# Heavy-Tailed Large Deviations and Sharp Characterization of Global Dynamics of SGDs in Deep Learning

Chang-Han Rhee

Northwestern University

Machine Learning and Data Science Seminar, Oxford

April 8, 2024

Based on the joint works with

Mihail Bazhba, Jose Blanchet, Bohan Chen, Sewoong Oh, Zhe Su, Xingyu Wang, and Bert Zwart

# Team

Mihail Bazhba
U. of Amsterdam

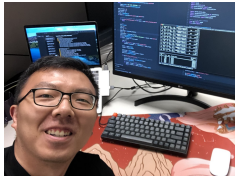
Jose Blanchet
Stanford

Bohan Chen
Munich Re

Sewoong Oh
U. of Washington

Zhe Su
Northwestern

Bert Zwart
CWI



Xingyu Wang
Northwestern

# Generalization Mystery of Deep Learning

# Empirical Success of Deep Neural Networks (DNNs)

"Deep Learning is eating the world."

- Jorge Nocedal

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.

Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.

Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation:

Training Set

Test Set [image source]

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation: choice of numerical optimization algorithm matters

Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation: choice of numerical optimization algorithm matters
- Stochastic Gradient Descent (SGD) turns out to work well.

Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation: choice of numerical optimization algorithm matters

  Better than deterministic gradient descent.
- Stochastic Gradient Descent (SGD) turns out to work well.

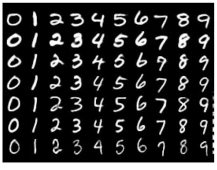
Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation: choice of numerical optimization algorithm matters
- Stochastic Gradient Descent (SGD) turns out to work well.

Better than deterministic gradient descent. In fact, the more noise, the better!



Training Set

Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Algorithmic Regularazation: choice of numerical optimization algorithm matters
- Stochastic Gradient Descent (SGD) turns out to work well.

Better than deterministic gradient descent. In fact, the more noise, the better!

**A Central Mystery of Deep Learning**

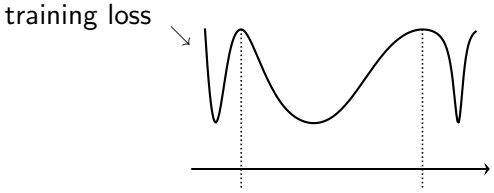
# Heavy Tails may have something to do with the Mystery

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

# Heavy Tails may have something to do with the Mystery

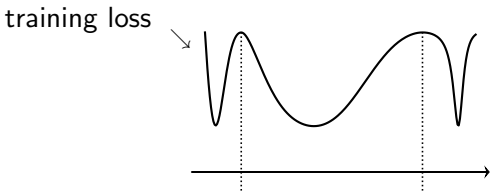
- Popular explanation: SGD somehow finds flat local minima.



# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

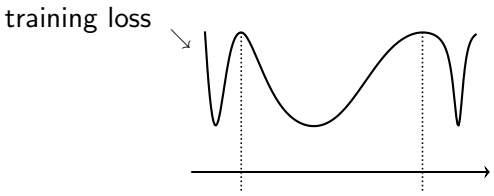


# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how?

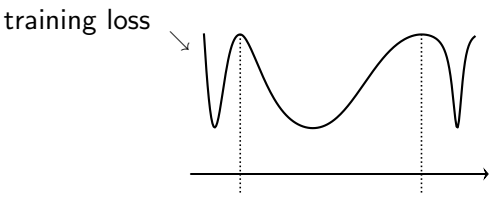


# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how?



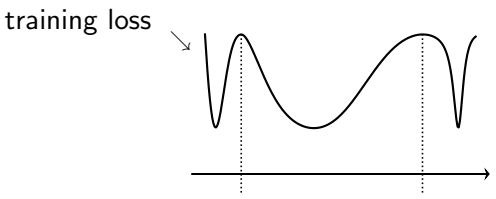

It takes more than $10^{600}$ time steps for SGD to escape from any of these.

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how? Previous attempts to explain had been unsatisfactory.




It takes more than $10^{600}$ time steps for SGD to escape from any of these.

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how? Previous attempts to explain had been unsatisfactory.

- Evidence/arguments for heavy-tails in training DNNs with SGDs!

  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how? Previous attempts to explain had been unsatisfactory.

- Evidence/arguments for heavy-tails in training DNNs with SGDs!

  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

**Heavy-tailed SGD escapes local minima and prefers flat local minima.**

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

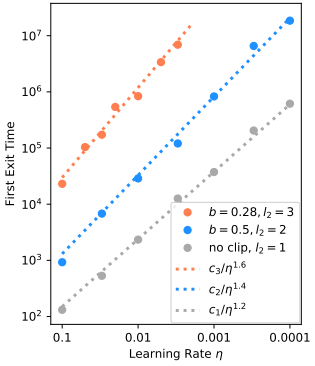
- But how? Previous attempts to explain had been unsatisfactory.

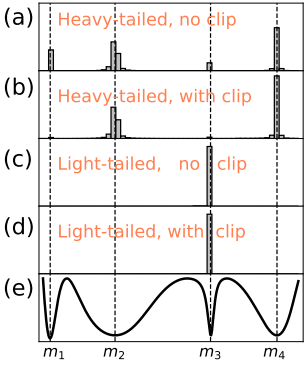
- Evidence/arguments for heavy-tails in training DNNs with SGDs!

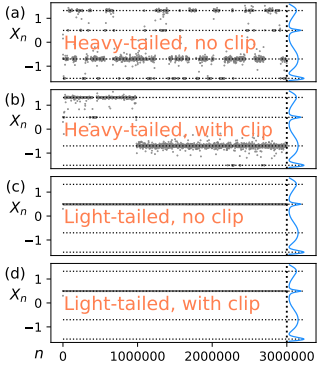
  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

**However, when SGD is heavy-tailed, one often truncates gradients.**

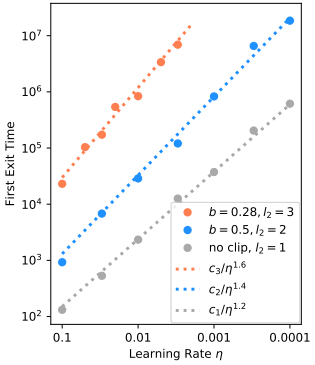
# Entirely Different Global Dynamics Depending on Tail Behaviors

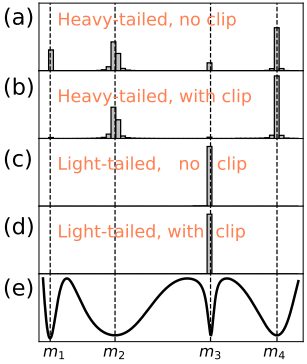


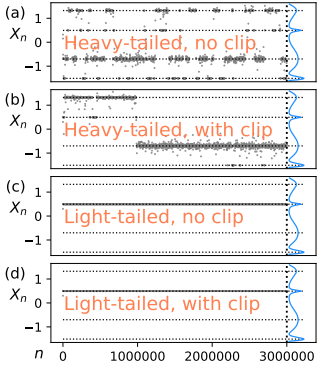



# Entirely Different Global Dynamics Depending on Tail Behaviors

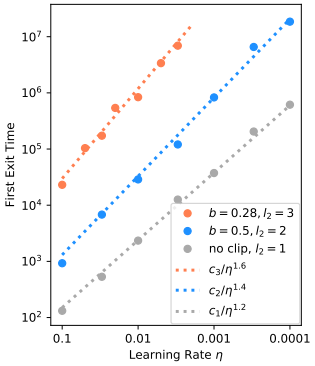


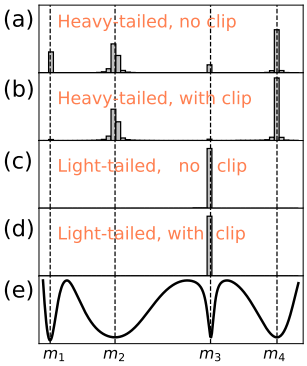



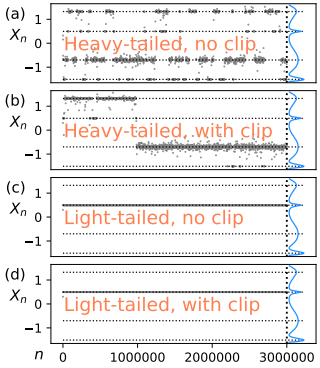




**Explanation?**

# Entirely Different Global Dynamics Depending on Tail Behaviors




**Explanation? Catastrophe Principle.**

# Heavy Tails and Catastrophe Principle

# 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

# 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

**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

**Heavy-Tailed Distributions**

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

**Systemwide rare events**

**arise because**

**EVERYTHING goes wrong.**

**(Conspiracy Principle)**

**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.**

# Insurance Example: Capital Reserve

$$Y(t) = c + pt - \sum_{i=1}^{N(t)} X_i$$

# Insurance Example: Capital Reserve

Initial Capital

$$Y(t) = c + pt - \sum_{i=1}^{N(t)} X_i$$

# Insurance Example: Capital Reserve

Initial Capital

$$Y(t) = c + pt - \sum_{i=1}^{N(t)} X_i$$

Premium

# Insurance Example: Capital Reserve

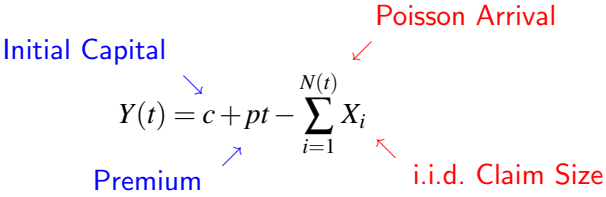
Initial Capital

$$Y(t) = c + pt - \sum_{i=1}^{N(t)} X_i$$

Premium

i.i.d. Claim Size

# Insurance Example: Capital Reserve



# Insurance Example: Capital Reserve

$$Y(t) = c + pt - \sum_{i=1}^{N(t)} X_i$$

# Insurance Example: Capital Reserve

$$\bar{Y}_n(t) = c + pt - \sum_{i=1}^{N(nt)} X_i/n$$

# Insurance Example: Capital Reserve

$$\bar{Y}_n(t) = c + pt - \sum_{i=1}^{N(nt)} X_i/n$$

Large $n$: analysis of large loss over a long time period

# Typical Scenario

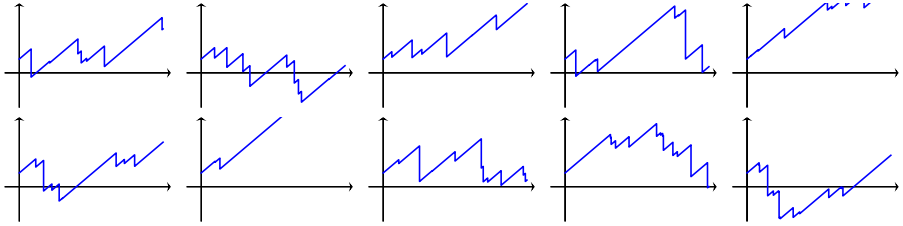
# Typical Scenario

Sample paths of $\bar{Y}_n$:

n=10 & claim sizes are **light-tailed**

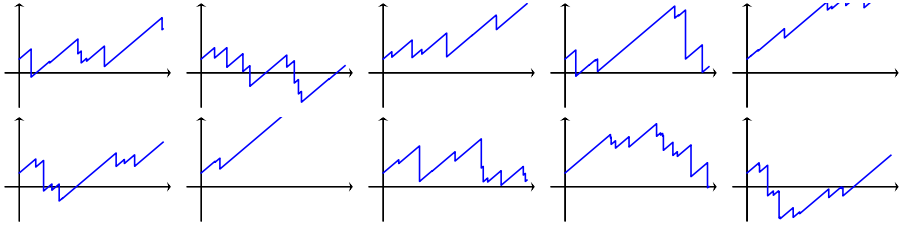
# Typical Scenario

Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**



# Typical Scenario

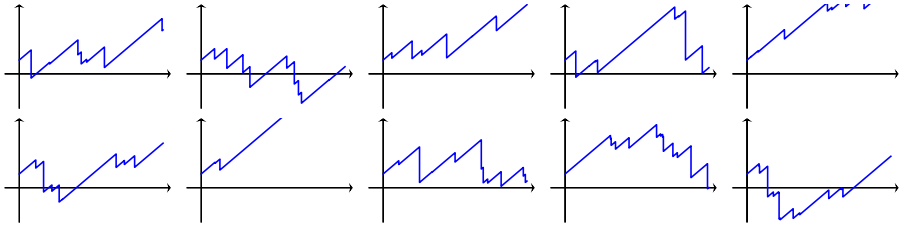
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**



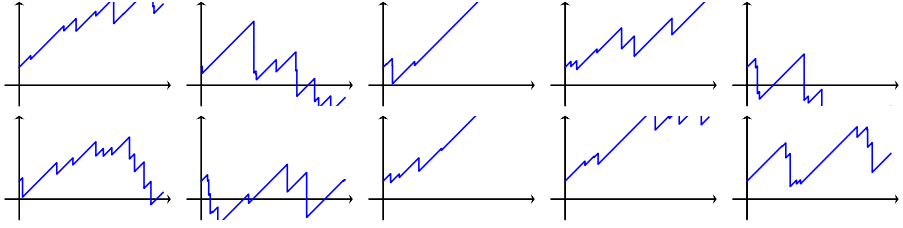
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

# Typical Scenario

Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**

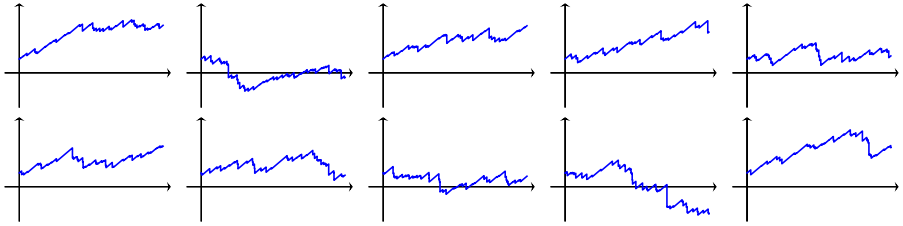


Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

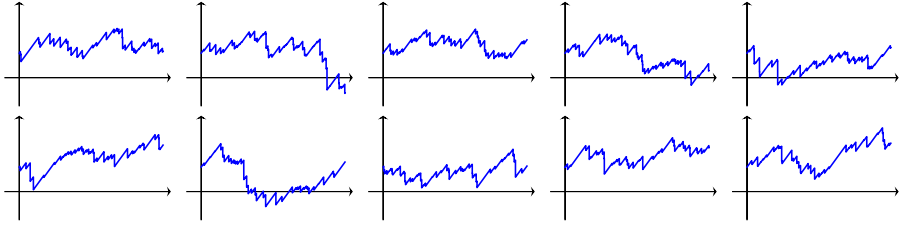


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=50 & claim sizes are **light-tailed**

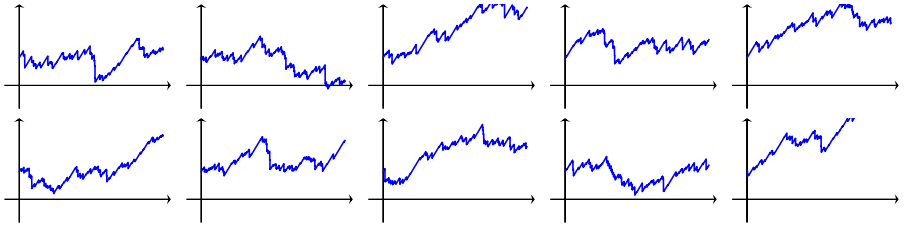


Sample paths of $\bar{Y}_n$: n=50 & claim sizes are **heavy-tailed**

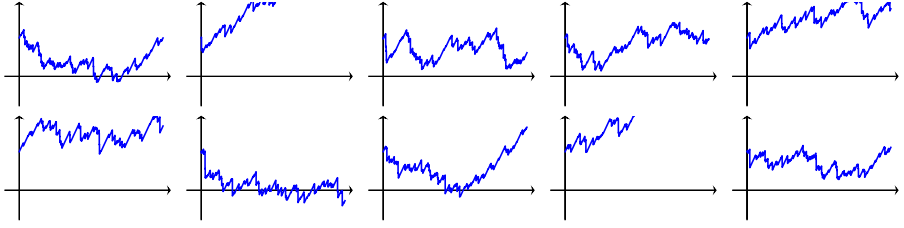


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=100 & claim sizes are **light-tailed**

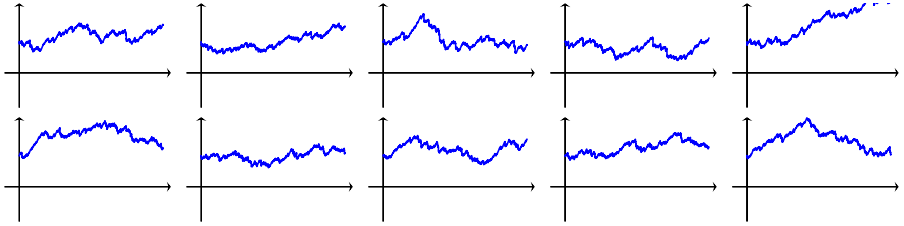


Sample paths of $\bar{Y}_n$: n=100 & claim sizes are **heavy-tailed**

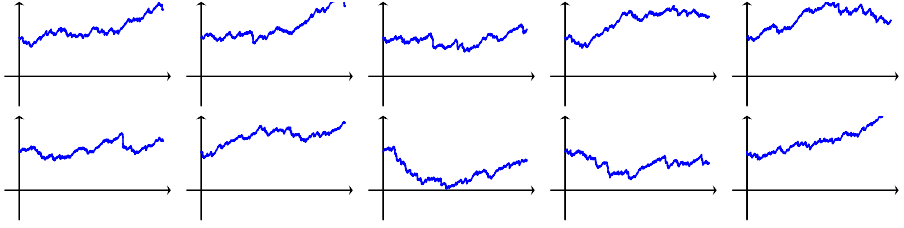


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=500 & claim sizes are **light-tailed**

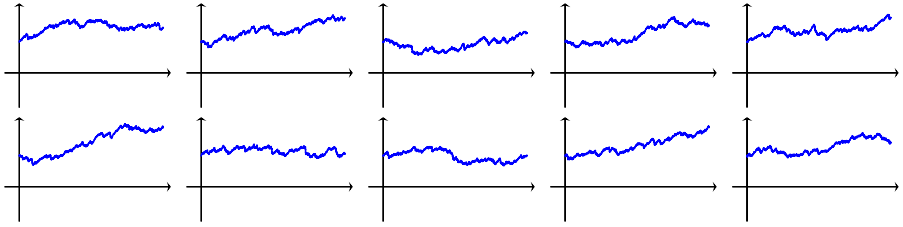


Sample paths of $\bar{Y}_n$: n=500 & claim sizes are **heavy-tailed**

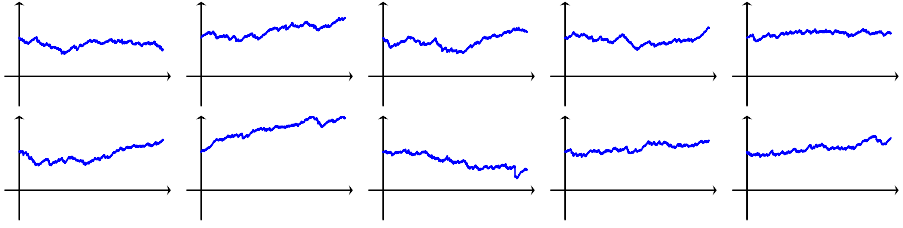


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=1000 & claim sizes are **light-tailed**

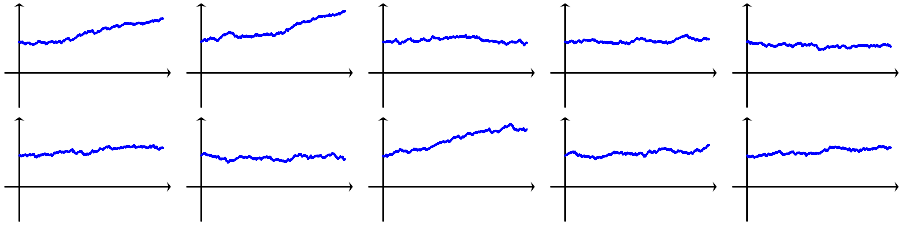


Sample paths of $\bar{Y}_n$: n=1000 & claim sizes are **heavy-tailed**

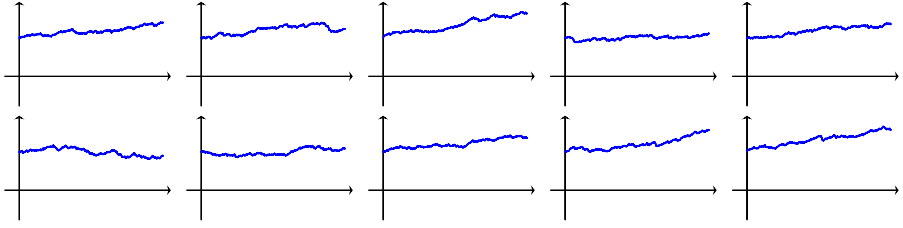


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=2500 & claim sizes are **light-tailed**

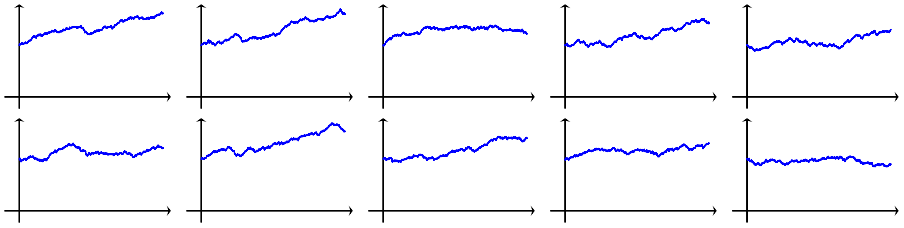


Sample paths of $\bar{Y}_n$: n=2500 & claim sizes are **heavy-tailed**

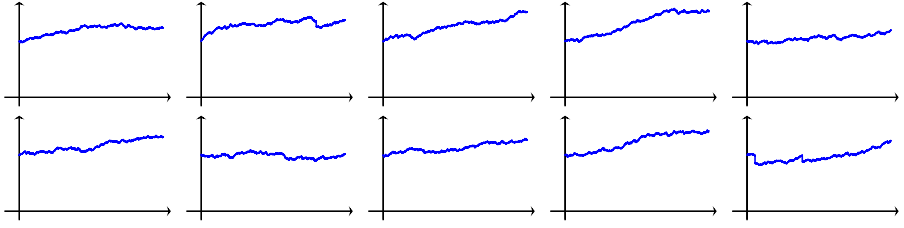


# Typical Scenario

Sample paths of $\bar{Y}_n$: n=5000 & claim sizes are **light-tailed**



Sample paths of $\bar{Y}_n$: n=5000 & claim sizes are **heavy-tailed**

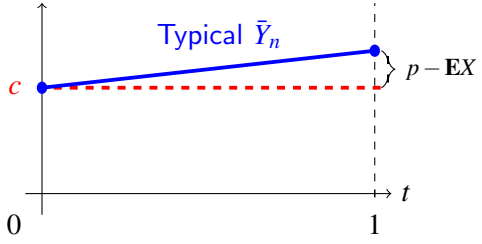


# Typical Scenario

That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.

