil r/b \rceil$

First Exit Time Analysis

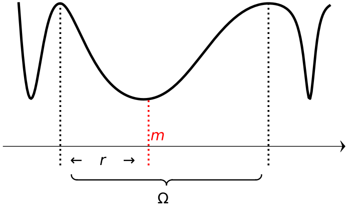


First Exit Time Analysis



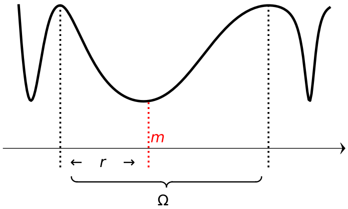
- **First Exit Time:** $\sigma^{\eta} \triangleq \min\{j \geq 0 : X^{\eta}_{j} \notin \Omega\}$

First Exit Time Analysis



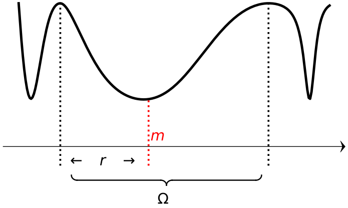
- **First Exit Time:** $\sigma^{\eta} \triangleq \min\{j \geq 0 : \ X_j^{\eta} \notin \Omega\}$
- **Effective Width** (Min Distance for Escape): $r \triangleq \inf_{x \notin \Omega} |x - m|$.

First Exit Time Analysis



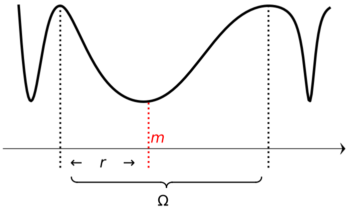
- **First Exit Time:** $\sigma^\eta \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- **Effective Width** (Min Distance for Escape): $r \triangleq \inf_{x \notin \Omega} |x - m|$.
- **Relative Width** (Min # of jumps for Escape): $l^* \triangleq \lceil r/b \rceil$.

First Exit Time Analysis



- **First Exit Time:** $\sigma^\eta \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- **Effective Width** (Min Distance for Escape): $r \triangleq \inf_{x \notin \Omega} |x - m|$.
- **Relative Width** (Min # of jumps for Escape): $l^* \triangleq \lceil r/b \rceil$.
- **(Wang, Oh, Rhee, 2021+)** As $\eta \downarrow 0$, $\sigma^\eta \lambda(\eta) \Rightarrow Exp(q)$.

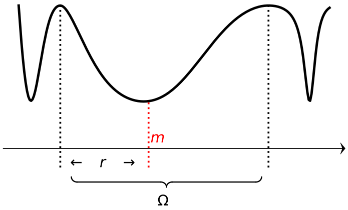
First Exit Time Analysis



- **First Exit Time:** $\sigma^\eta \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- **Effective Width** (Min Distance for Escape): $r \triangleq \inf_{x \notin \Omega} |x - m|$.
- **Relative Width** (Min # of jumps for Escape): $l^* \triangleq \lceil r/b \rceil$.
- **(Wang, Oh, Rhee, 2021+)** As $\eta \downarrow 0$, $\sigma^\eta \lambda(\eta) \Rightarrow Exp(q)$.

$$\big(\lambda(\eta) \approx O(\eta^{\alpha+(l^*-1)(\alpha-1)}), \text{ deterministic}\big)$$

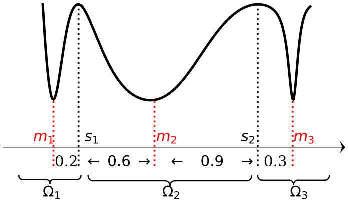
First Exit Time Analysis



- **First Exit Time:** $\sigma^\eta \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- **Effective Width** (Min Distance for Escape): $r \triangleq \inf_{x \notin \Omega} |x - m|$.
- **Relative Width** (Min # of jumps for Escape): $l^* \triangleq \lceil r/b \rceil$.