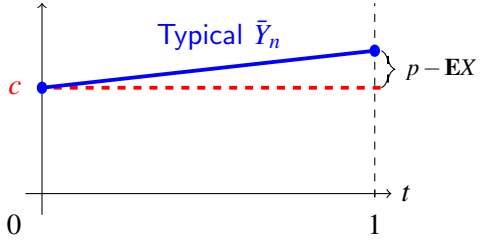
# Typical Scenario

That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.



# Typical Scenario

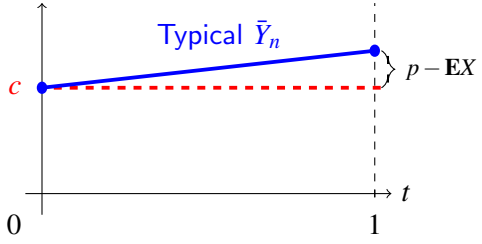
That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.




**Typically, your business will flourish**

# Typical Scenario

That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.

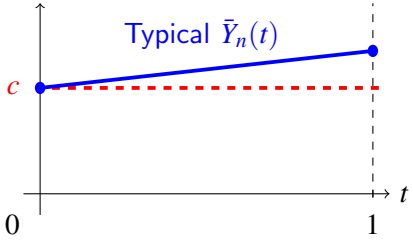



**Typically, your business will flourish**

**regardless of the tail distributions.**

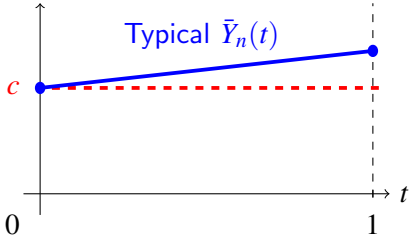
## What about atypical cases?

# A Rare Event: Bankruptcy



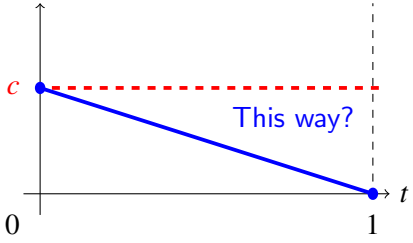
# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



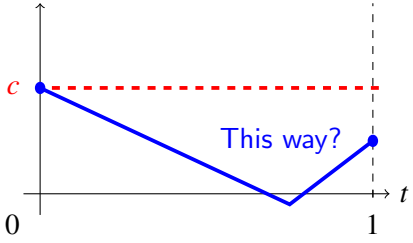
# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



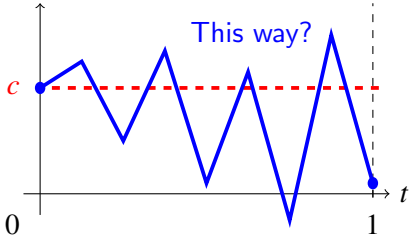
# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



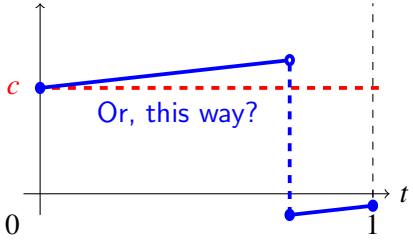
# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



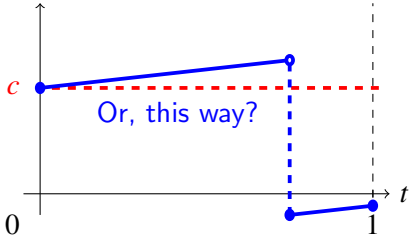
# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



# A Rare Event: Bankruptcy

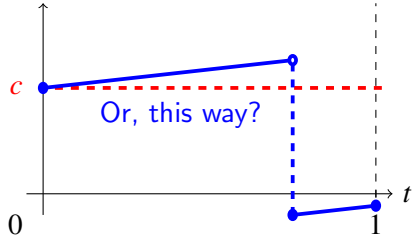
Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)




**Are we going to see clear patterns?**

# A Rare Event: Bankruptcy

Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)

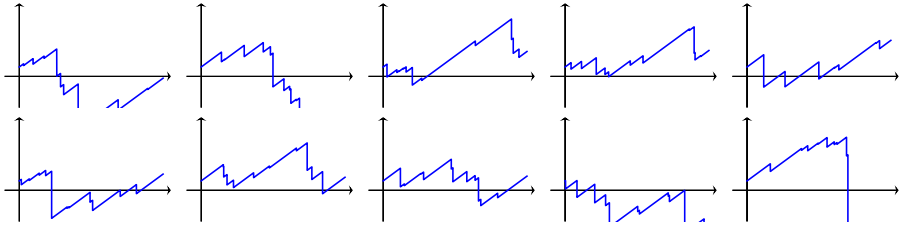



**Are we going to see clear patterns?**

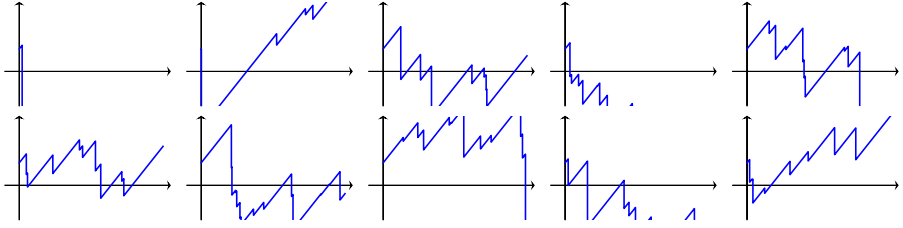
**Do they depend on the tail distributions?**

# A Rare Event: Bankruptcy

Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **light-tailed** claims:



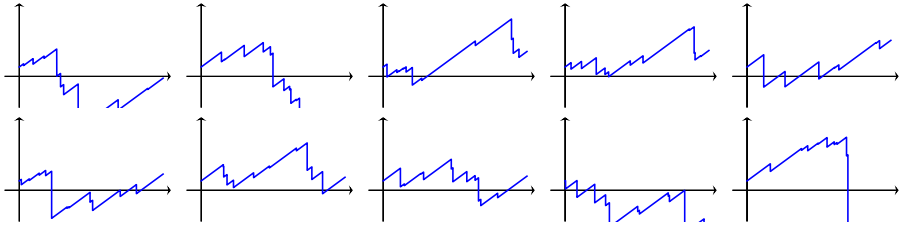
Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **heavy-tailed** claims:



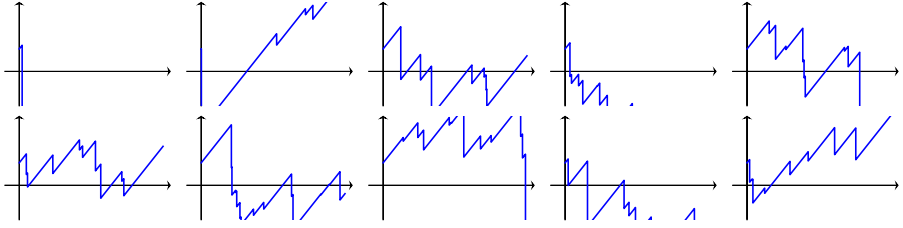
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **light-tailed** claims:



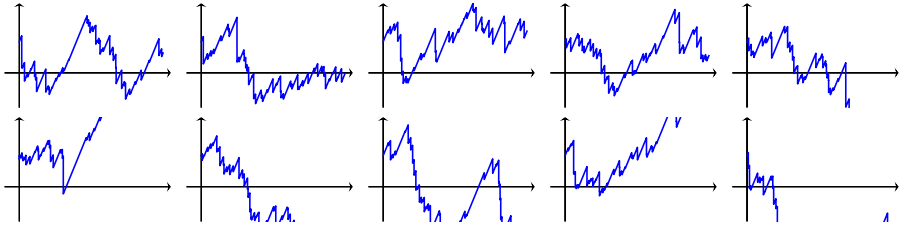
Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **heavy-tailed** claims:



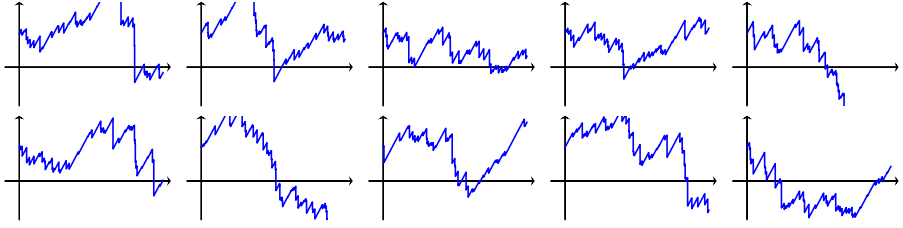
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{50}$ conditional on $B$ for **light-tailed** claims:



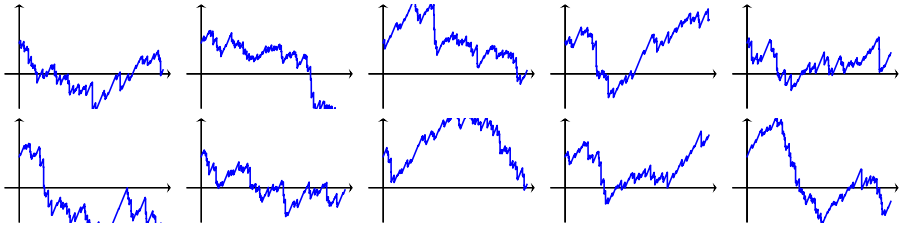
Sample paths of $\bar{Y}_{50}$ conditional on $B$ for **heavy-tailed** claims:



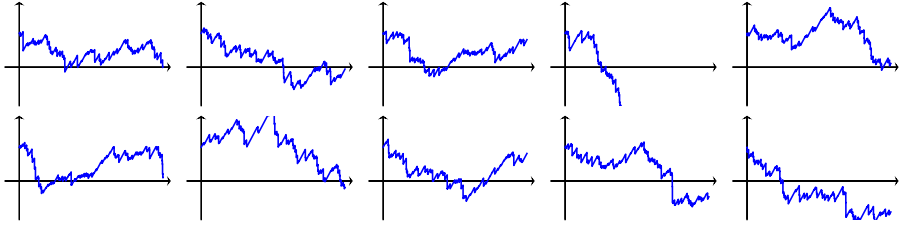
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{100}$ conditional on $B$ for **light-tailed** claims:



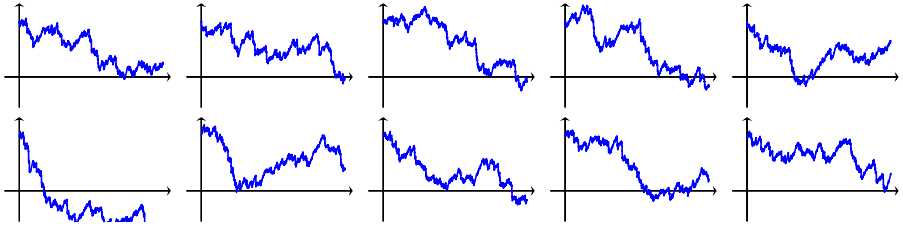
Sample paths of $\bar{Y}_{100}$ conditional on $B$ for **heavy-tailed** claims:



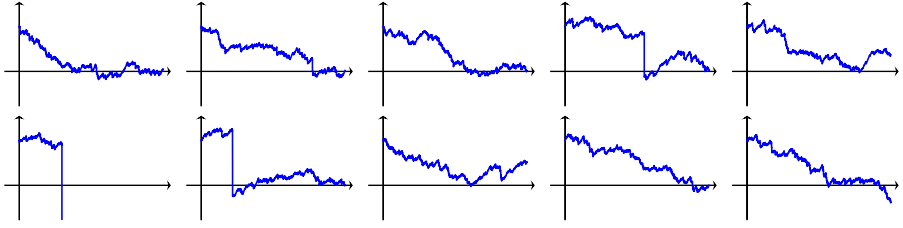
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{500}$ conditional on $B$ for **light-tailed** claims:



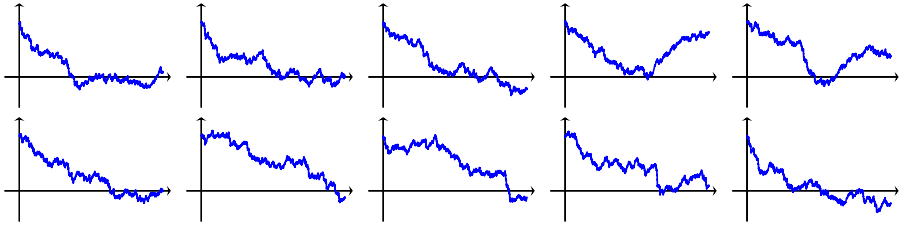
Sample paths of $\bar{Y}_{500}$ conditional on $B$ for **heavy-tailed** claims:



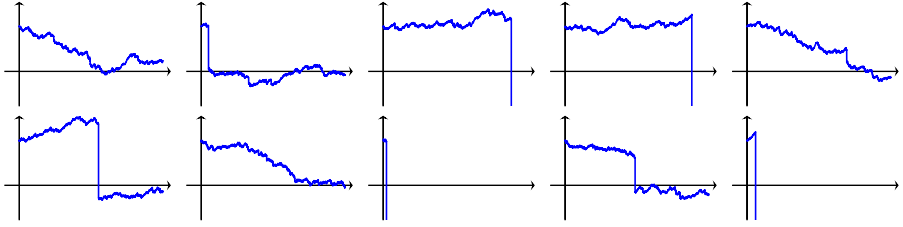
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{1000}$ conditional on $B$ for **light-tailed** claims:



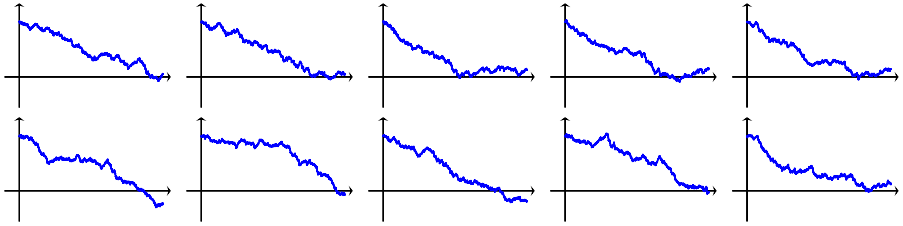
Sample paths of $\bar{Y}_{1000}$ conditional on $B$ for **heavy-tailed** claims:



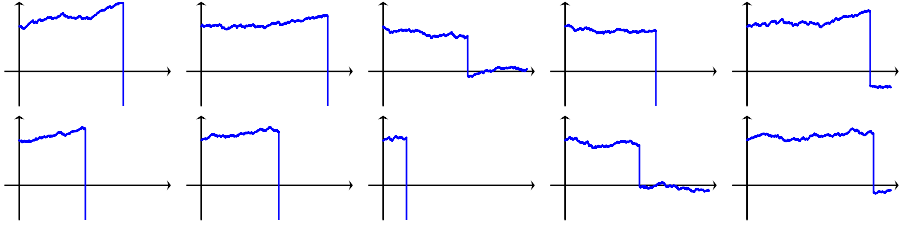
# A Rare Event: Bankruptcy

Bankruptcy

Sample paths of $\bar{Y}_{2500}$ conditional on $B$ for **light-tailed** claims:



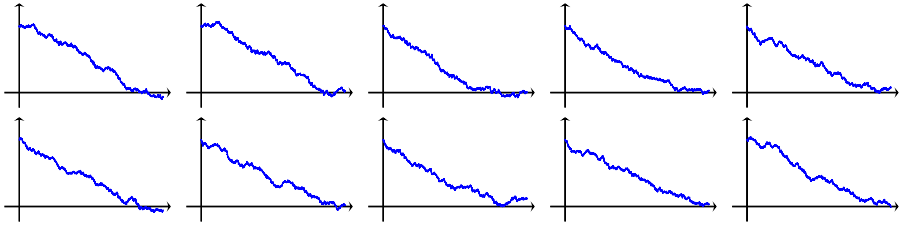
Sample paths of $\bar{Y}_{2500}$ conditional on $B$ for **heavy-tailed** claims:



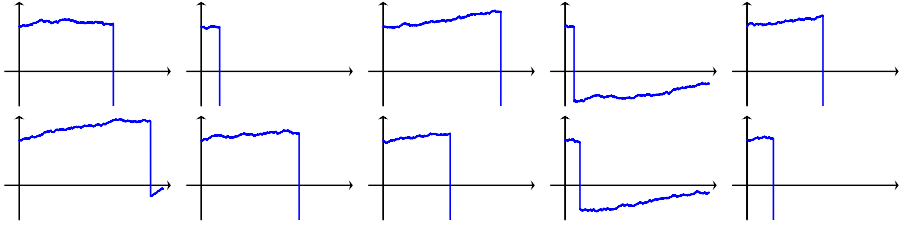
# A Rare Event: Bankruptcy

Bankruptcy

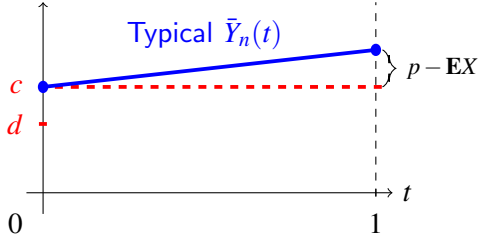
Sample paths of $\bar{Y}_{5000}$ conditional on $B$ for **light-tailed** claims:



Sample paths of $\bar{Y}_{5000}$ conditional on $B$ for **heavy-tailed** claims:

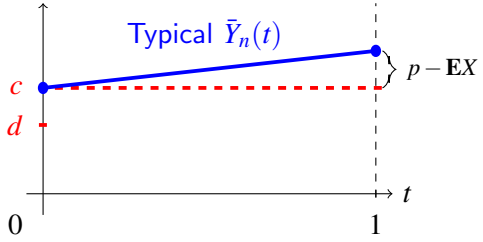


# Bankruptcy Despite of Reinsurance



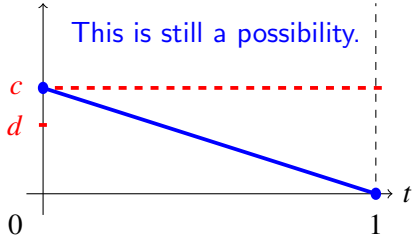
# Bankruptcy Despite of Reinsurance

Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



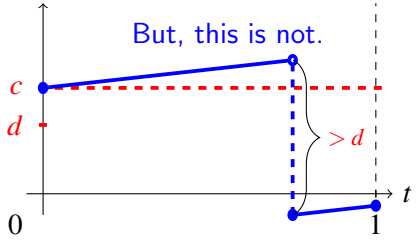
# Bankruptcy Despite of Reinsurance

Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



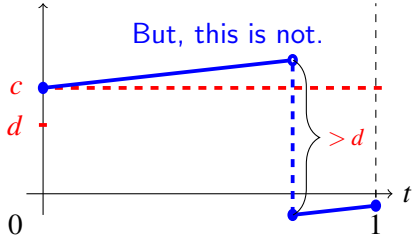
# Bankruptcy Despite of Reinsurance

Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



# Bankruptcy Despite of Reinsurance

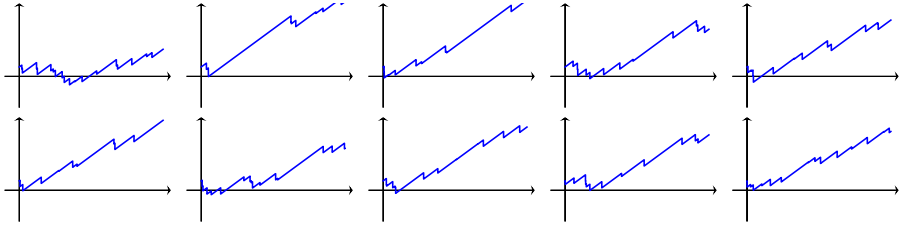
Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



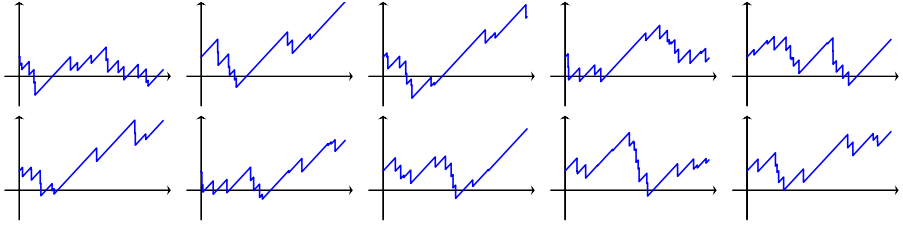

**How does the pattern change in this case?**

# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{10}$ conditional on $R$ for **light-tailed** claims:

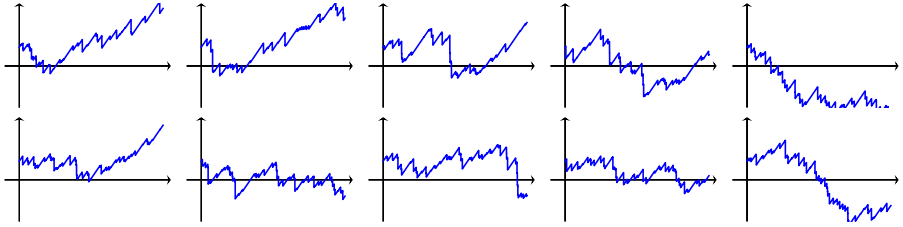


Sample paths of $\bar{Y}_{10}$ conditional on $R$ for **heavy-tailed** claims:

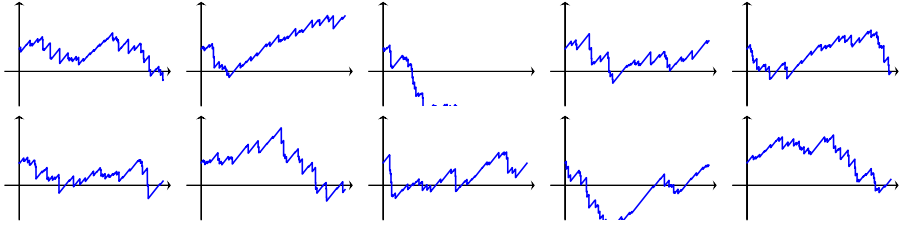


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{50}$ conditional on $R$ for **light-tailed** claims:

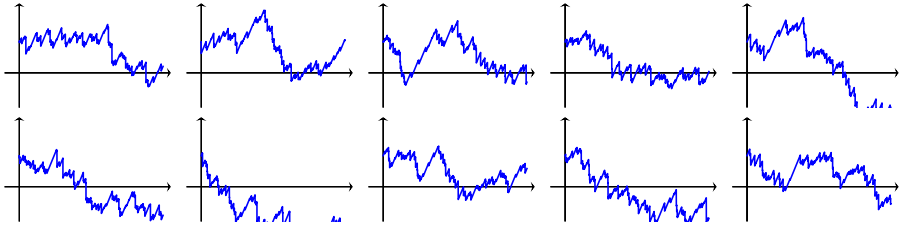


Sample paths of $\bar{Y}_{50}$ conditional on $R$ for **heavy-tailed** claims:

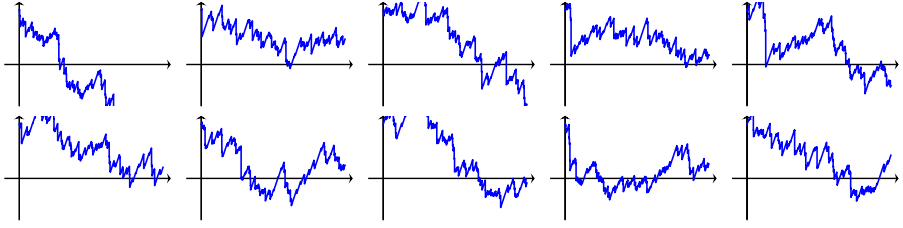


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{100}$ conditional on $R$ for **light-tailed** claims:

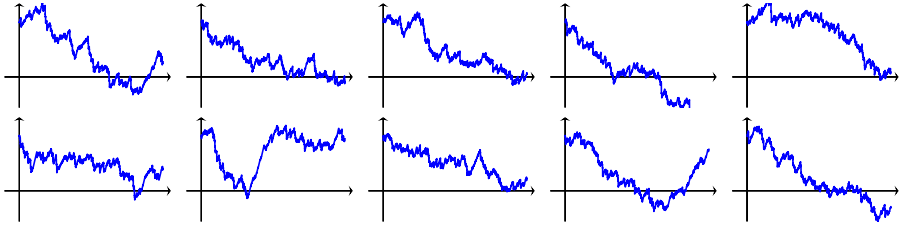


Sample paths of $\bar{Y}_{100}$ conditional on $R$ for **heavy-tailed** claims:

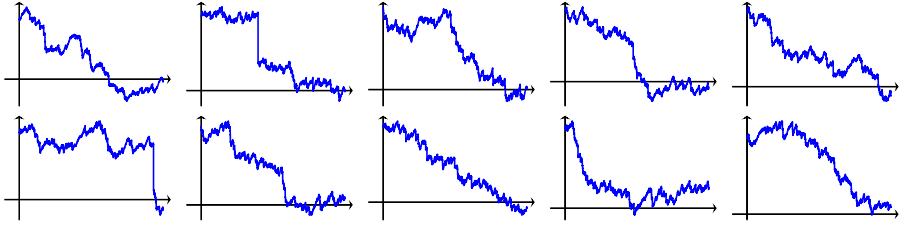


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{500}$ conditional on $R$ for **light-tailed** claims:

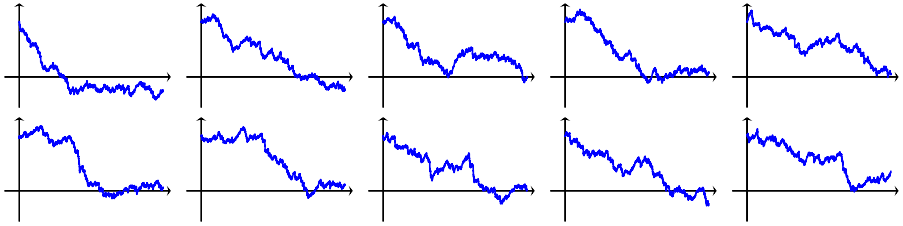


Sample paths of $\bar{Y}_{500}$ conditional on $R$ for **heavy-tailed** claims:

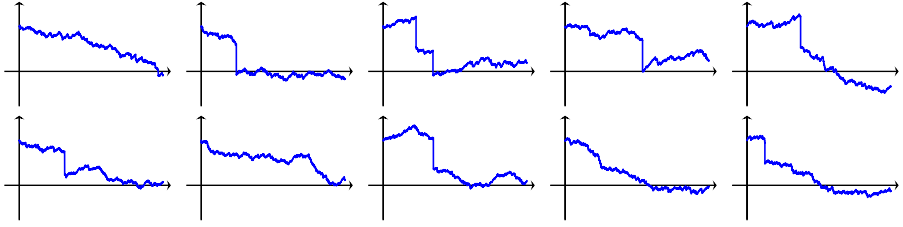


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{1000}$ conditional on $R$ for **light-tailed** claims:

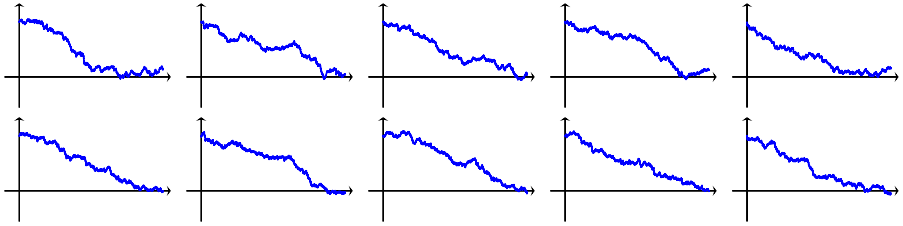


Sample paths of $\bar{Y}_{1000}$ conditional on $R$ for **heavy-tailed** claims:

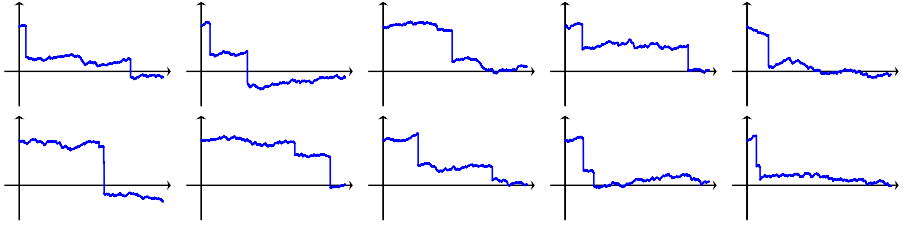


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{2500}$ conditional on $R$ for **light-tailed** claims:

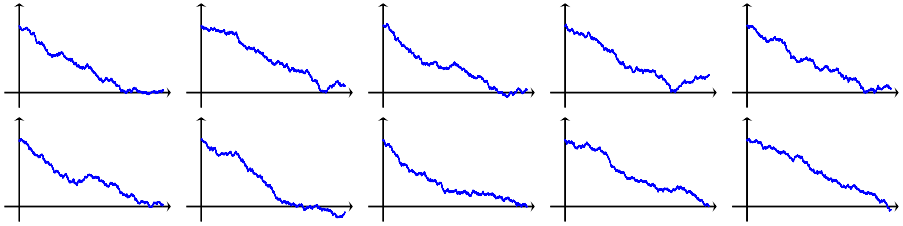


Sample paths of $\bar{Y}_{2500}$ conditional on $R$ for **heavy-tailed** claims:

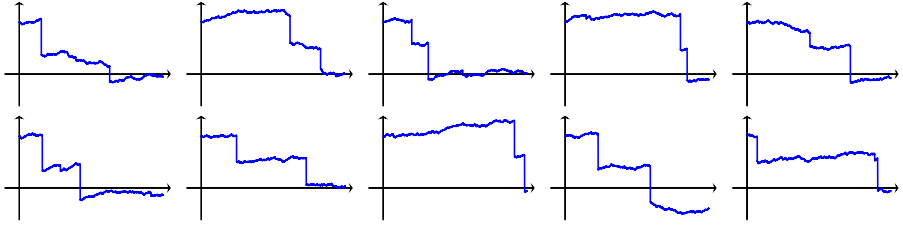


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **light-tailed** claims:

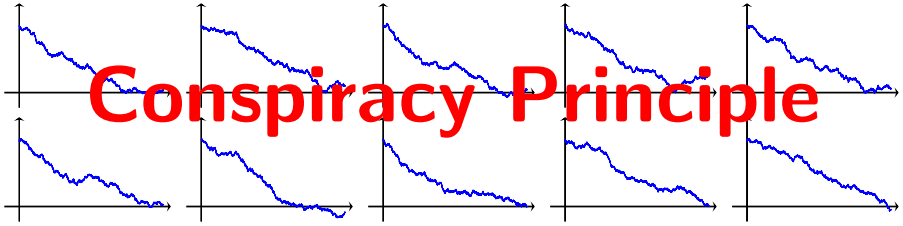


Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **heavy-tailed** claims:

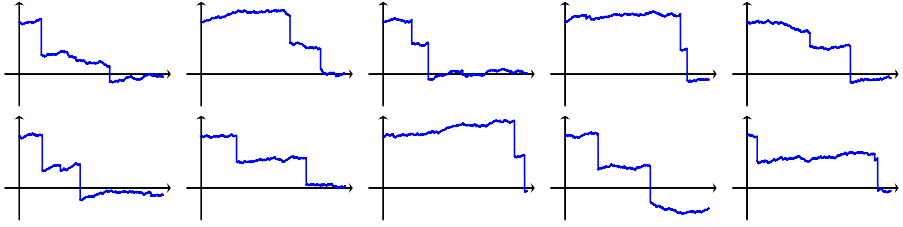


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **light-tailed** claims:

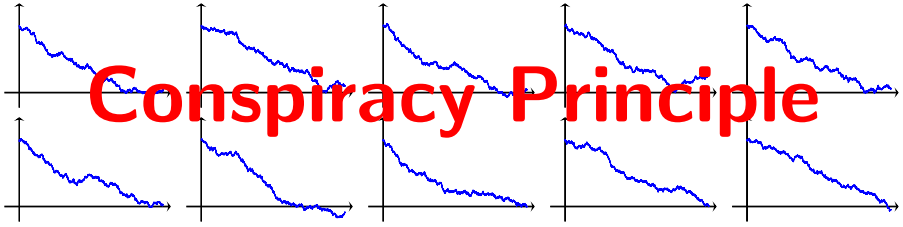



Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **heavy-tailed** claims:

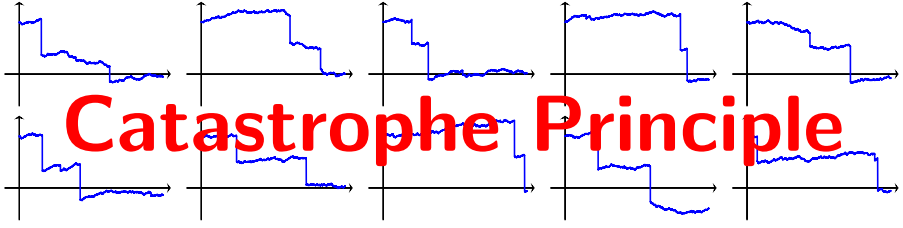


# Bankruptcy Despite of Reinsurance

Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **light-tailed** claims:




Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **heavy-tailed** claims:



## Sample Path Large Deviations:

## Characterization of Conspiracy and Catastrophe Principles

# Light-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \Lambda(\lambda) = \log \mathbf{E}\exp(\lambda X_1) < \infty, \quad \forall \lambda > 0$$

# Light-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \Lambda(\lambda) = \log \mathbf{E}\exp(\lambda X_1) < \infty, \quad \forall \lambda > 0$$

**Theorem (Mogulskii, 1993)**

*For any measurable* $A \subseteq \mathbb{D}$

$$-\inf_{f\in A^\circ} I(f) \le \liminf_{n\to\infty} \frac{1}{n}\log \mathbf{P}(\bar{S}_n \in A) \le \limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}(\bar{S}_n \in A) \le -\inf_{f\in A^-} I(f)$$

- $I(f) = \begin{cases} \int_0^1 \Lambda^*\big(\dot{f}(t)\big)dt & \text{if } f \in \mathcal{AC},\ f(0)=0 \\ 0 & \text{otherwise} \end{cases}$

# Light-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \Lambda(\lambda) = \log \mathbf{E}\exp(\lambda X_1) < \infty, \quad \forall \lambda > 0$$

**Theorem (Mogulskii, 1993)**

*For any measurable $A \subseteq \mathbb{D}$*

$$\mathbf{P}(\bar{S}_n \in A) \overset{\log}{\sim} \exp\Big( - n \cdot \inf_{f \in A} I(f) \Big)$$

- $I(f) = \begin{cases} \displaystyle\int_0^1 \Lambda^*\big(\dot{f}(t)\big)dt & \text{if } f \in \mathcal{AC},\ f(0) = 0 \\ 0 & \text{otherwise} \end{cases}$

# Implication: The Conspiracy Principle

**Theorem (Ganesh et al. 2003; Lemma 4.2)**

*Suppose that $X_n$ satisfies an LDP and*

$$\inf_{f \in A} I(f) = k < \infty.$$

*Under certain regularity conditions,*

$$\mathbf{P}(\bar{S}_n \notin B | \bar{S}_n \in A) \to 0$$

*for any neighborhood $B$ of $\{x \in A : I(x) = k\}$.*

# Implication: The Conspiracy Principle

**Theorem (Ganesh et al. 2003; Lemma 4.2)**

*Suppose that $X_n$ satisfies an LDP and*

$$\inf_{f \in A} I(f) = k < \infty.$$

*Under certain regularity conditions,*

$$\mathbf{P}(\bar{S}_n \notin B | \bar{S}_n \in A) \to 0$$

*for any neighborhood $B$ of $\{x \in A : I(x) = k\}$.*

**Rigorous Characterization of Conspiracy Principle**

# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (R., Blanchet, Zwart, 2019)**

*For "general" $A \subseteq \mathbb{D}$*

$$C(A^\circ) \leq \liminf_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{n^{-\alpha \mathcal{J}(A)}} \leq \limsup_{n\to\infty} \frac{\mathbf{P}(\bar{S}_n \in A)}{n^{-\alpha \mathcal{J}(A)}} \leq C(A^-).$$

- $\mathcal{J}(A)$*: min #jumps for step functions to be inside* $A$
- $C(\cdot)$*: a measure*

# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (R., Blanchet, Zwart, 2019)**

*For "general" $A \subseteq \mathbb{D}$*

$$\mathbf{P}(\bar{S}_n \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

- $\mathcal{J}(A)$*: min #jumps for step functions to be inside* $A$

# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_i\text{: centered iid r.v. with } \mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (R., Blanchet, Zwart, 2019)**

*For "general" $A \subseteq \mathbb{D}$*

LD power index

$$\mathbf{P}(\bar{S}_n \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

- $\mathcal{J}(A)$*: min #jumps for step functions to be inside* $A$

# Implication: The Catastrophe Principle

**Theorem (R., Blanchet, Zwart. 2019)**

*Under certain regularity conditions on $A$,*

$$\mathbf{P}(\bar{S}_n \in \cdot | \bar{S}_n \in A) \to \mathbf{P}(\bar{S}_{|A} \in \cdot) = \frac{C(\cdot \cap A)}{C(A)}$$

$\bar{S}_{|A}$*: a (random) piecewise-constant function with* $\mathcal{J}(A)$ *jumps.*

# Implication: The Catastrophe Principle

**Theorem (R., Blanchet, Zwart. 2019)**

*Under certain regularity conditions on $A$,*

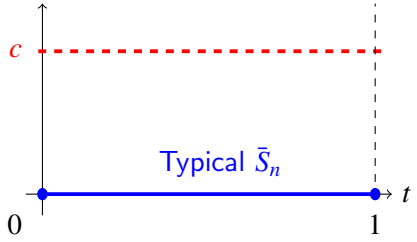
$$\mathbf{P}(\bar{S}_n \in \cdot | \bar{S}_n \in A) \to \mathbf{P}(\bar{S}_{|A} \in \cdot) = \frac{C(\cdot \cap A)}{C(A)}$$

$\bar{S}_{|A}$*: a (random) piecewise-constant function with* $\mathcal{J}(A)$ *jumps.*

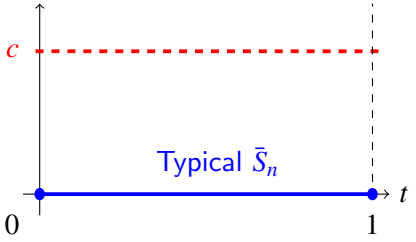
**Rigorous Characterization of Catastrophe Principle**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ crosses level } c \text{ on } [0,1] \,\} = A$$



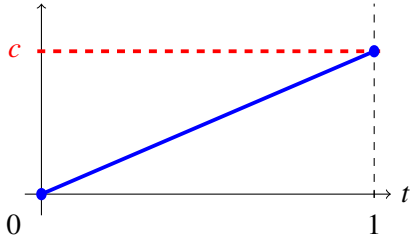

**Light-Tailed Claim Size**



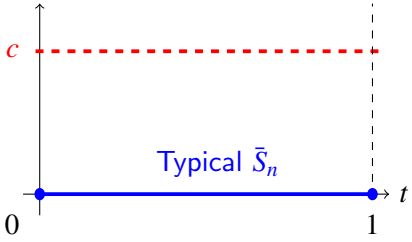

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ crosses level } c \text{ on } [0,1] \,\} = A$$



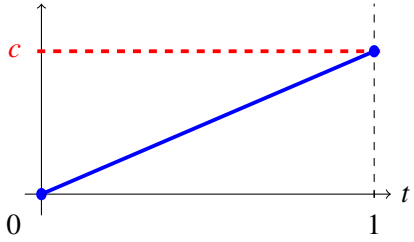

**Light-Tailed Claim Size**



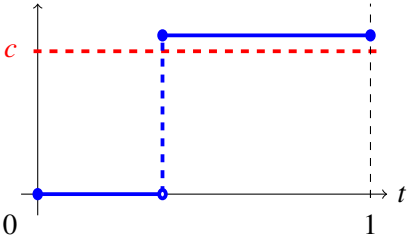

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ crosses level } c \text{ on } [0,1] \,\} = A$$



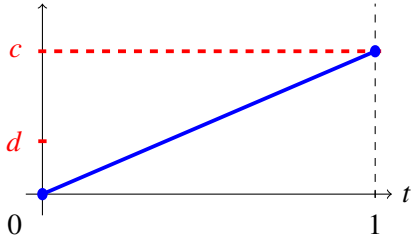

**Light-Tailed Claim Size**



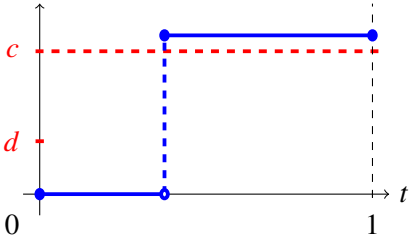

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} :\ f \text{ crosses level } c \text{ on } [0,1] \ \&\ \text{jump sizes} \le d \,\} = A$$