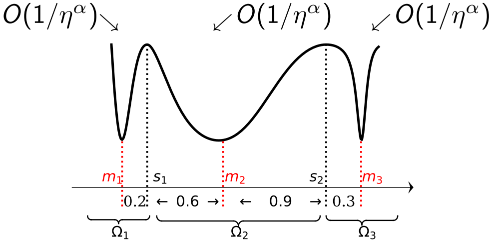
$$\sigma^\eta \sim O(1/\lambda(\eta)) \approx O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$$

Elimination of Narrow Minima



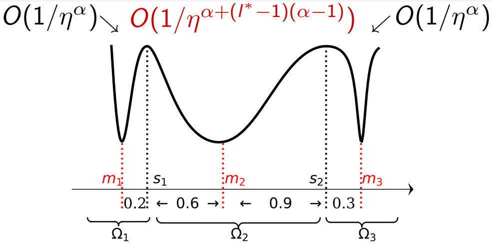
Without Clipping

Elimination of Narrow Minima



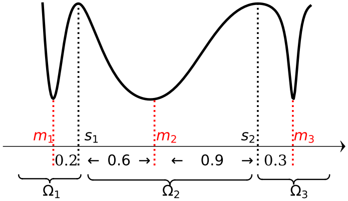
Without Clipping

Elimination of Narrow Minima



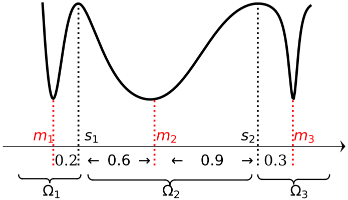
With Clipping

Elimination of Narrow Minima



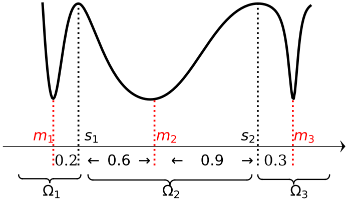
- **Min # of jumps for escape**: l_i^*

Elimination of Narrow Minima



- **Min # of jumps for escape**: l_i^* (Example: set $b = 0.5$)

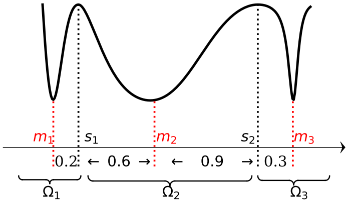
Elimination of Narrow Minima



$$l_1^* = 1, \quad l_2^* = 2, \quad l_3^* = 1$$

- **Min # of jumps for escape**: l_i^* (Example: set $b = 0.5$)

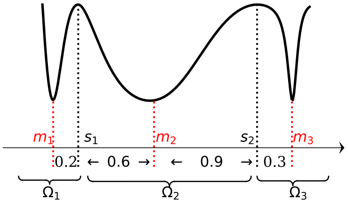
Elimination of Narrow Minima



$$l_1^* = 1, \quad l_2^* = 2, \quad l_3^* = 1$$

- **Min # of jumps for escape**: l_i^* (Example: set $b = 0.5$)
- **Set of Widest Minima:** $m_i \in M^{\text{wide}}$ iff $l_i^* = \max_j l_j^*$.

Elimination of Narrow Minima



$l_1^* = 1, \quad l_2^* = 2, \quad l_3^* = 1$

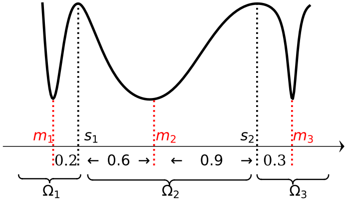
- **Min # of jumps for escape**: l_i^* (Example: set $b = 0.5$)
- **Set of Widest Minima:** $m_i \in M^{\text{wide}}$ iff $l_i^* = \max_j l_j^*$.

Theorem (Wang, Oh, Rhee, 2021+)

Under structural conditions on loss landscape, for any $t > 0$ and $\beta > 1 + (\alpha - 1)\max_i l_i^*$,

$$\frac{1}{\lfloor t/\eta^\beta \rfloor} \int_0^{\lfloor t/\eta^\beta \rfloor} 1\Big\{X^\eta_{\lfloor u \rfloor} \in \bigcup_{j: m_j \notin M^{wide}} \Omega_j\Big\} du \xrightarrow{\text{P}} 0 \text{ as } \eta \downarrow 0.$$

Elimination of Narrow Minima



$$l_1^* = 1, \quad l_2^* = 2, \quad l_3^* = 1$$

- **Min # of jumps for escape**: l_i^* (Example: set $b = 0.5$)
- **Set of Widest Minima:** $m_i \in M^{\text{wide}}$ iff $l_i^* = \max_j l_j^*$.