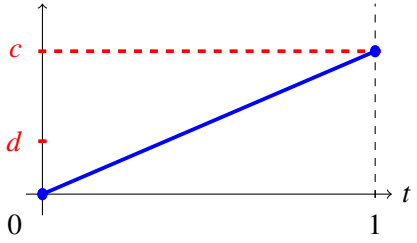

**Light-Tailed Claim Size**



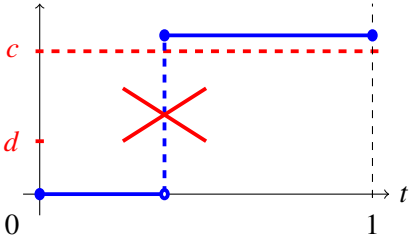

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} :\ f \text{ crosses level } c \text{ on } [0,1] \ \&\ \text{jump sizes} \le d \,\} = A$$



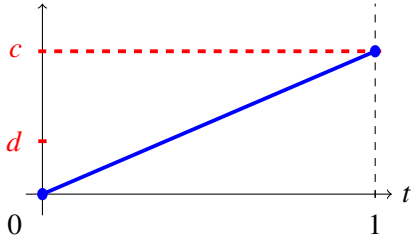

**Light-Tailed Claim Size**



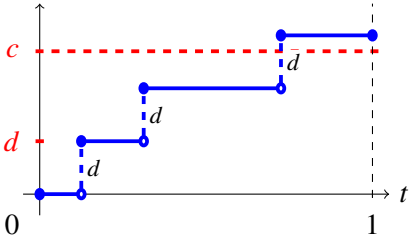

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} :\ f \text{ crosses level } c \text{ on } [0,1] \ \&\ \text{jump sizes} \le d \,\} = A$$



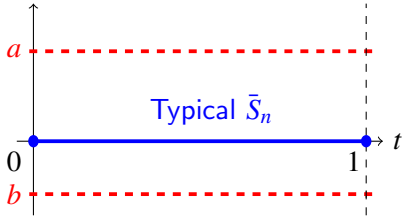

**Light-Tailed Claim Size**



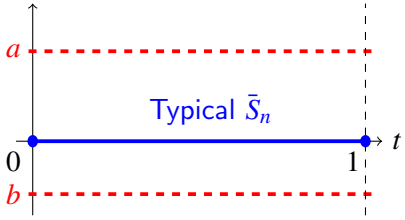

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ hits below } b \text{ on } [0,1] \text{ and ends up above } a \,\} = A$$



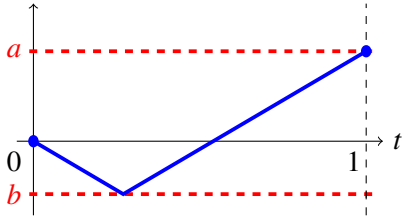

**Light-Tailed Increments**



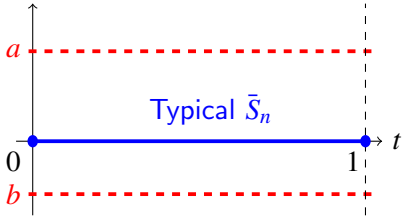

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ hits below } b \text{ on } [0,1] \text{ and ends up above } a \,\} = A$$



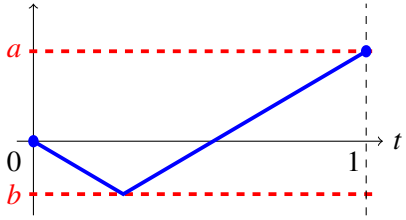

**Light-Tailed Increments**



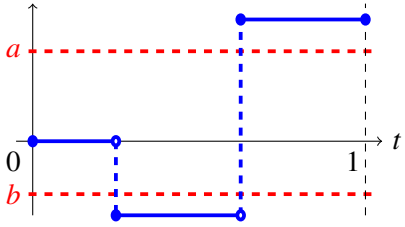

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ hits below } b \text{ on } [0,1] \text{ and ends up above } a \,\} = A$$



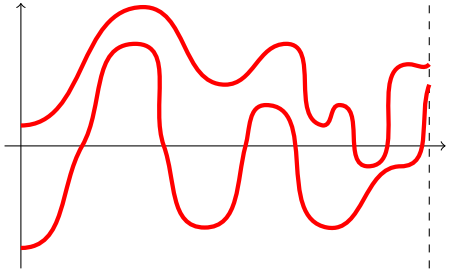

**Light-Tailed Increments**



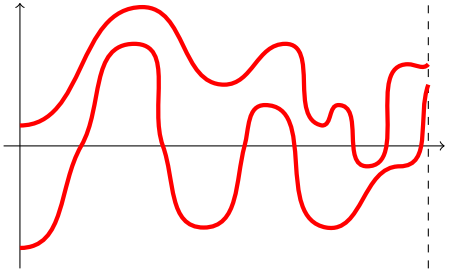

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$\bar{S}_n \in \{ f \in \mathbb{D} : f \text{ lies between the two red curves} \}$



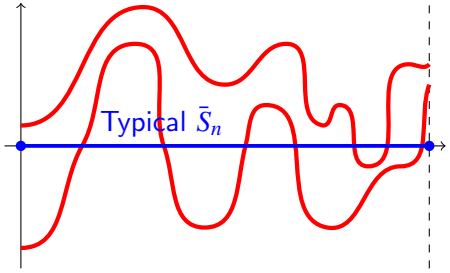
**Light-Tailed Increments**



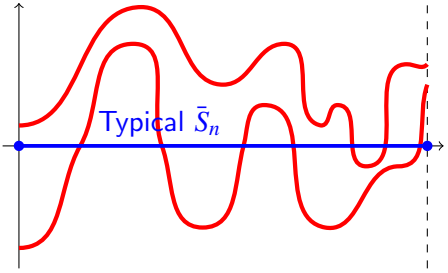
**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$\bar{S}_n \in \{ f \in \mathbb{D} : f \text{ lies between the two red curves} \}$



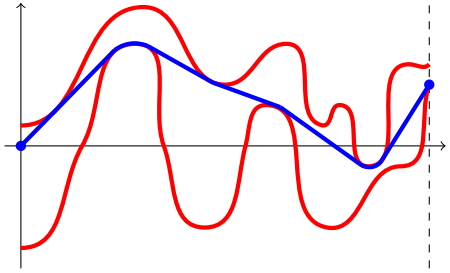

**Light-Tailed Increments**



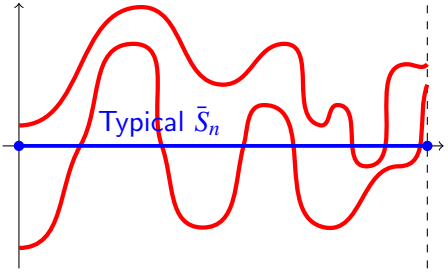

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$\bar{S}_n \in \{ f \in \mathbb{D} : f \text{ lies between the two red curves} \}$



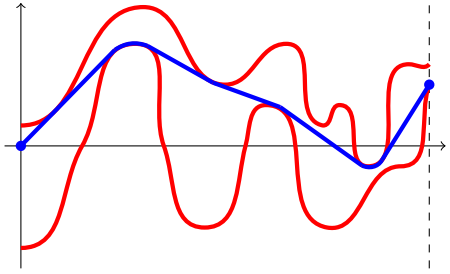
**Light-Tailed Increments**



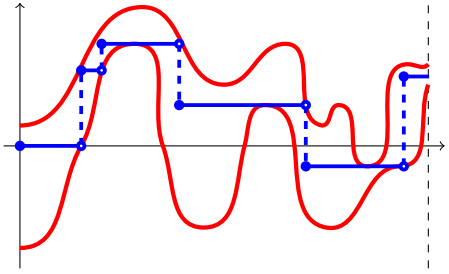

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

$\bar{S}_n \in \{ f \in \mathbb{D} : f \text{ lies between the two red curves} \}$

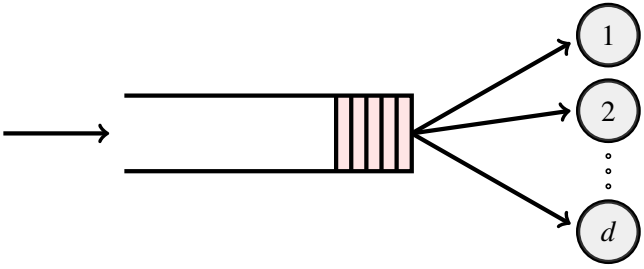


**Light-Tailed Increments**

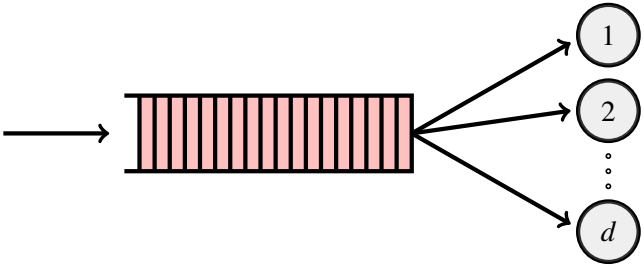


**Heavy-Tailed Increments**

Congestion of Multiple Server Queue:

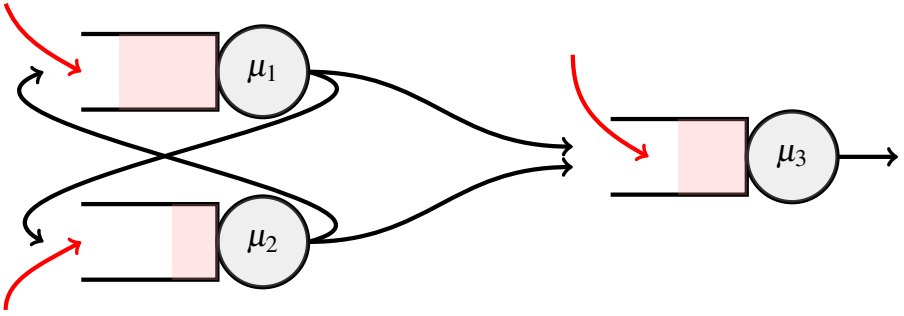


Congestion of Multiple Server Queue:



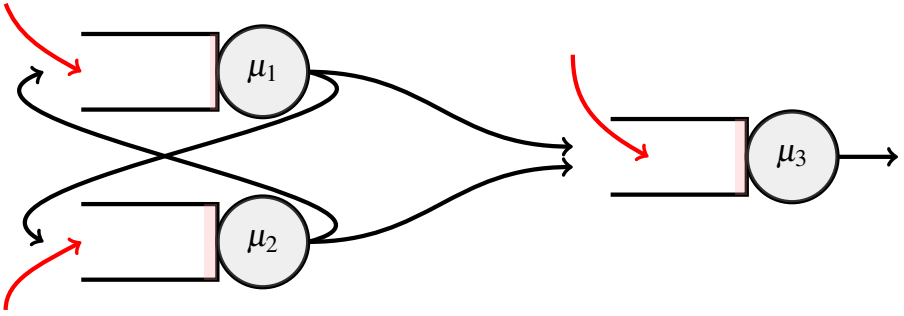
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:

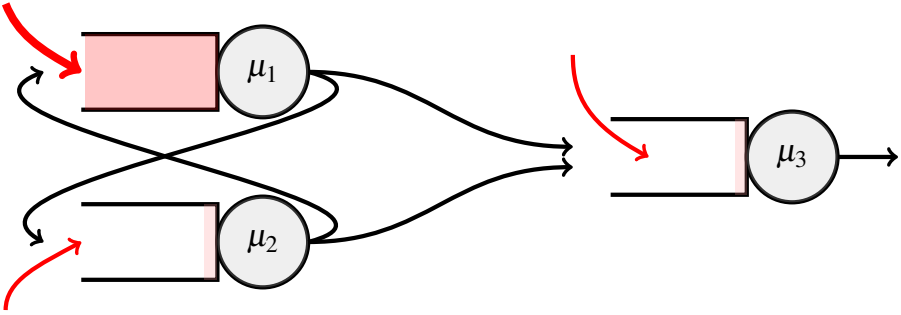


# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:

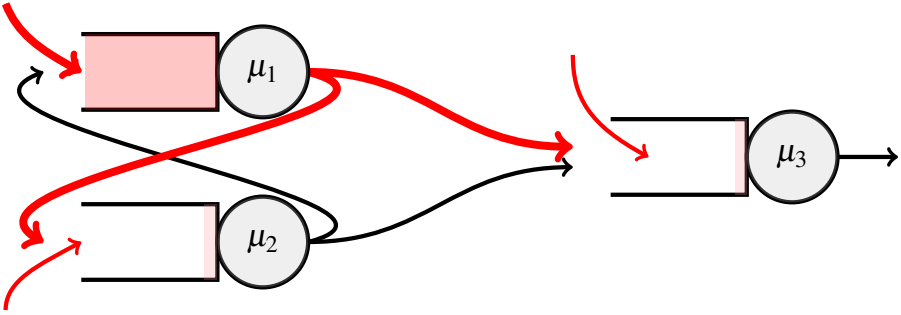


Congestion of Stochastic Fluid Networks:



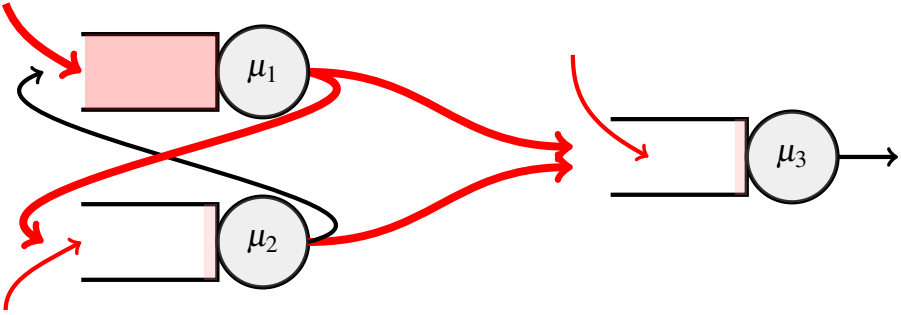
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:

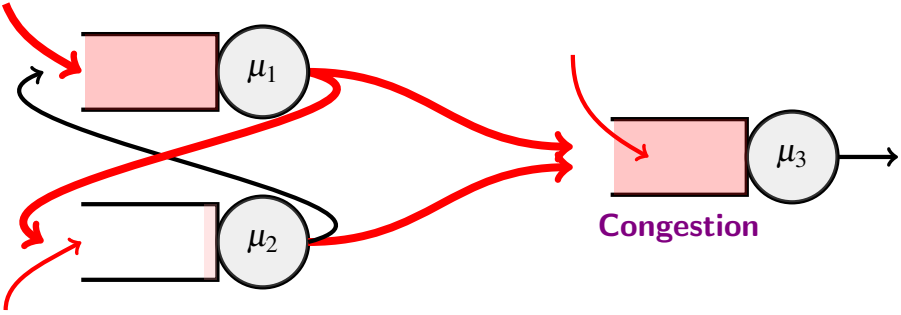


# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



# The point is

**Many heavy-tailed rare events can be written as**

$$\{\bar{S}_n \in A\}$$

**and the decay rate is determined by $\mathcal{J}(A)$, i.e.,**

$$\mathbf{P}(\bar{S}_n \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

# Catastrophe Principle Extends to General Heavy-Tailed Systems

Heavy-Tailed Large Deviations for

- Continuous-Time Processes
  R., Blanchet, Zwart (2019), Bazhba, Blanchet, R., Zwart (2020), Wang, R. (2024+)

- Processes with Spatial and Temporal Correlations
  Bazhba, R., Zwart (2022), Chen, R., Zwart (2024), Bazhba, Blanchet, R., Zwart (2024+), Su, R. (2024+), Wang, R. (2024+)

**Minimal # Jumps added to Typical Paths Characterize the Catastrophe Principle**

# Back to Stochastic Gradient Descent

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W_{k+1} = W_k - \eta \left( f'(W_k) \right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W_{k+1} = W_k - \eta \left( \tilde{f}'(W_k) \right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W_{k+1} = W_k - \eta \left( f'(W_k) + Z_k \right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W_{k+1} = W_k - \eta \left( f'(W_k) + Z_k \right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W^{\eta}{}_{k+1} = W^{\eta}{}_{k} - \eta\left(f'(W^{\eta}{}_{k}) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W^{\eta}(\eta(k+1)) = W^{\eta}(\eta k) - \eta\left(f'(W^{\eta}(\eta k)) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W^{\eta}(\eta(k+1)) = W^{\eta}(\eta k) - \eta\left(f'(W^{\eta}(\eta k)) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

Then

$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W^{\eta}(\eta(k+1)) = W^{\eta}(\eta k) - \eta\left(f'(W^{\eta}(\eta k)) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

Then

$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

where

$$\mathrm{d}w(t) = -f'(w(t))\mathrm{d}t$$

# Stochastic Gradient Descent

Minimizing loss function $f$:

$$W^{\eta}(\eta(k+1)) = W^{\eta}(\eta k) - \eta\left(f'(W^{\eta}(\eta k)) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

Then

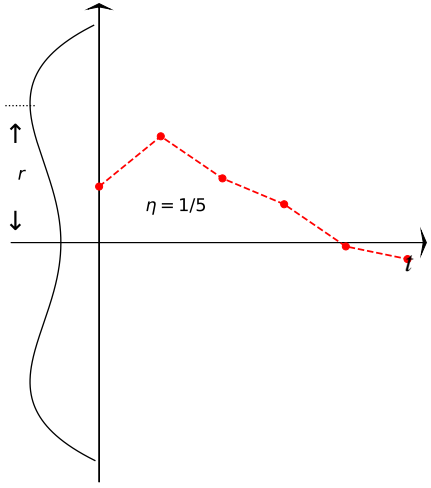
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

Gradient Flow

where

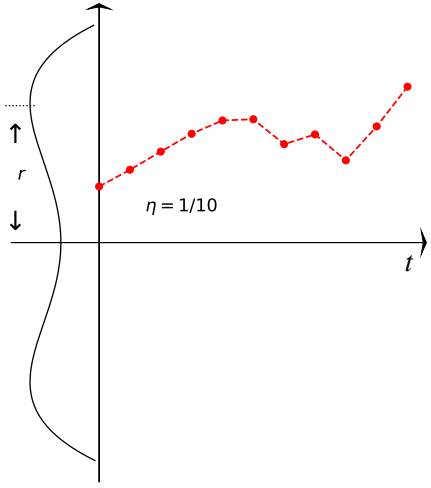
$$\mathrm{d}w(t) = -f'(w(t))\mathrm{d}t$$

# Gradient Flow: Law of Large Numbers



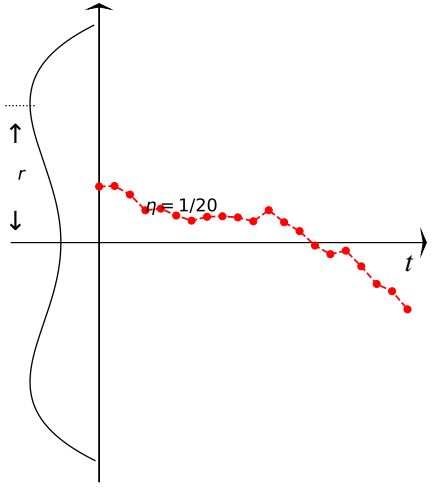

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



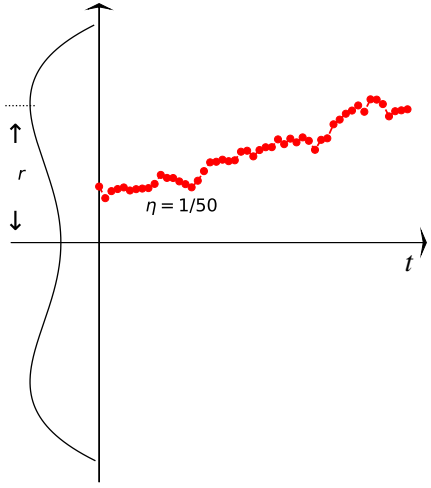

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



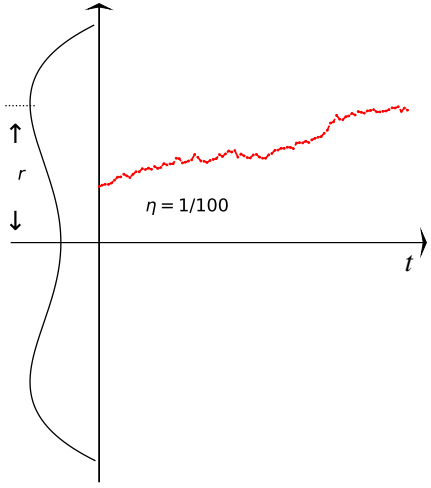

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



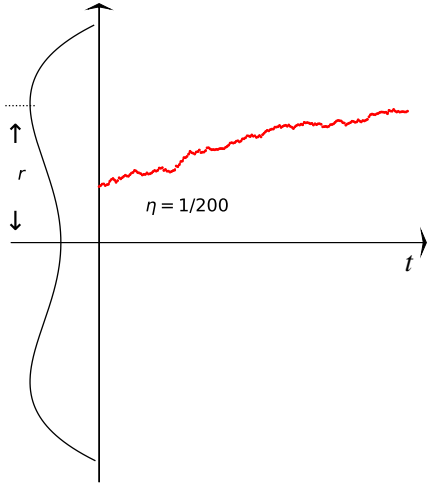

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



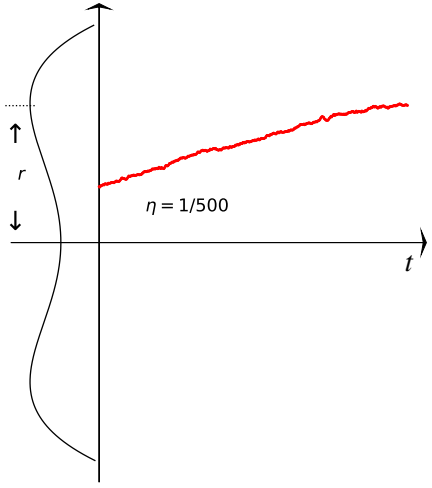

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



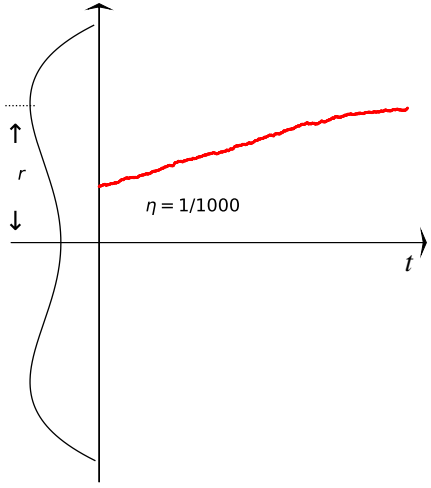

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



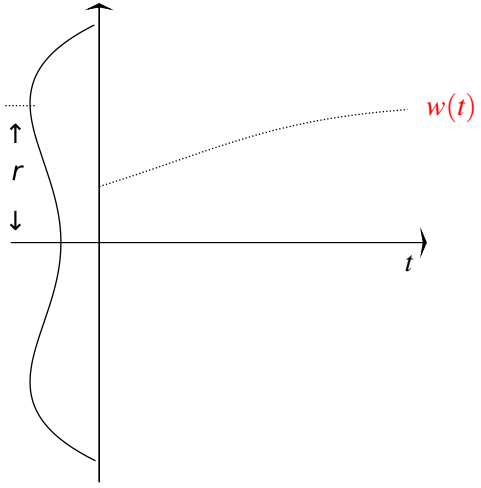

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



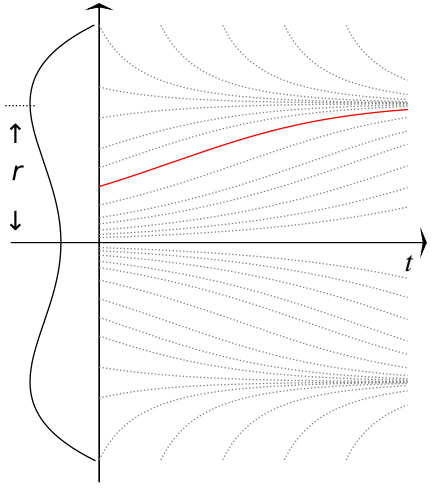

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



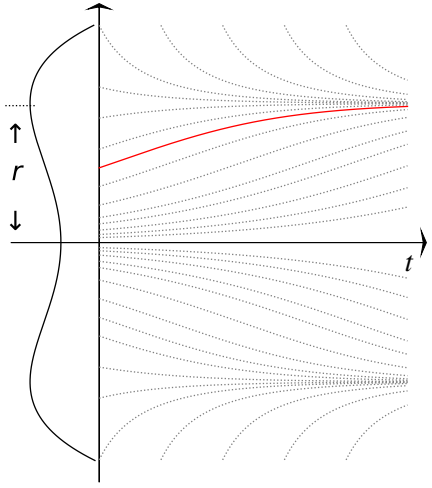

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



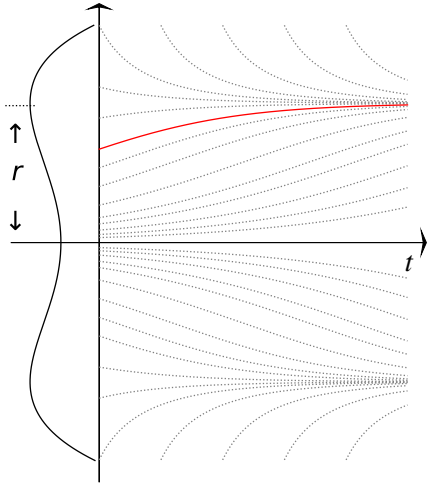

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



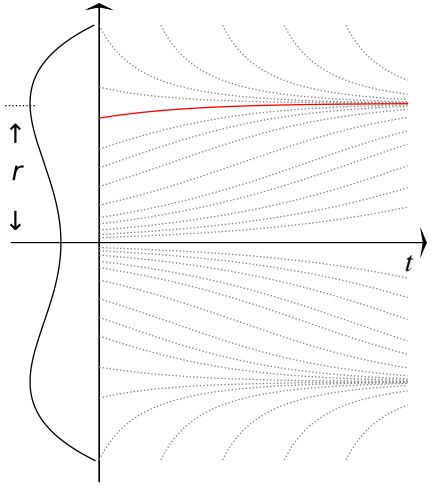

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



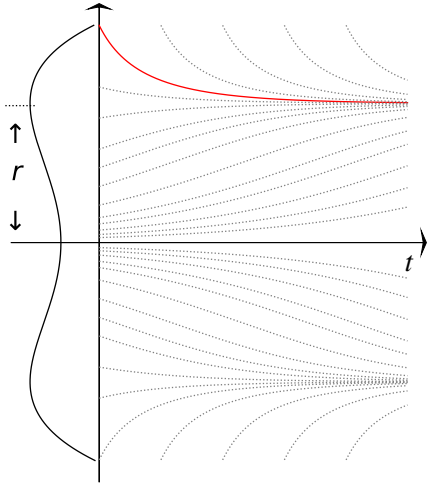

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



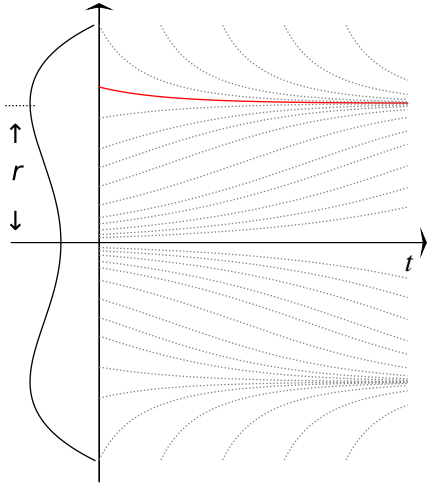

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



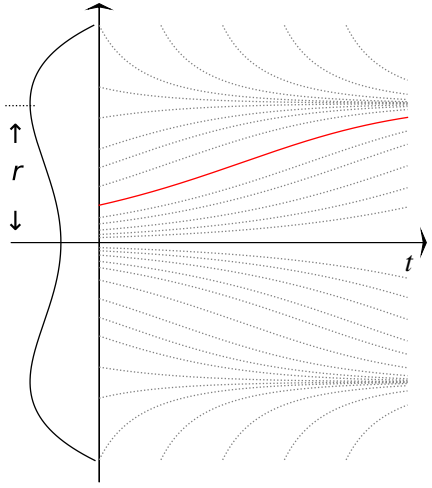

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



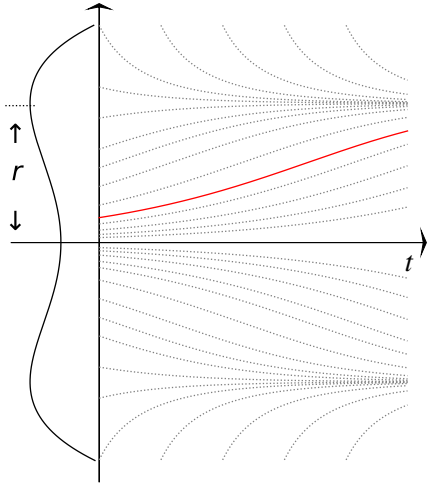

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



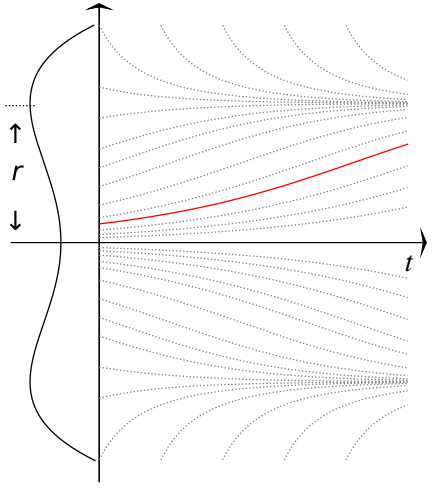

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



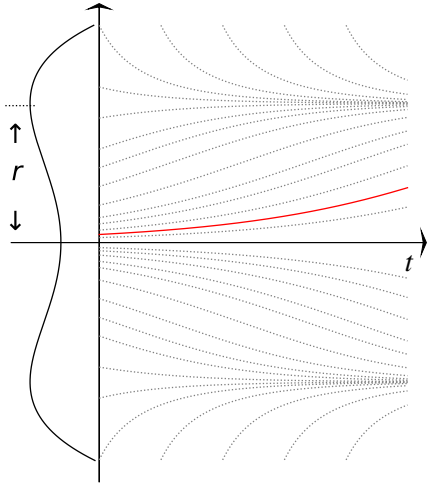

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



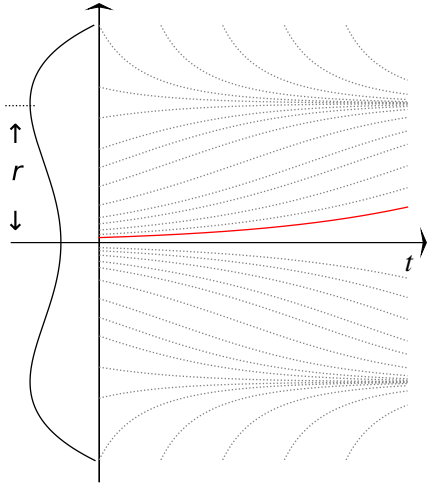

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



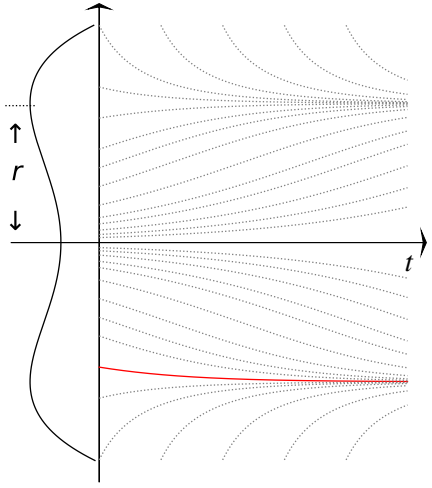

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



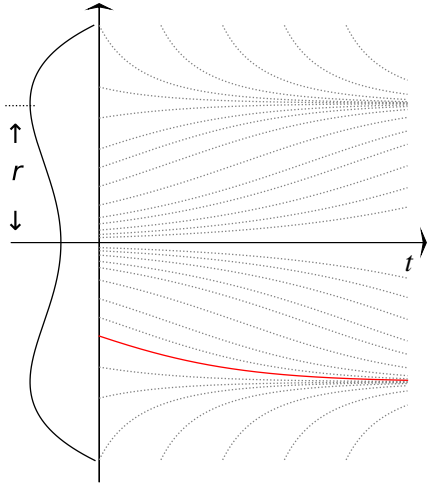

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



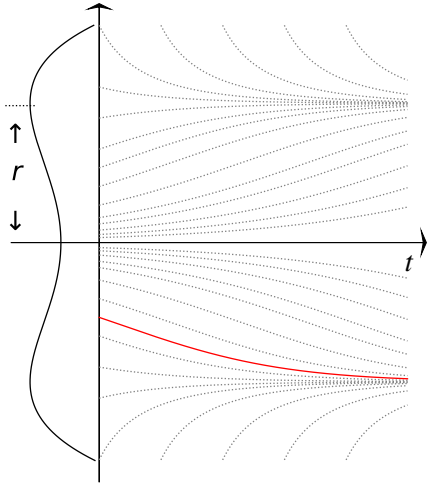

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



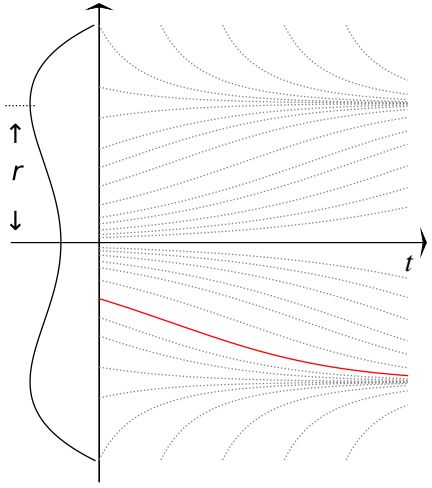

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



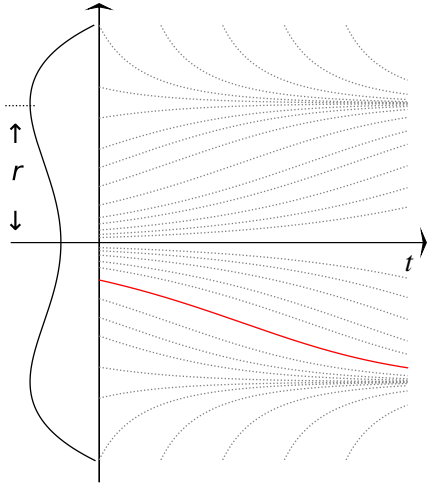

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers




(90° rotation of) $f$

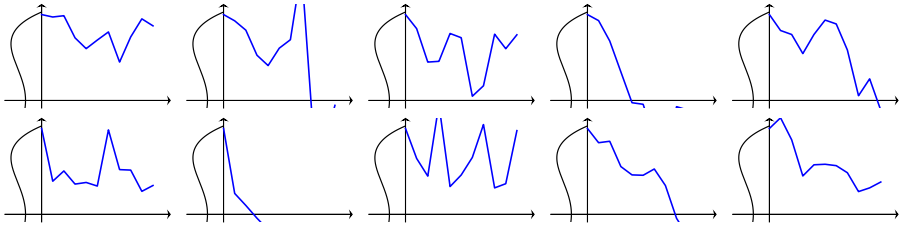
# Gradient Flow: Law of Large Numbers



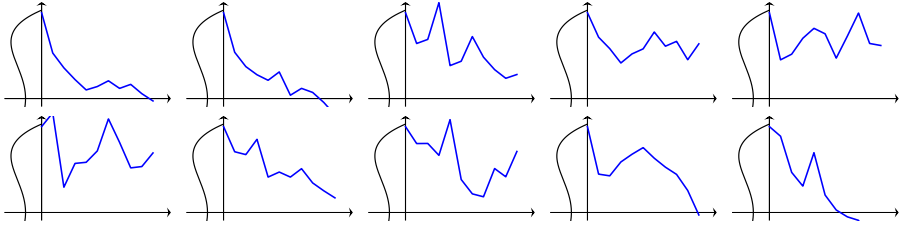

(90° rotation of) $f$

# Typical Scenario

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

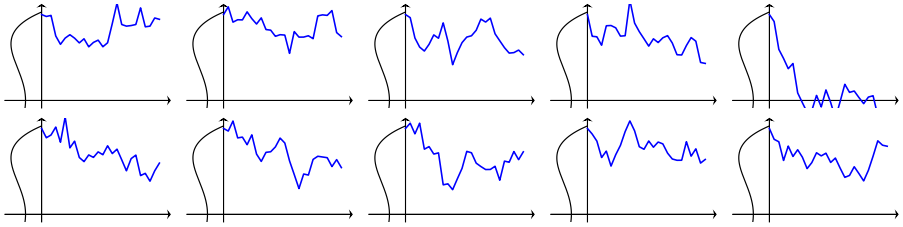


Trajectory of SGD $W^\eta$: $\eta = 1/10$ & noises are **heavy-tailed**

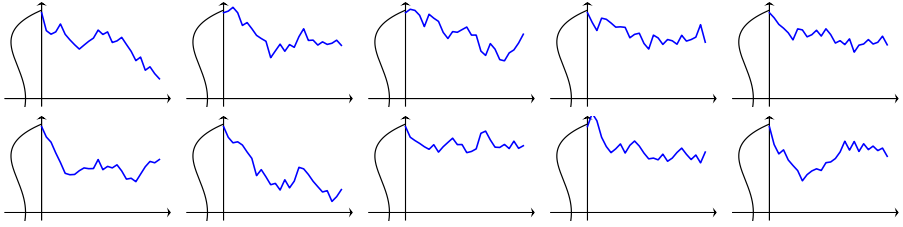


# Typical Scenario

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

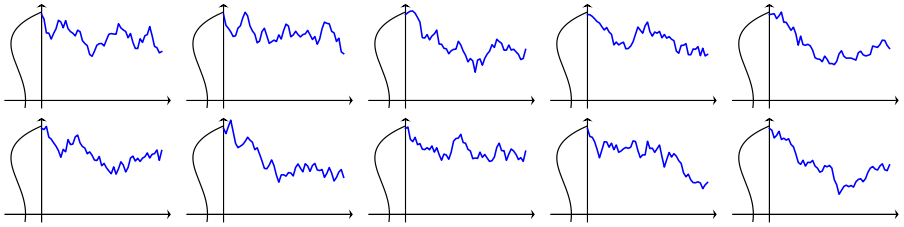


Trajectory of SGD $W^\eta$: $\eta = 1/25$ & noises are **heavy-tailed**

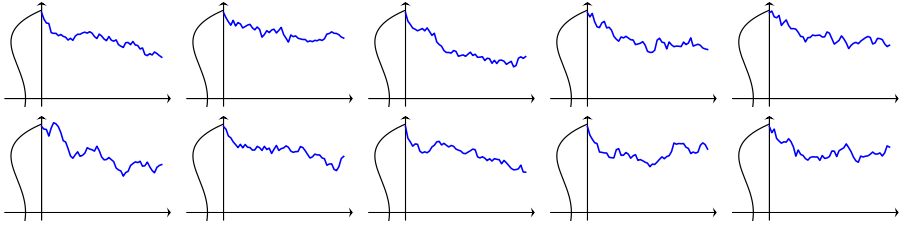


# Typical Scenario

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

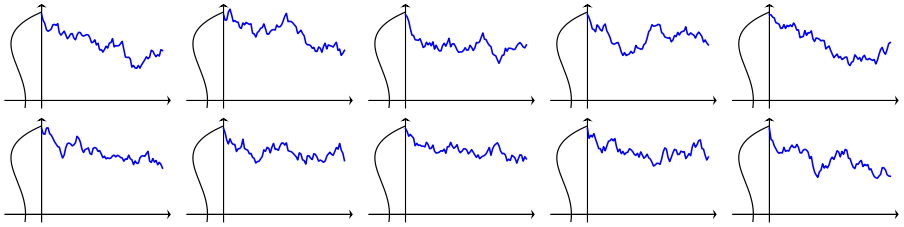


Trajectory of SGD $W^\eta$: $\eta = 1/50$ & noises are **heavy-tailed**

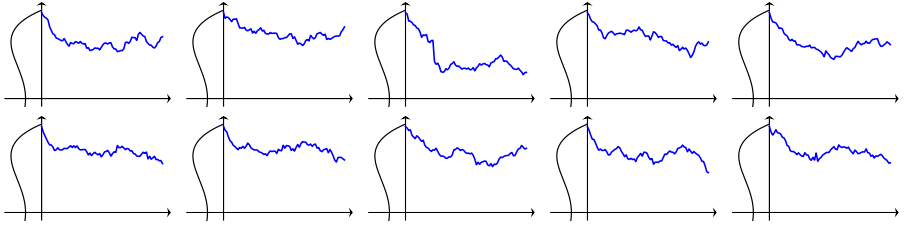


# Typical Scenario

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

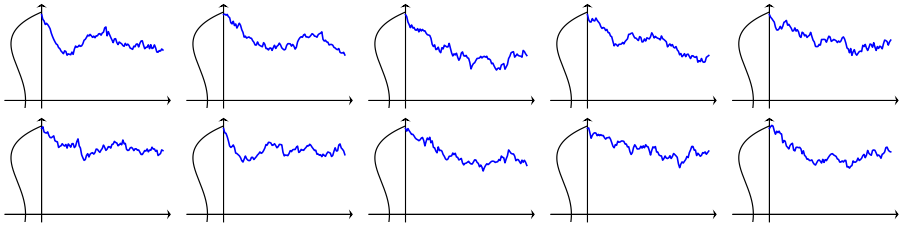


Trajectory of SGD $W^\eta$: $\eta = 1/75$ & noises are **heavy-tailed**

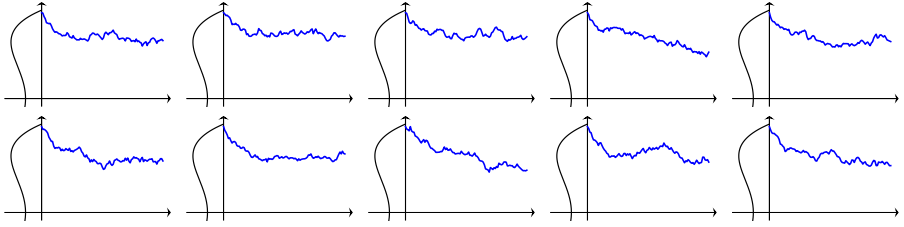


# Typical Scenario

Trajectory of SGD $W^\eta$: $\eta = 1/100$ & noises are **light-tailed**

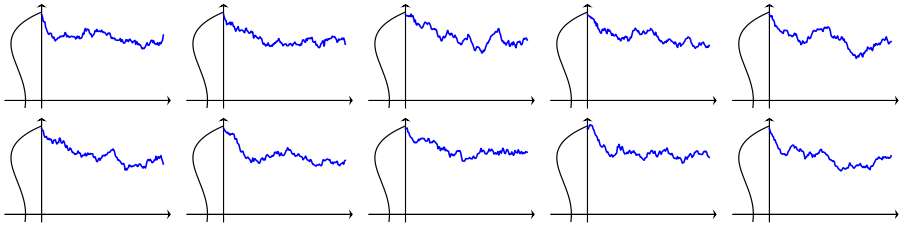


Trajectory of SGD $W^\eta$: $\eta = 1/100$ & noises are **heavy-tailed**

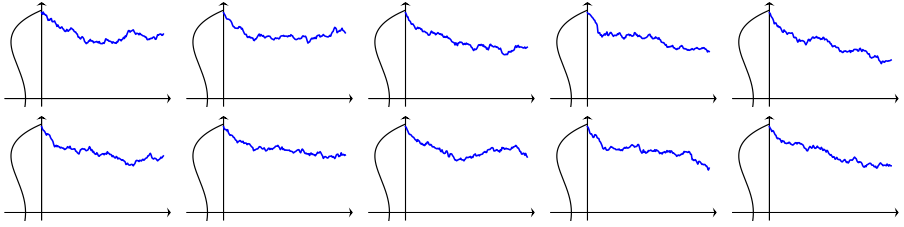


# Typical Scenario

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

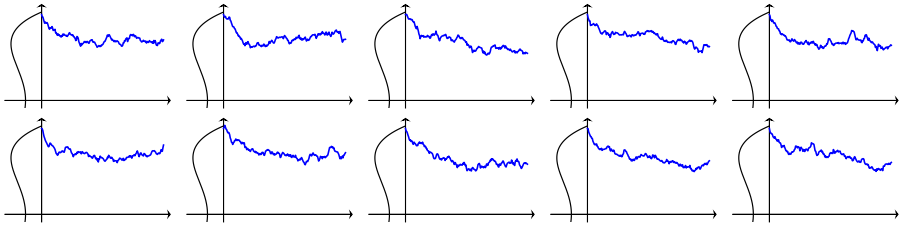


Trajectory of SGD $W^\eta$: $\eta = 1/150$ & noises are **heavy-tailed**

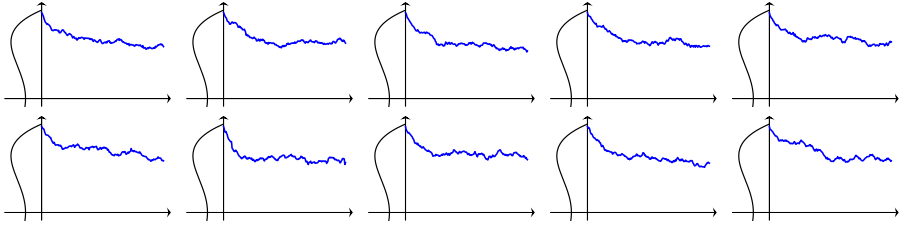


# Typical Scenario

Trajectory of SGD $W^\eta$: $\eta = 1/200$ & noises are **light-tailed**



Trajectory of SGD $W^\eta$: $\eta = 1/200$ & noises are **heavy-tailed**



# Heavy-Tailed Large Deviations for SGD

**Theorem (Wang, Su, R., 2022+)**

*For "general" $A \subseteq \mathbb{D}$*

$$C(A^\circ) \leq \liminf_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq \limsup_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq C(A^-).$$