Theorem (Wang, Oh, Rhee, 2021+)

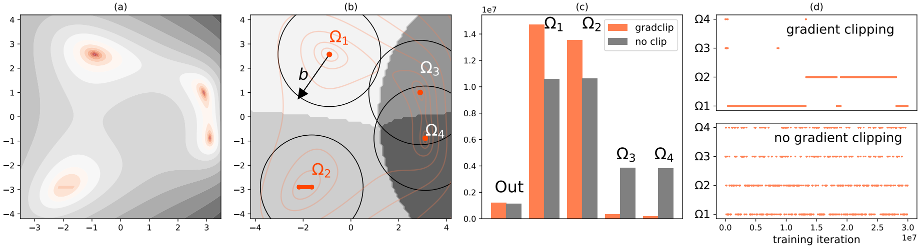
Under structural conditions on loss landscape, for any $t > 0$ and $\beta > 1 + (\alpha - 1)\max_i l_i^*$,

$$\frac{1}{\lfloor t/\eta^\beta \rfloor} \int_0^{\lfloor t/\eta^\beta \rfloor} \mathbb{1}\Big\{ X^\eta_{\lfloor u \rfloor} \in \bigcup_{j: m_j \notin M^{\text{wide}}} \Omega_j \Big\} du \xrightarrow{\text{P}} 0 \text{ as } \eta \downarrow 0.$$

↖Proportion of time at narrow minima

$\mathbb{R}^d$ Case

- Same Elimination Effect in $\mathbb{R}^d$



New Training Algorithm

Truncated Heavy-tailed SGD in Deep Learning

- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights;
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights;

  Gradient Clipping ↙
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights;
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + \text{"Heavy-tailed Noise"}$$

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + \text{“Heavy-tailed Noise”}$$

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + \text{"Heavy-tailed Noise"}$$

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + \text{“Heavy-tailed Noise”}$$

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$
- **Heavy-tail Inflation:** $Z\big(g_{\text{SB}}(X) - g_{\text{GD}}(X)\big)$ for some heavy-tailed Z

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + Z\Big(- g_{\text{GD}}(X) + g_{\text{SB}*}(X)\Big)$$

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$
- **Heavy-tail Inflation:** $Z\big(g_{\text{SB}}(X) - g_{\text{GD}}(X)\big)$ for some heavy-tailed Z

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + Z\Big(- g_{\text{LB}}(X) + g_{\text{SB}*}(X)\Big)$$

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$
- **Heavy-tail Inflation:** $Z\big(g_{\text{SB}}(X) - g_{\text{GD}}(X)\big)$ for some heavy-tailed Z

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + Z\big(- g_{\text{LB}}(X) + g_{\text{SB*}}(X)\big)$$

Same or independent batches?

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$
- **Heavy-tail Inflation:** $Z\big(g_{\text{SB}}(X) - g_{\text{GD}}(X)\big)$ for some heavy-tailed Z

Truncated Heavy-tailed SGD in Deep Learning

- X: current weights; **GD**: gradient descent; **SB**: small batch; g_{XX}: gradient under method XX.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SB}}(X) + Z\Big(- g_{\text{LB}}(X) + g_{\text{SB*}}(X)\Big)$$

$\nwarrow$ $\nearrow$

Same or independent batches? $\Rightarrow$two versions

- **Gradient noise:** $g_{\text{SB}}(X) - g_{\text{GD}}(X)$
- **Heavy-tail Inflation:** $Z\big(g_{\text{SB}}(X) - g_{\text{GD}}(X)\big)$ for some heavy-tailed Z

Experiments

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

Experiments

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Expected Sharpness:** Zhu et al. (2019); Neyshabur et al. (2017b)

Experiments

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Expected Sharpness:** Zhu et al. (2019); Neyshabur et al. (2017b)
  - Consistent results under other sharpness metrics

Experiments

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Expected Sharpness:** Zhu et al. (2019); Neyshabur et al. (2017b)
  - Consistent results under other sharpness metrics
- **Flatter geometry & Improved generalization performance**

Experiments

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Expected Sharpness:** Zhu et al. (2019); Neyshabur et al. (2017b)
  - Consistent results under other sharpness metrics
- **Flatter geometry & Improved generalization performance**
- Requires both **heavy-tailed** noise and **truncation**