- $\mathcal{J}(A)$*: min #jumps added to* $w(\cdot)$ *for it to be inside* $A$
- $C(\cdot)$*: a measure*

# Heavy-Tailed Large Deviations for SGD

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, Su, R., 2022+)**

*For "general" $A \subseteq \mathbb{D}$*

$$C(A^\circ) \leq \liminf_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq \limsup_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq C(A^-).$$

- $\mathcal{J}(A)$: *min #jumps added to $w(\cdot)$ for it to be inside $A$*
- $C(\cdot)$: *a measure*

# Heavy-Tailed Large Deviations for SGD

gradient noise

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, Su, R., 2022+)**

*For "general" $A \subseteq \mathbb{D}$*

$$C(A^\circ) \leq \liminf_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq \limsup_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in A)}{\eta^{\alpha \mathcal{J}(A)}} \leq C(A^-).$$

- $\mathcal{J}(A)$: *min #jumps added to $w(\cdot)$ for it to be inside $A$*
- $C(\cdot)$: *a measure*

# Heavy-Tailed Large Deviations for SGD

gradient noise

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

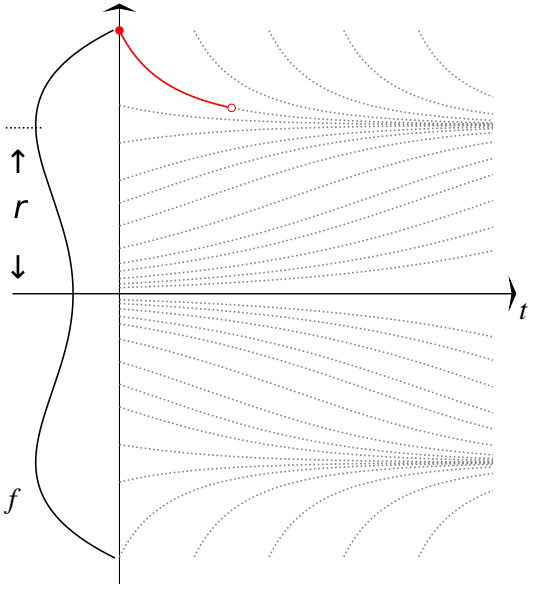
**Theorem (Wang, Su, R., 2022+)**

*For "general" $A \subseteq \mathbb{D}$*

$$\mathbf{P}(W^\eta \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

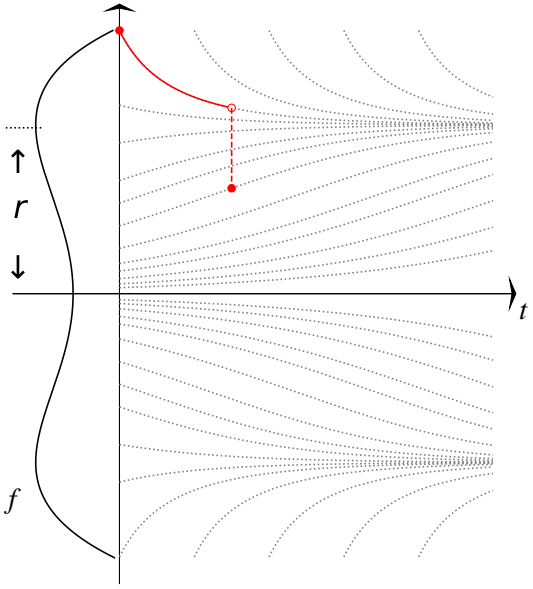
- $\mathcal{J}(A)$: *min #jumps added to $w(\cdot)$ for it to be inside $A$*

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



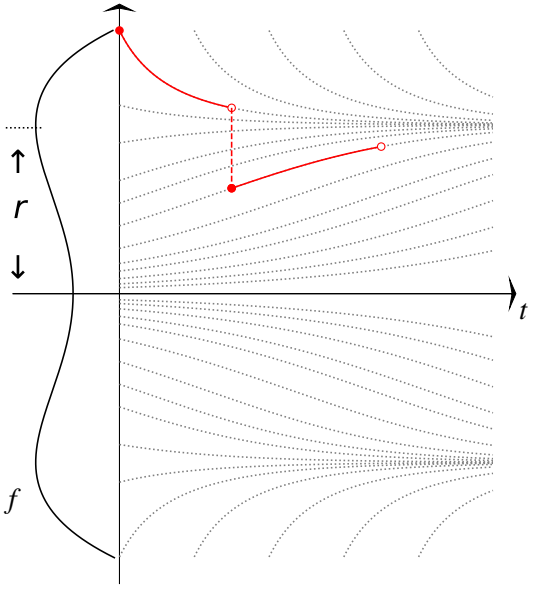

(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



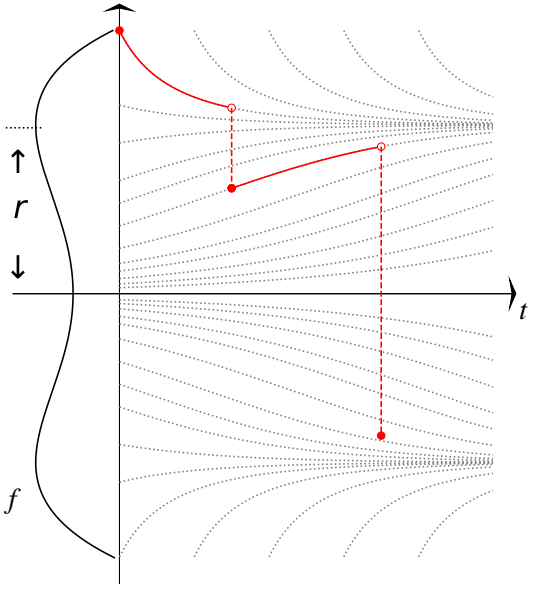

(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



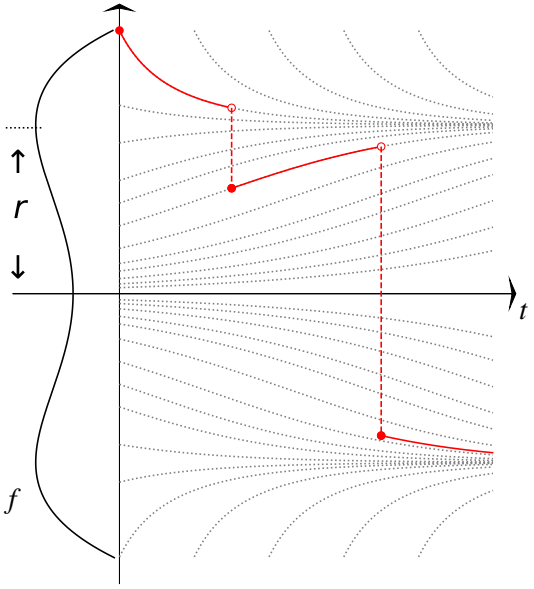

(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



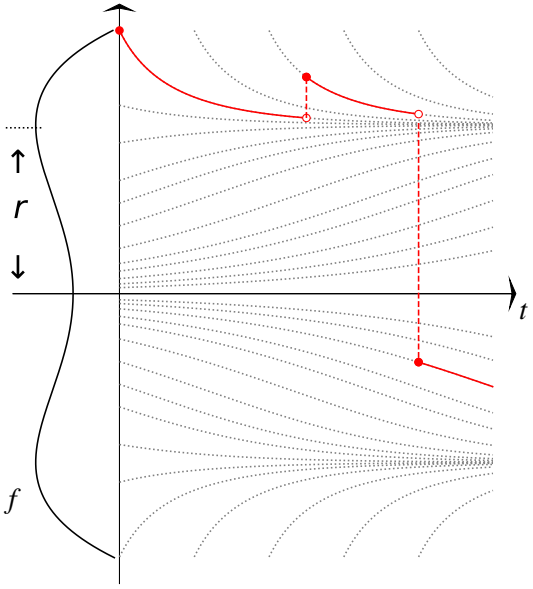

(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



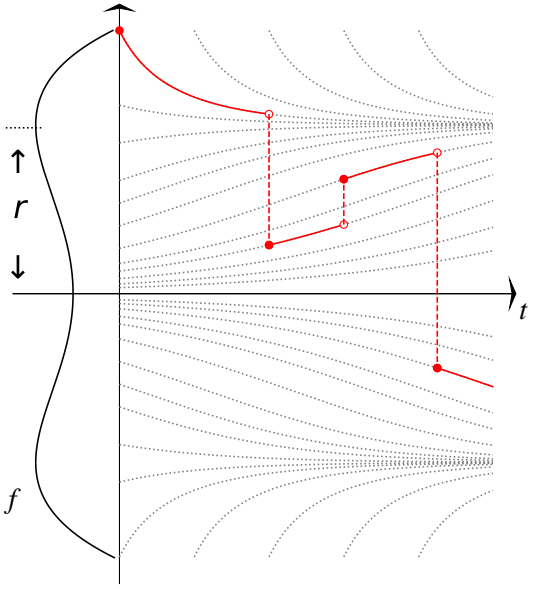

(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow




(90° rotation of) $f$

# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow




(90° rotation of) $f$

# Recall: Heavy-Tailed Large Deviations for SGD

**Theorem (Wang, Su, R., 2022+)**

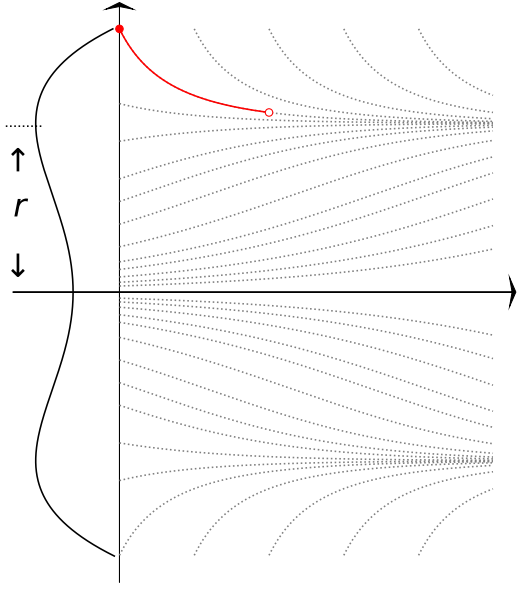
*For "general" $A \subseteq \mathbb{D}$*

$$\mathbf{P}(W^{\eta} \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

- $\mathcal{J}(A)$: *min #jumps added to $w(\cdot)$ for it to be inside $A$*

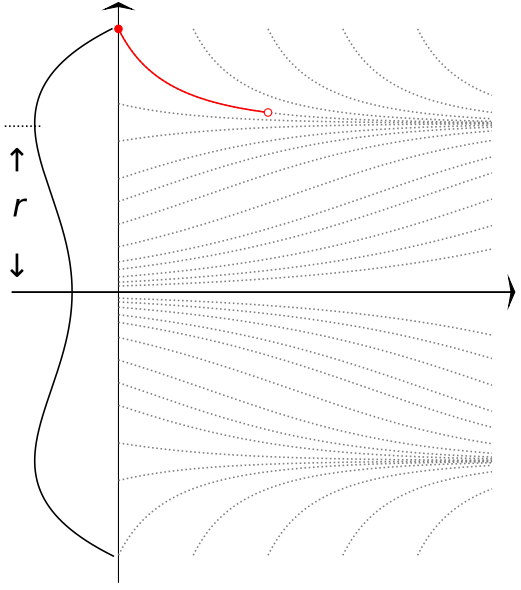
# Catastrophe Principle: Most Likely Escape Route

Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



# Catastrophe Principle: Most Likely Escape Route

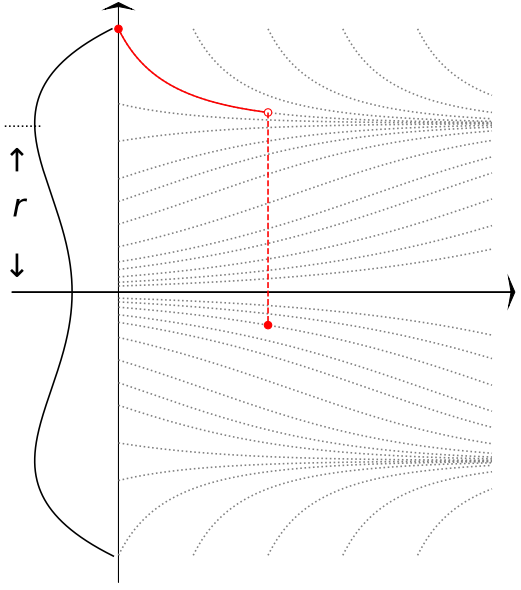
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

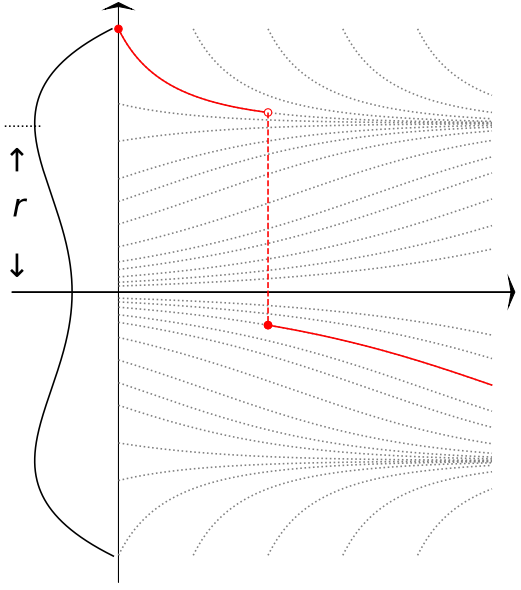
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

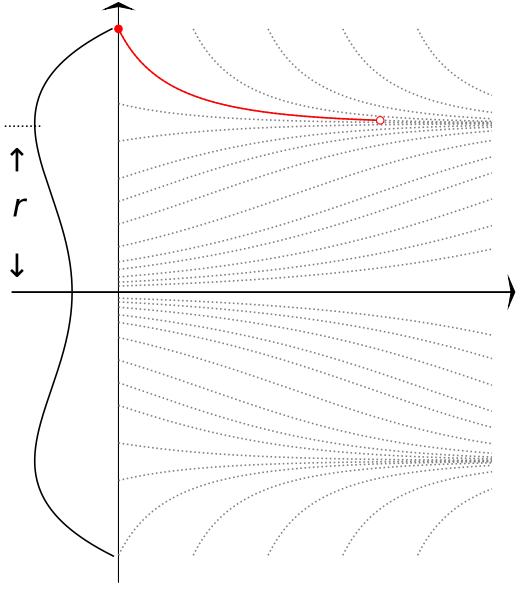
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

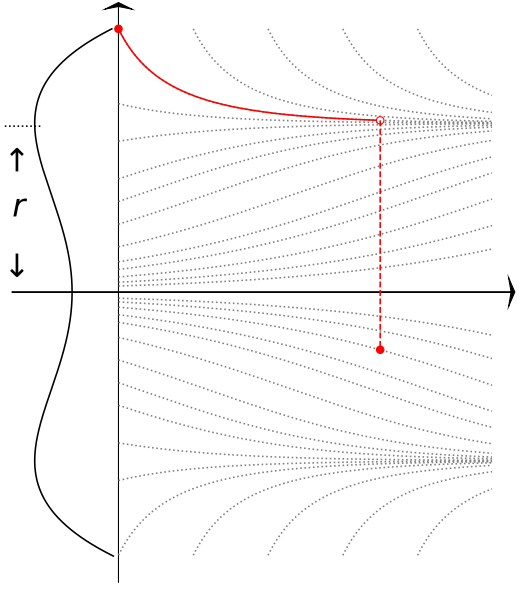
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

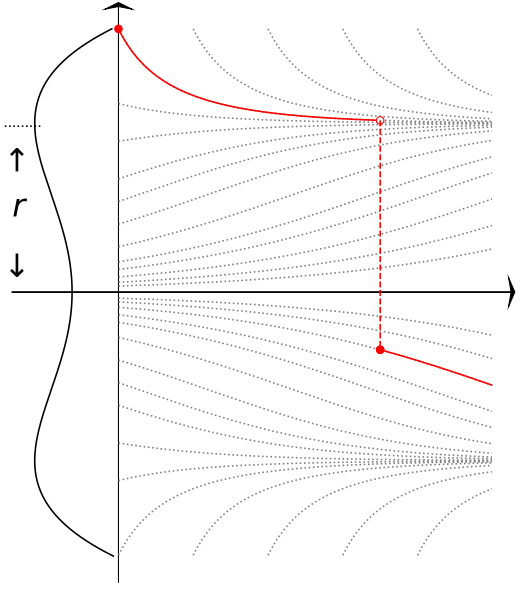
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

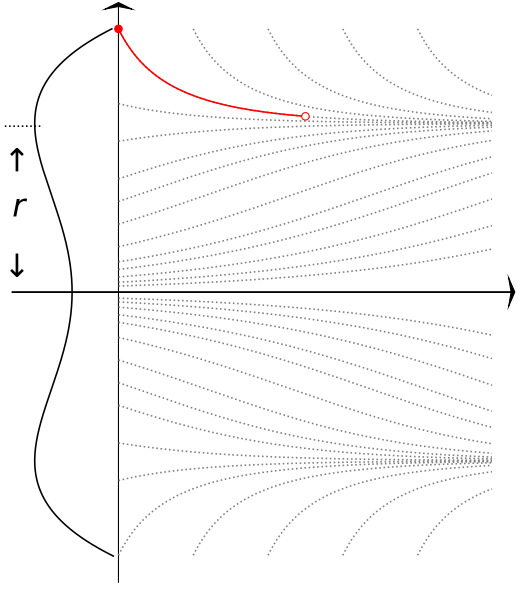
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

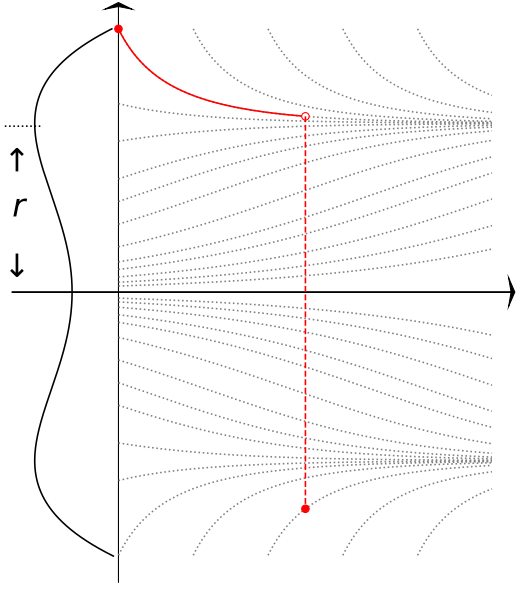
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