Experiments

CIFAR10-VGG11	SB + Clip	Our 1	Our 2
Test Accuracy	89.54%	**90.76%**	90.45%
Expected Sharpness	0.167	**0.085**	0.096
PAC-Bayes Sharpness	1.31×10^4	$\mathbf{9 \times 10^3}$	10^4
Maximal Sharpness	1.66×10^4	1.29×10^4	$\mathbf{1.22 \times 10^4}$
CIFAR100-VGG16	**SB + Clip**	**Our 1**	**Our 2**
Test Accuracy	56.32%	**65.44%**	62.99%
Expected Sharpness	0.857	**0.441**	0.479
PAC-Bayes Sharpness	2.49×10^4	$\mathbf{1.9 \times 10^4}$	1.98×10^4
Maximal Sharpness	2.75×10^4	$\mathbf{2.12 \times 10^4}$	2.16×10^4

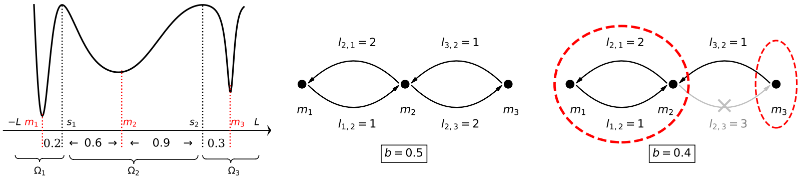
- **More training techniques:** Data augmentation, learning rate scheduler.

Conclusion

- **Theoretical Contribution**
  - Rigorously established that truncated heavy-tailed noises can eliminate sharp minima
  - Catastrophe principle, first exit time analysis, and metastability for heavy-tailed SGD
- **Algorithmic Contribution**
  - Proposed a tail-inflation strategy to find flatter solution with better generalization

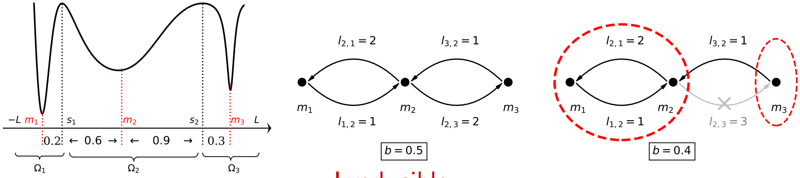
Remarks on Technical Results

- **"Regularity conditions"**



Remarks on Technical Results

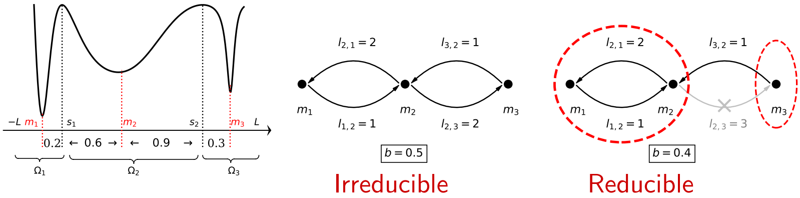
- **“Regularity conditions”**



Irreducible

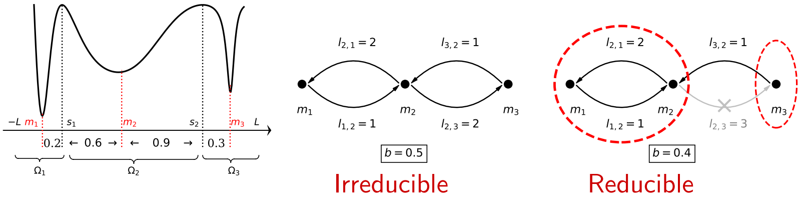
Remarks on Technical Results

- **"Regularity conditions"**



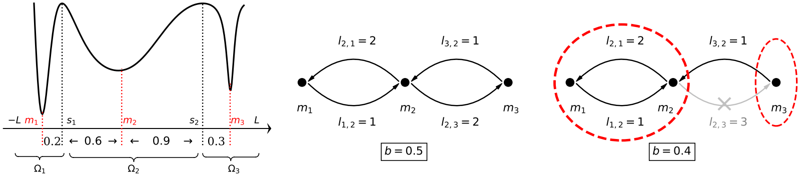
Remarks on Technical Results

- **"Regularity conditions"**: Irreducibility



Remarks on Technical Results

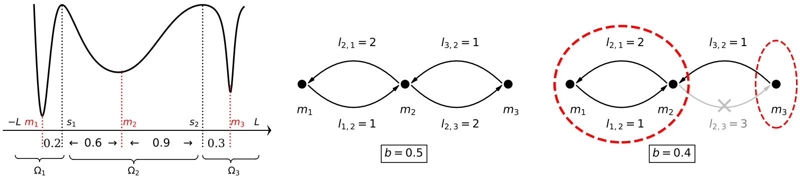
- **"Regularity conditions"**: Irreducibility



  - We established similar results for the reducible case.