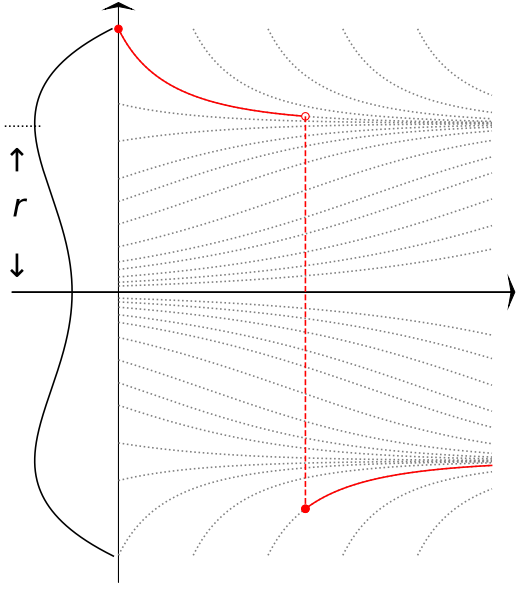
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$




$\mathcal{J}(A) = 1$

# Catastrophe Principle: Most Likely Escape Route

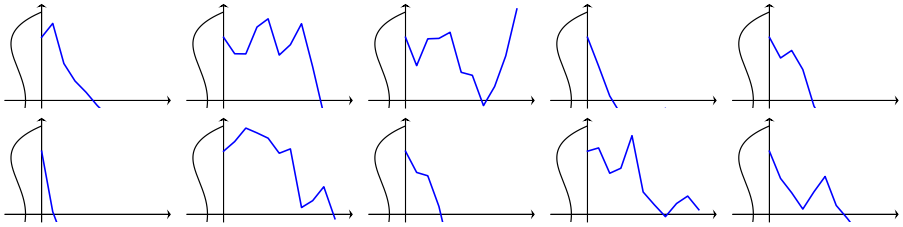
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



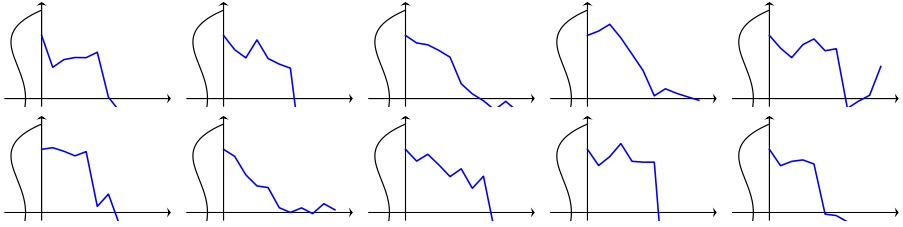

$\mathcal{J}(A) = 1$

# Catastrophe Principle Dictates SGD's Escape Route

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

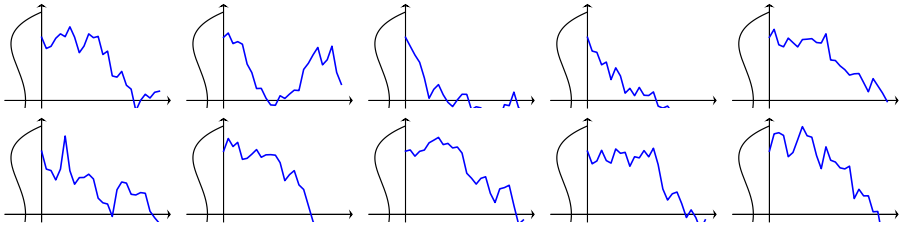


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

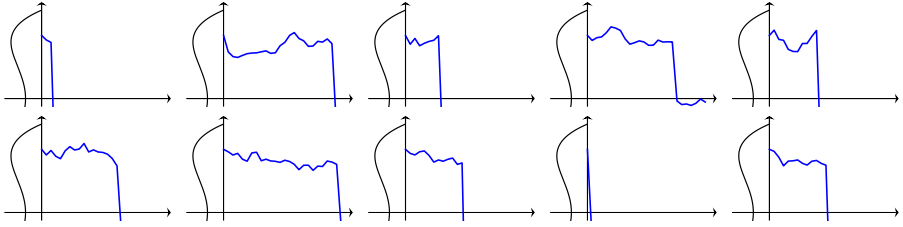


# Catastrophe Principle Dictates SGD's Escape Route

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

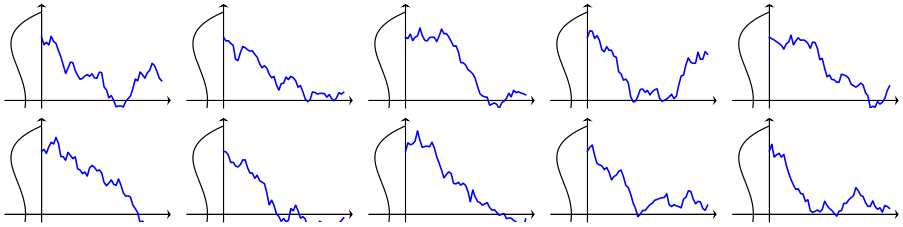


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

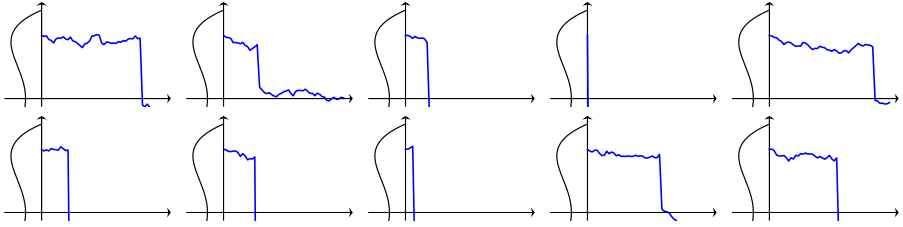


# Catastrophe Principle Dictates SGD's Escape Route

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

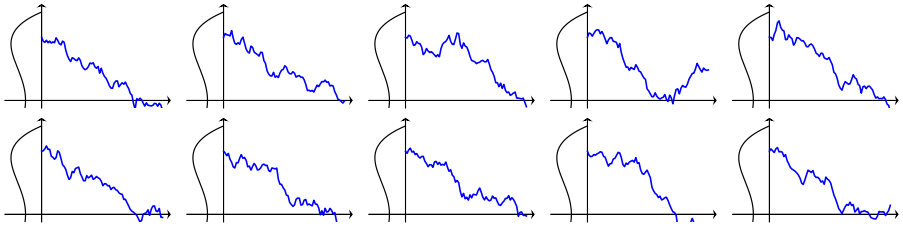


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/50$

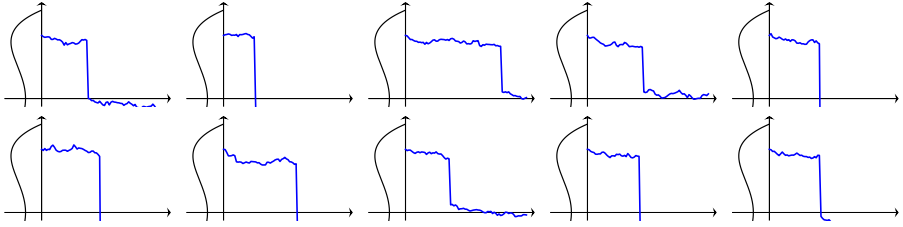


# Catastrophe Principle Dictates SGD's Escape Route

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

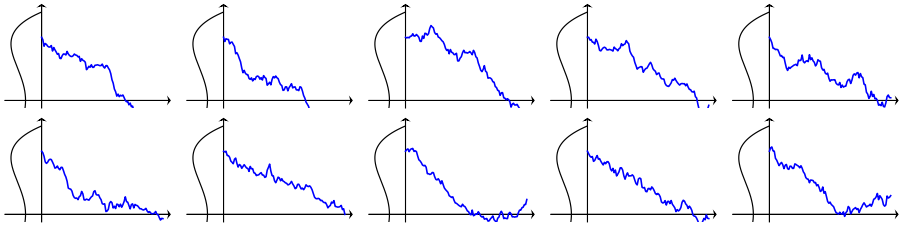


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

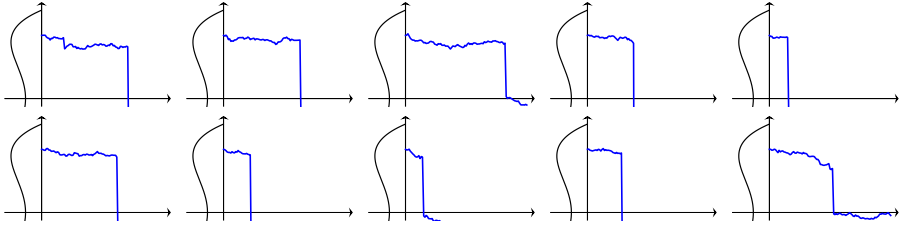


# Catastrophe Principle Dictates SGD's Escape Route

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

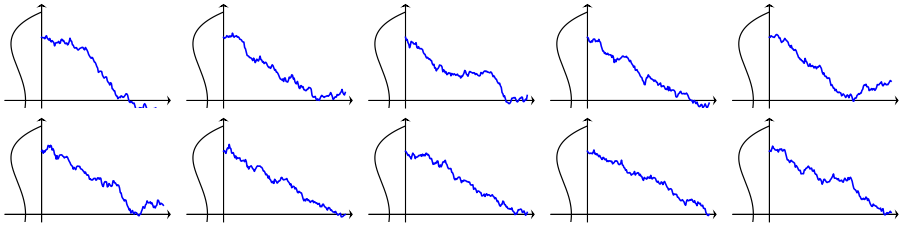


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

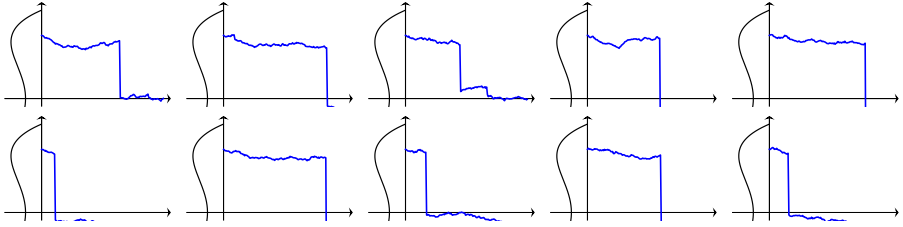


# Catastrophe Principle Dictates SGD's Escape Route

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

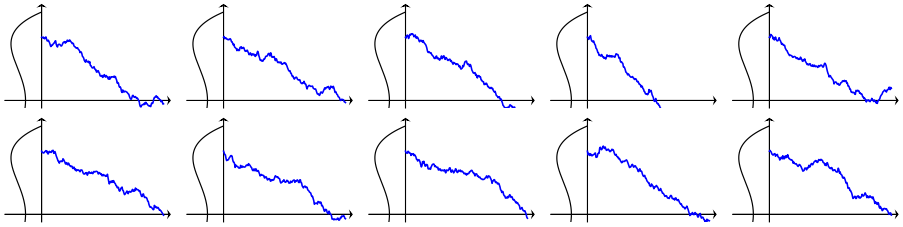


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

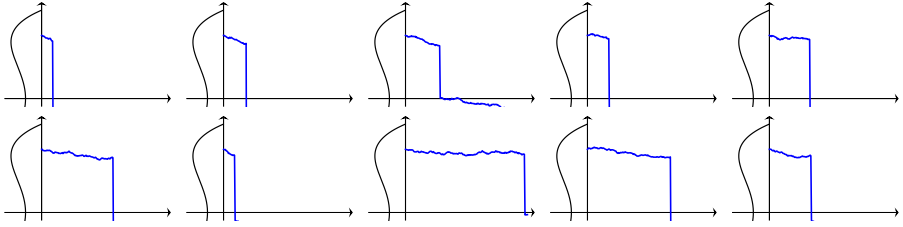


# Catastrophe Principle Dictates SGD's Escape Route

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



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



# Truncated Version of Stochastic Gradient Descent

SGD

$$W_{k+1}^{\eta} = W_k^{\eta} - \qquad \eta\left(f'(W_k^{\eta}) + Z_k\right) \qquad k = 0, 1, 2, \ldots$$

# Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

$$W_{k+1}^{\eta} = W_k^{\eta} - \varphi_c\left(\eta\left(f'(W_k^{\eta}) + Z_k\right)\right) \qquad k = 0, 1, 2, \ldots$$

# Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

$$W_{k+1}^{\eta} = W_k^{\eta} - \varphi_c\left(\eta\left(f'(W_k^{\eta}) + Z_k\right)\right) \qquad k = 0, 1, 2, \ldots$$

where

$$\varphi_c(x) = \frac{x}{|x|}\min\{c, |x|\}.$$

# Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

$$W_{k+1}^{\eta} = W_k^{\eta} - \varphi_c\left(\eta\left(f'(W_k^{\eta}) + Z_k\right)\right) \qquad k = 0, 1, 2, \ldots$$

where

$$\varphi_c(x) = \frac{x}{|x|}\min\{c, |x|\}.$$

Then, again,

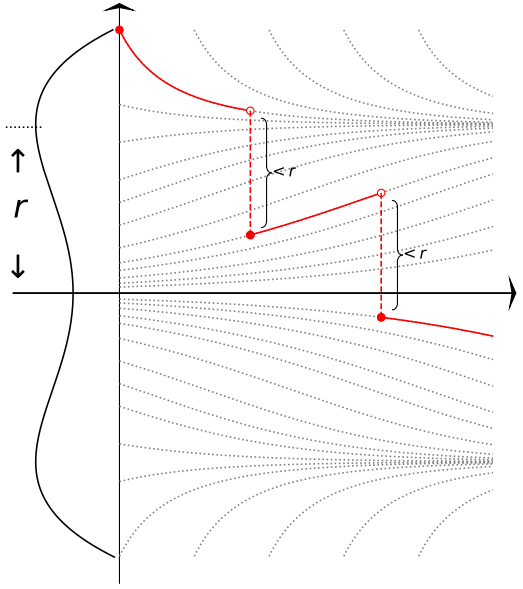
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

where

$$\mathrm{d}w(t) = -f'(w(t))\mathrm{d}t.$$

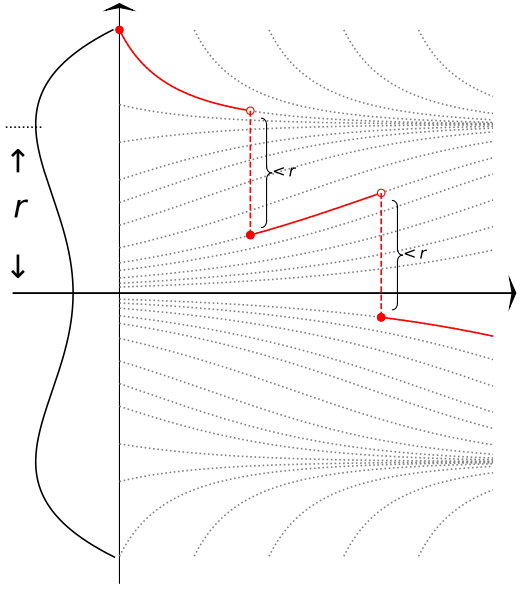
# How does $\mathcal{J}(A)$ change?

If $r \in (c, 2c)$



# How does $\mathcal{J}(A)$ change?

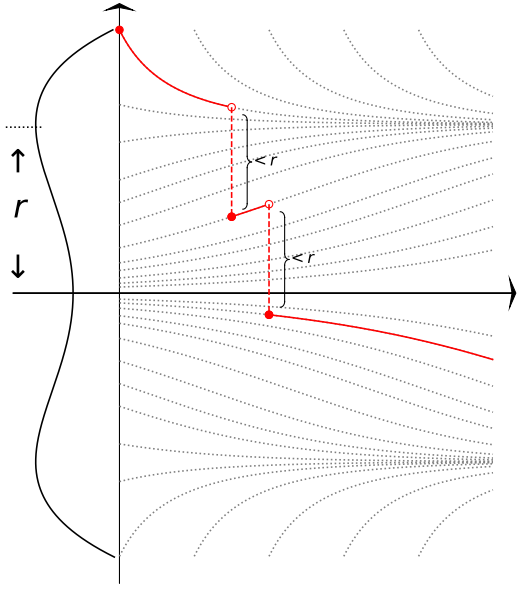
If $r \in (c, 2c)$




$\mathcal{J}(A) = 2$

# How does $\mathcal{J}(A)$ change?

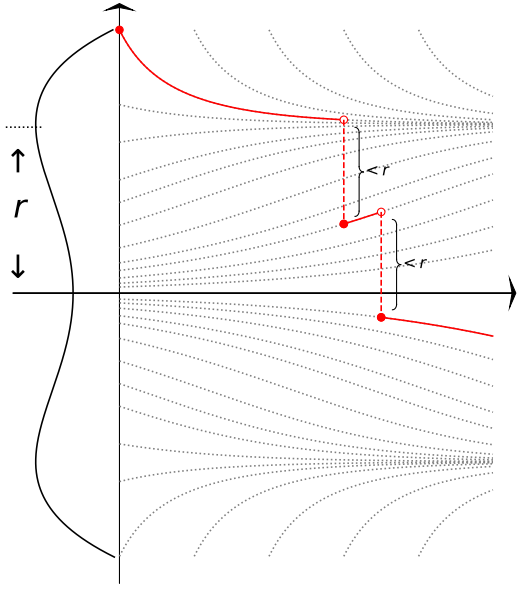
If $r \in (c, 2c)$




$\mathcal{J}(A) = 2$

# How does $\mathcal{J}(A)$ change?

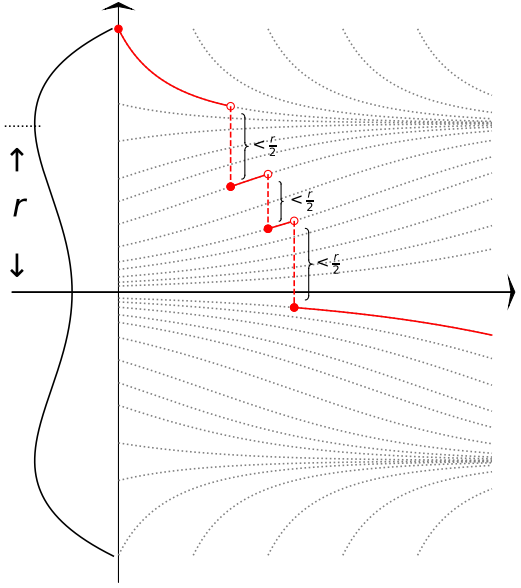
If $r \in (c, 2c)$




$\mathcal{J}(A) = 2$

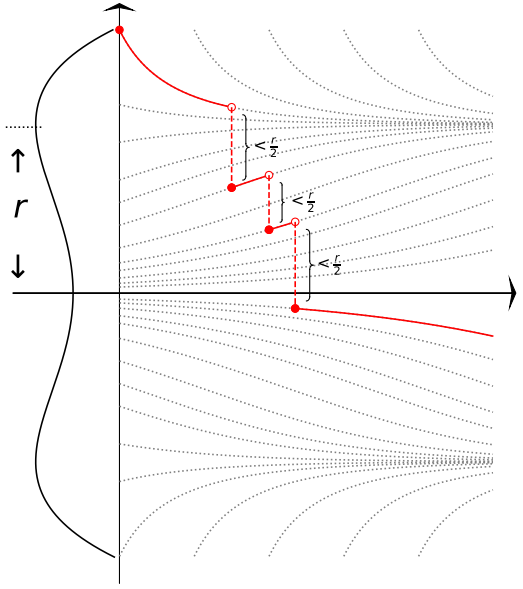
# How does $\mathcal{J}(A)$ change?

If $r \in (2c, 3c)$



# How does $\mathcal{J}(A)$ change?

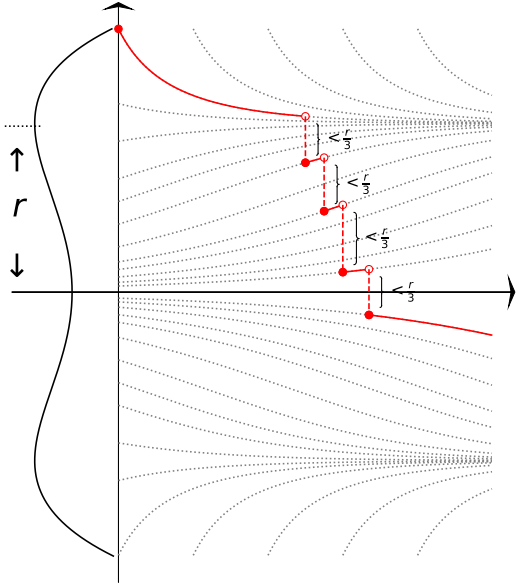
If $r \in (2c, 3c)$




$\mathcal{J}(A) = 3$

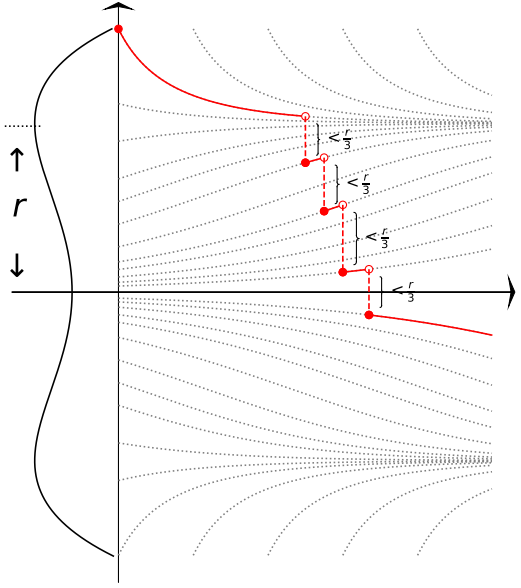
# How does $\mathcal{J}(A)$ change?