Remarks on Technical Results

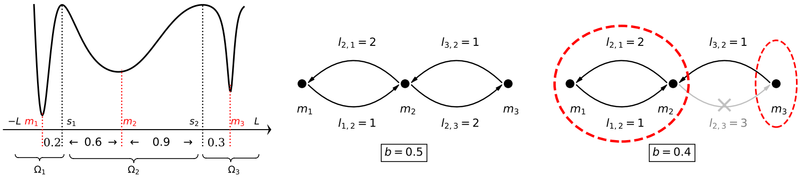
- **"Regularity conditions"**: Irreducibility



  - We established similar results for the reducible case.
- $\mathbb{R}^d$ **Extension**
  - First exit time results in $\mathbb{R}^d$

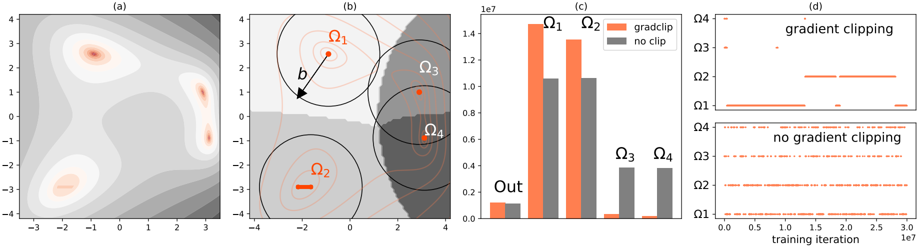
Remarks on Technical Results

- **“Regularity conditions”**: Irreducibility



  - We established similar results for the reducible case.

- $\mathbb{R}^d$ **Extension**
  - First exit time results in $\mathbb{R}^d$
  - $\mathbb{R}^d$ simulation experiments



First Exit Time Analysis

- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis

- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

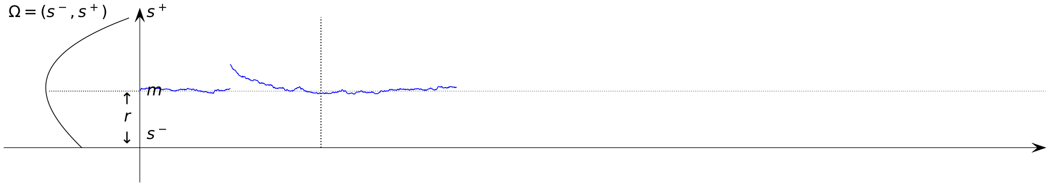
First Exit Time Analysis

- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis

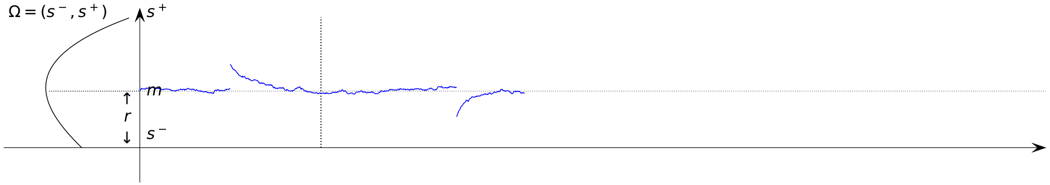
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



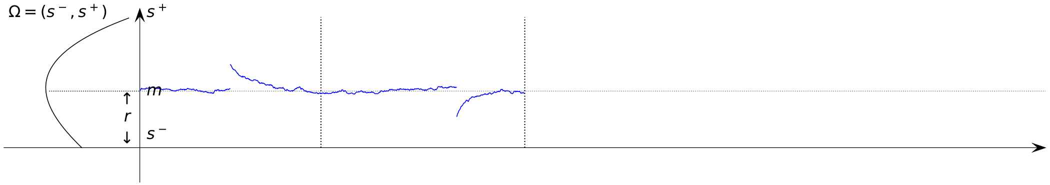
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



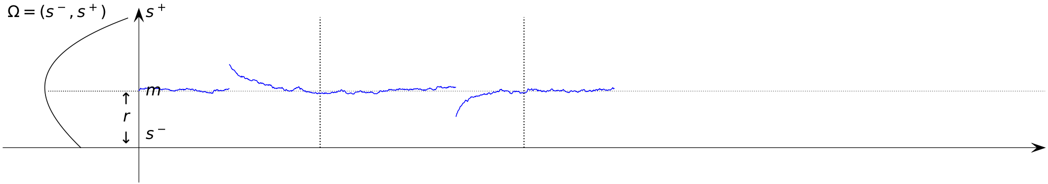
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



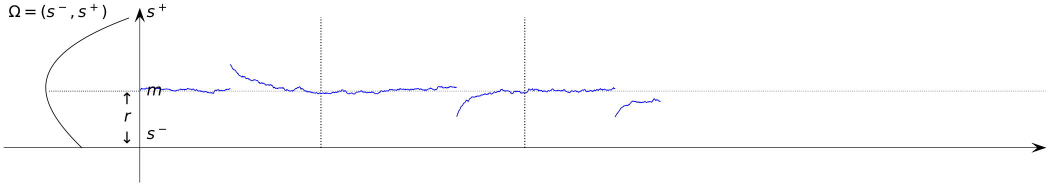
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 :\ X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



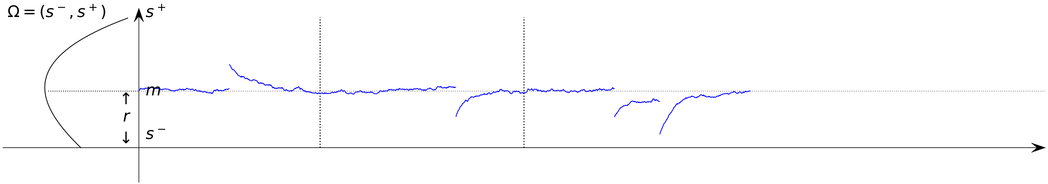
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



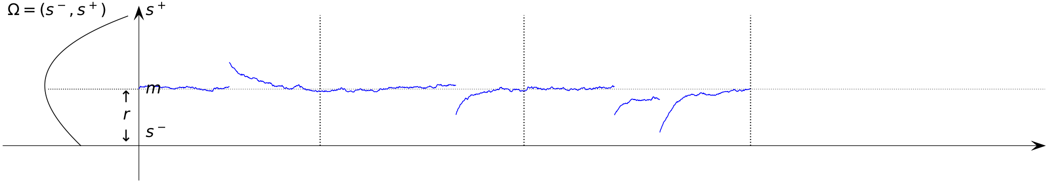
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



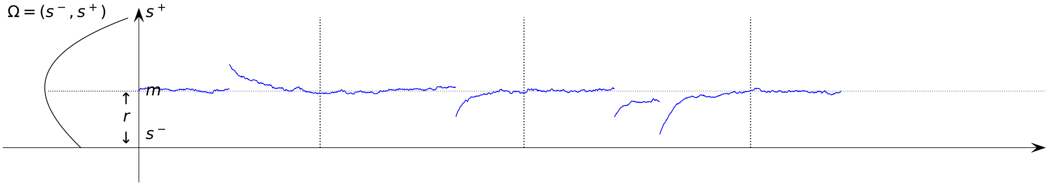
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



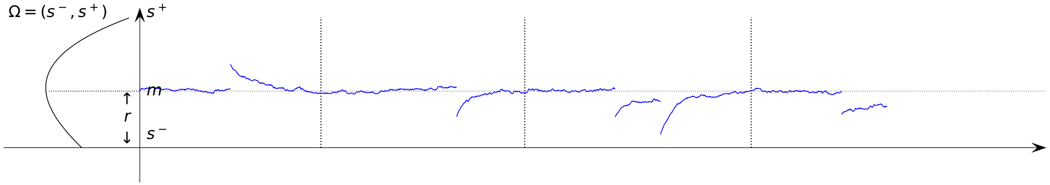
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



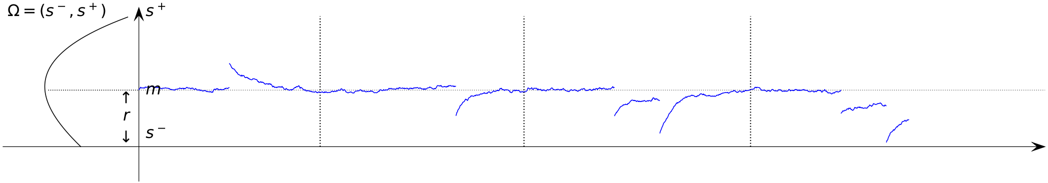
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



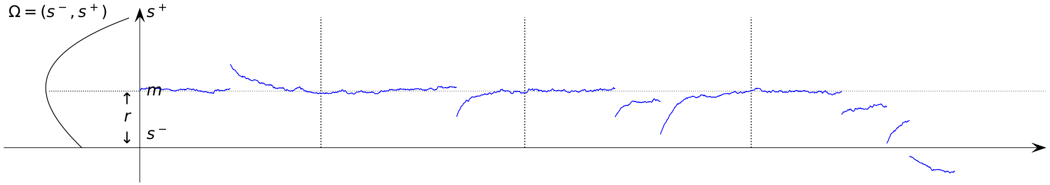
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



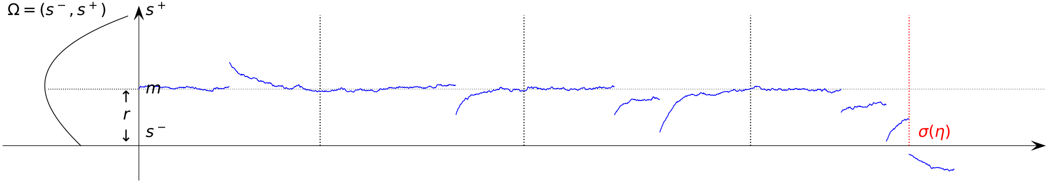
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



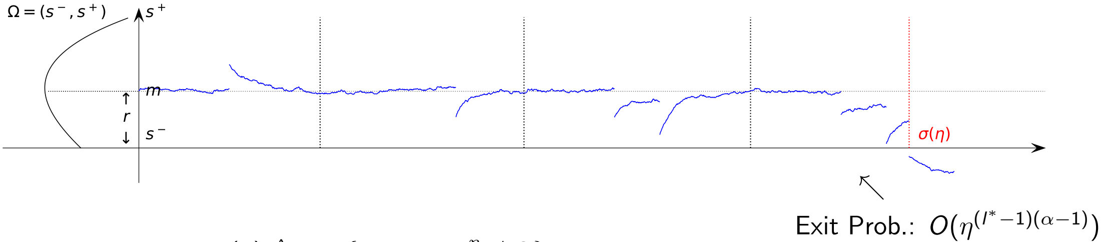
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



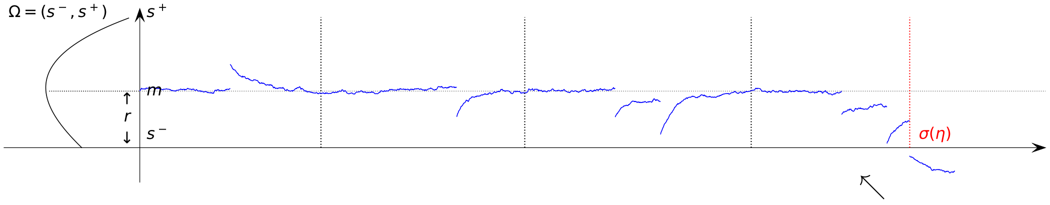
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

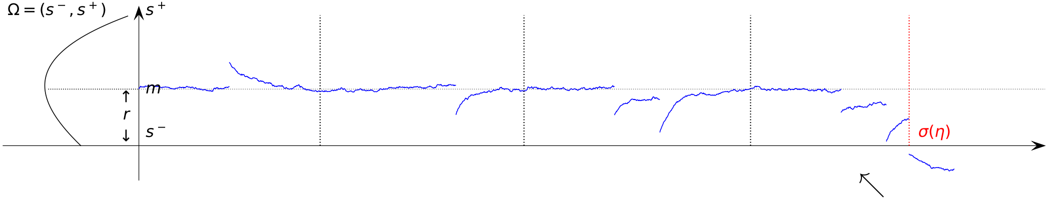
First Exit Time Analysis



Exit Prob.: $O(\eta^{(l^*-1)(\alpha-1)})$
Duration: $O(1/\eta^{\alpha})$

- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

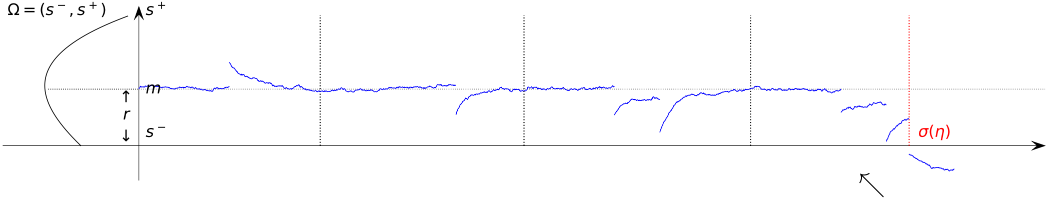
First Exit Time Analysis



Exit Prob.: $O(\eta^{(l^*-1)(\alpha-1)})$
Duration: $O(1/\eta^{\alpha})$
$\Rightarrow \sigma(\eta) \sim O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$

- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^{\eta} \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

First Exit Time Analysis



Exit Prob.: $O(\eta^{(l^*-1)(\alpha-1)})$
Duration: $O(1/\eta^{\alpha})$
$\Rightarrow \sigma(\eta) \sim O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$

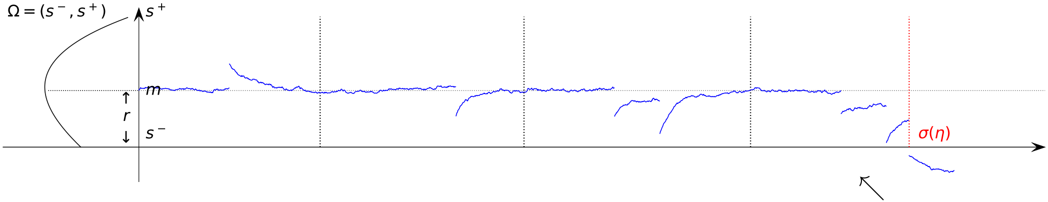
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^{\eta} \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

Theorem (Wang, Oh, Rhee, 2021)

For (Lebesgue) almost every $b > 0$, there exist some $q > 0$ and $\lambda(\eta) \in RV_{\alpha+(l^*-1)(\alpha-1)}(\eta)$ such that

$$\sigma(\eta)\lambda(\eta) \Rightarrow Exp(q) \text{ as } \eta \downarrow 0.$$

First Exit Time Analysis



Exit Prob.: $O(\eta^{(l^*-1)(\alpha-1)})$

Duration: $O(1/\eta^{\alpha})$

$\Rightarrow \sigma(\eta) \sim O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$

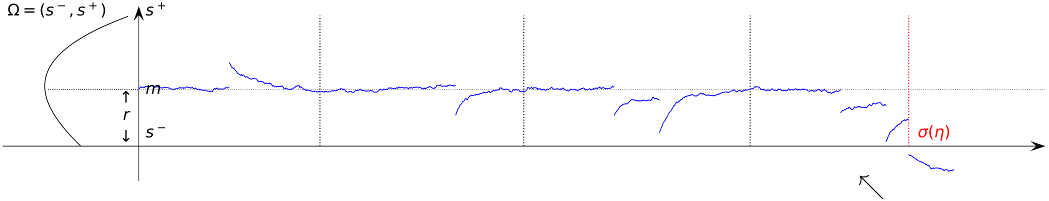
- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

Theorem (Wang, Oh, Rhee, 2021)

For (Lebesgue) almost every $b > 0$, there exist some $q > 0$ and $\lambda(\eta) \approx O(\eta^{\alpha+(l^*-1)(\alpha-1)})$ such that

$$\sigma(\eta)\lambda(\eta) \Rightarrow Exp(q) \text{ as } \eta \downarrow 0.$$

First Exit Time Analysis



- **First Exit Time:** $\sigma(\eta) \triangleq \min\{j \geq 0 : X_j^\eta \notin \Omega\}$
- $l^* \triangleq \lceil r/b \rceil$

Exit Prob.: $O(\eta^{(l^*-1)(\alpha-1)})$
Duration: $O(1/\eta^\alpha)$
$\Rightarrow \sigma(\eta) \sim O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$

Theorem (Wang, Oh, Rhee, 2021)

For (Lebesgue) almost every $b > 0$, there exist some $q > 0$ and $\lambda(\eta) \approx O(\eta^{\alpha+(l^*-1)(\alpha-1)})$ such that

$$\sigma(\eta)\lambda(\eta) \Rightarrow Exp(q) \text{ as } \eta \downarrow 0.$$

$$\sigma(\eta) \sim O(1/\lambda(\eta)) \approx O(1/\eta^{\alpha+(l^*-1)(\alpha-1)})$$