If $r \in (3c, 4c)$



# How does $\mathcal{J}(A)$ change?

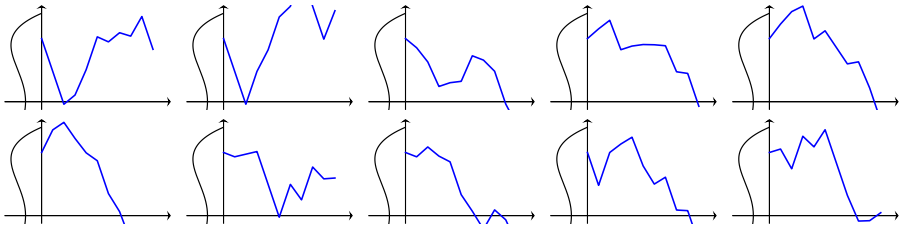
If $r \in (3c, 4c)$



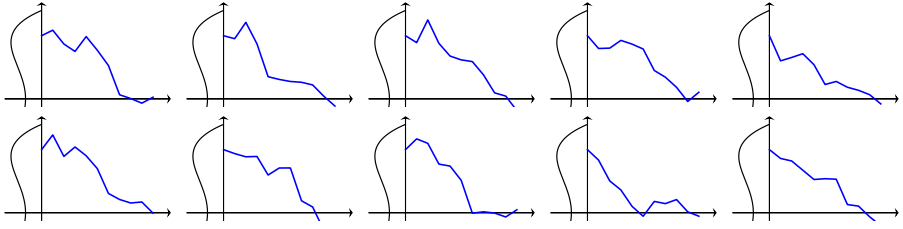

$\mathcal{J}(A) = 4$

# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

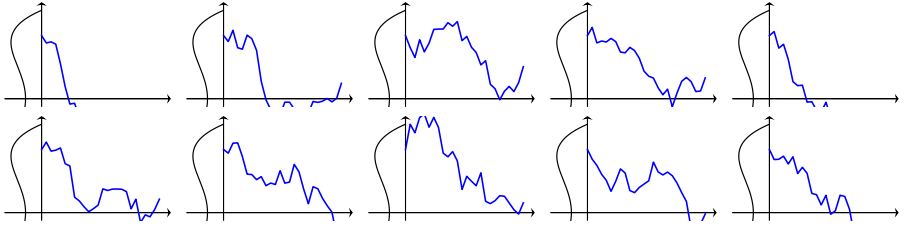


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

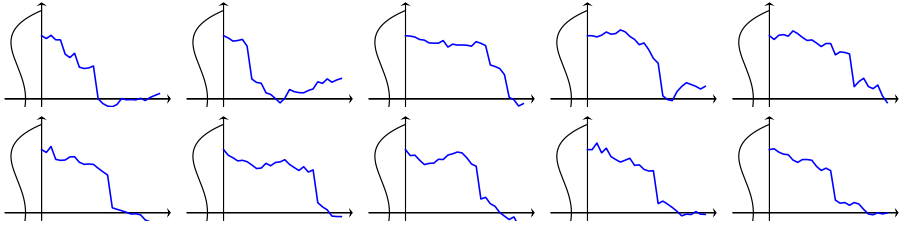


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

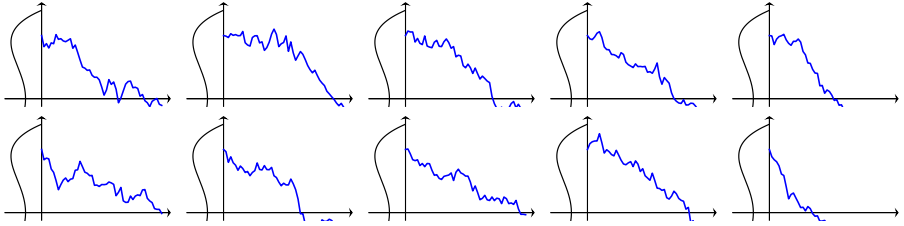


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

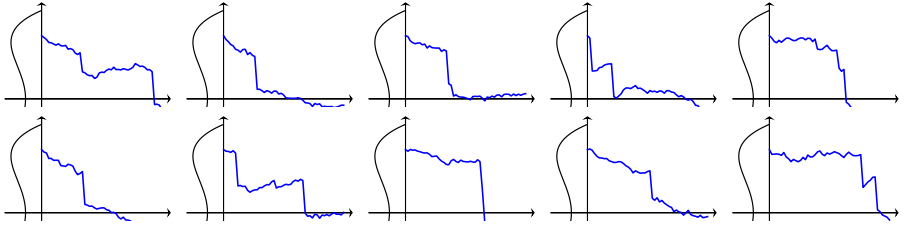


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

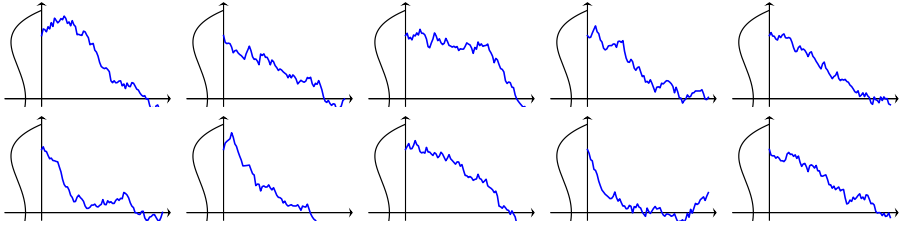


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

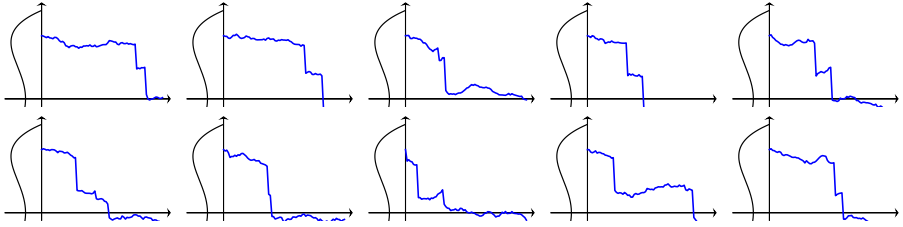


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

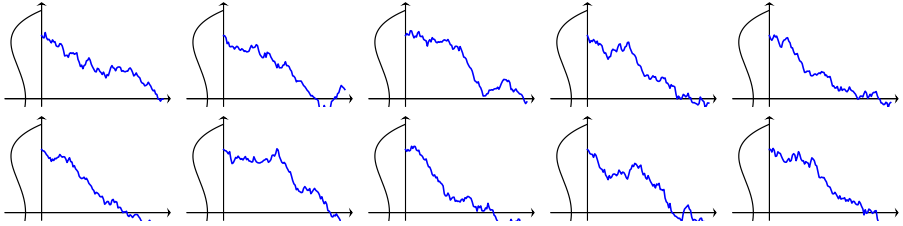


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

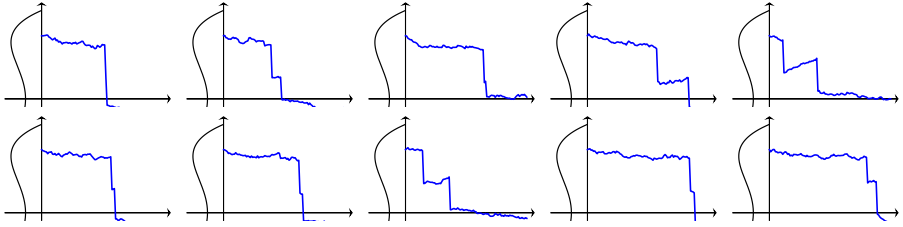


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

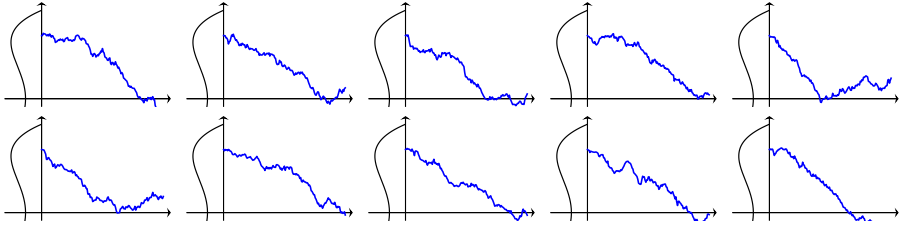


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

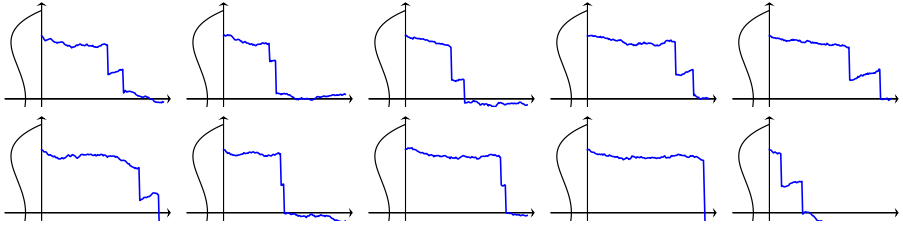


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

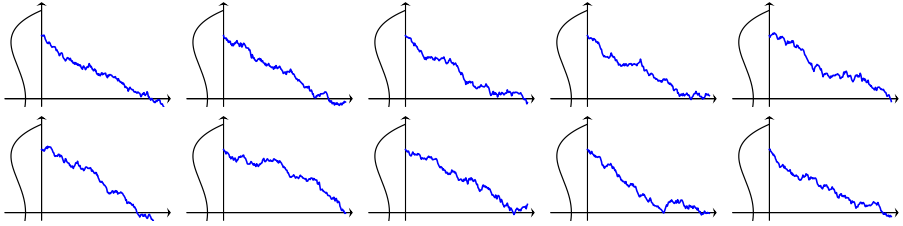


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

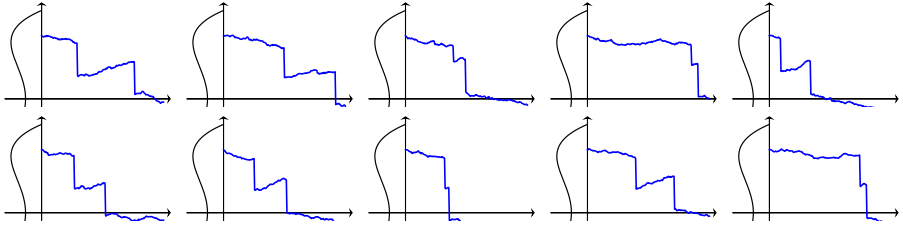


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

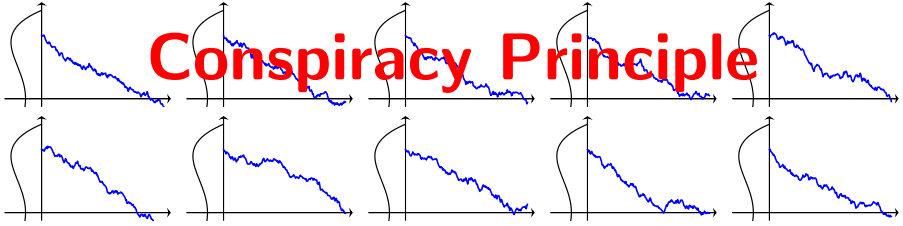


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/200$

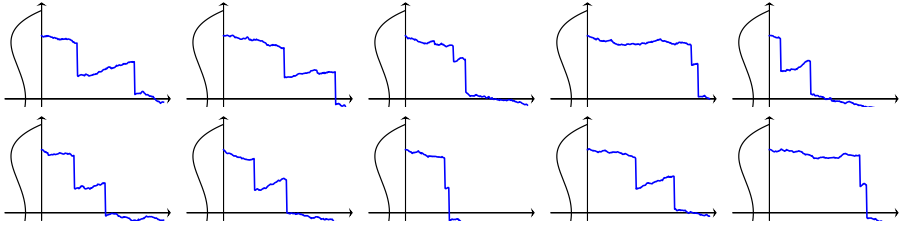


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

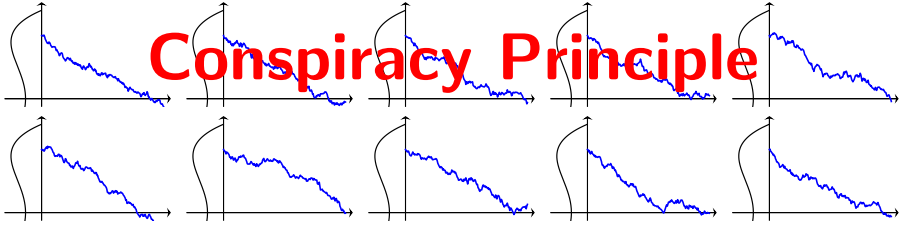



Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/200$

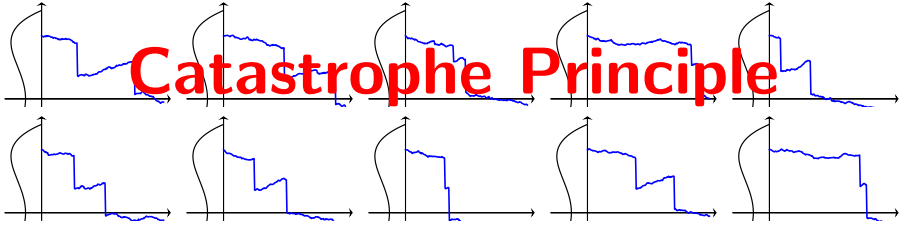


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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




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



# Metastability of SGD

# Heavy-Tailed Large Deviations for SGD

**Theorem (Wang, R., 2024+)**

*For "general" $B \subseteq \mathbb{D}[0,T]$*

$$C(B^\circ) \leq \liminf_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq \limsup_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq C(B^-).$$

- $\mathcal{J}(B)$: *min #jumps added to $w(\cdot)$ for it to be inside $B$*
- $C(\cdot)$: *a measure*

# Heavy-Tailed Large Deviations for SGD

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, R., 2024+)**

*For "general" $B \subseteq \mathbb{D}[0,T]$*

$$C(B^\circ) \leq \liminf_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq \limsup_{\eta \to 0} \frac{\mathbf{P}(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq C(B^-).$$

- $\mathcal{J}(B)$: *min #jumps added to $w(\cdot)$ for it to be inside $B$*
- $C(\cdot)$: *a measure*

# Heavy-Tailed Large Deviations for SGD

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, R., 2024+)**

*For "general" $B \subseteq \mathbb{D}[0,T]$*

$$C(B^\circ) \leq \lim_{\varepsilon \to 0} \liminf_{\eta \to 0} \frac{\inf_{x \in A(\varepsilon)} \mathbf{P}_x(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq \lim_{\varepsilon \to 0} \limsup_{\eta \to 0} \frac{\sup_{x \in A(\varepsilon)} \mathbf{P}_x(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq C(B^-).$$

- $\mathcal{J}(B)$: *min #jumps added to $w(\cdot)$ for it to be inside $B$*
- $C(\cdot)$: *a measure*

# Heavy-Tailed Large Deviations for SGD

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, R., 2024+)**

*For "general" $B \subseteq \mathbb{D}[0,T]$*

$$C(B^\circ) \leq \lim_{\varepsilon \to 0} \liminf_{\eta \to 0} \frac{\inf_{x \in A(\varepsilon)} \mathbf{P}_x(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq \lim_{\varepsilon \to 0} \limsup_{\eta \to 0} \frac{\sup_{x \in A(\varepsilon)} \mathbf{P}_x(W^\eta \in B)}{\eta^{\alpha \mathcal{J}(B)}} \leq C(B^-).$$

- $\mathcal{J}(B)$: *min #jumps added to $w(\cdot)$ for it to be inside $B$*
- $C(\cdot)$: *a measure*

**Locally Uniform Large Deviations over Asymptotic Atom** $\{A(\varepsilon) : \varepsilon > 0\}$

# $\mathbb{M}$-Convergence

$$\mathbb{M}(\mathbb{S}\setminus\mathbb{C}) = \{\nu(\cdot) : \nu(\mathbb{S}\setminus\mathbb{C}^{\varepsilon}) < \infty, \forall \varepsilon > 0; \quad \nu(\cdot) \text{ Borel measure on } \mathbb{S}\setminus\mathbb{C}\}$$

**Definition ($\mathbb{M}$-convergence; Lindskog, Resnick, Roy., 2014)**

Let $\mu^{\eta}, \mu \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$. We say that $\mu^{\eta}$ converges to $\mu$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ as $\eta \to 0$ if

$$\lim_{\eta\downarrow 0} |\mu^{\eta}(f) - \mu(f)| = 0 \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

# $\mathbb{M}$-Convergence

$\varepsilon$-fattening of $\mathbb{C}$

$$\mathbb{M}(\mathbb{S}\setminus\mathbb{C}) = \{\nu(\cdot) : \nu(\mathbb{S}\setminus\mathbb{C}^{\varepsilon}) < \infty, \forall \varepsilon > 0; \quad \nu(\cdot) \text{ Borel measure on } \mathbb{S}\setminus\mathbb{C}\}$$

**Definition ($\mathbb{M}$-convergence; Lindskog, Resnick, Roy., 2014)**

Let $\mu^{\eta}, \mu \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$. We say that $\mu^{\eta}$ converges to $\mu$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ as $\eta \to 0$ if

$$\lim_{\eta\downarrow 0} |\mu^{\eta}(f) - \mu(f)| = 0 \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

# $\mathbb{M}$-Convergence

$\varepsilon$-fattening of $\mathbb{C}$

$$\mathbb{M}(\mathbb{S}\setminus\mathbb{C}) = \{\nu(\cdot) : \nu(\mathbb{S}\setminus\mathbb{C}^{\varepsilon}) < \infty, \forall \varepsilon > 0; \quad \nu(\cdot) \text{ Borel measure on } \mathbb{S}\setminus\mathbb{C}\}$$

**Definition ($\mathbb{M}$-convergence; Lindskog, Resnick, Roy., 2014)**

Let $\mu^{\eta}, \mu \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$. We say that $\mu^{\eta}$ converges to $\mu$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ as $\eta \to 0$ if

$$\lim_{\eta\downarrow 0} |\mu^{\eta}(f) - \mu(f)| = 0 \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

"bounded continuous functions supported on $\mathbb{S}\setminus\mathbb{C}$"

# $\mathbb{M}$-Convergence

$\varepsilon$-fattening of $\mathbb{C}$

$$\mathbb{M}(\mathbb{S}\setminus\mathbb{C}) = \{\nu(\cdot) : \nu(\mathbb{S}\setminus\mathbb{C}^{\varepsilon}) < \infty, \forall \varepsilon > 0; \quad \nu(\cdot) \text{ Borel measure on } \mathbb{S}\setminus\mathbb{C}\}$$

**Definition ($\mathbb{M}$-convergence; Lindskog, Resnick, Roy., 2014)**

Let $\mu^{\eta}, \mu \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$. We say that $\mu^{\eta}$ converges to $\mu$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ as $\eta \to 0$ if

$$\lim_{\eta\downarrow 0} |\mu^{\eta}(f) - \mu(f)| = 0 \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

"bounded continuous functions supported on $\mathbb{S}\setminus\mathbb{C}$"

**Definition (Uniform $\mathbb{M}$-convergence; Wang, R., 2024+)**

Let $\Theta$ be a set of indices. Let $\mu_{\theta}^{\eta}, \mu_{\theta} \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$ and $\theta \in \Theta$. We say that $\mu_{\theta}^{\eta}$ converges to $\mu_{\theta}$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ uniformly in $\theta$ on $\Theta$ as $\eta \to 0$ if

$$\lim_{\eta\downarrow 0} \sup_{\theta\in\Theta} |\mu_{\theta}^{\eta}(f) - \mu_{\theta}(f)| = 0 \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

# Portmanteau Theorem for Uniform $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-Convergence

**Theorem (Wang, R., 2024+)**

*Let $\mu_\theta^\eta, \mu_\theta \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$ and $\theta \in \Theta$. Suppose that for any sequence $(\theta_n)_{n\geq 1}$, there exist a sub-sequence $(\theta_{n_k})_{k\geq 1}$ and $\theta^* \in \Theta$ s.t.*

$$\lim_{k\to\infty} \mu_{\theta_{n_k}}(f) = \mu_{\theta^*}(f) \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

# Portmanteau Theorem for Uniform $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-Convergence

**Theorem (Wang, R., 2024+)**

*Let $\mu_\theta^\eta, \mu_\theta \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$ and $\theta \in \Theta$. Suppose that for any sequence $(\theta_n)_{n\geq 1}$, there exist a sub-sequence $(\theta_{n_k})_{k\geq 1}$ and $\theta^* \in \Theta$ s.t.*

$$\lim_{k\to\infty} \mu_{\theta_{n_k}}(f) = \mu_{\theta^*}(f) \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

*Then the following statements are equivalent:*

- $\mu_\theta^\eta \to \mu_\theta$ *in* $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ *uniformly in* $\theta$ *on* $\Theta$ *as* $\eta \downarrow 0$;
- *For all* $\varepsilon > 0$, $F, G$ *bounded away from* $\mathbb{C}$,

$$\liminf_{\eta\downarrow 0} \inf_{\theta\in\Theta} \left(\mu_\theta^\eta(G) - \mu_\theta(G_\varepsilon)\right) \geq 0$$

$$\limsup_{\eta\downarrow 0} \sup_{\theta\in\Theta} \left(\mu_\theta^\eta(F) - \mu_\theta(F^\varepsilon)\right) \leq 0$$

# Portmanteau Theorem for Uniform $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-Convergence

**Theorem (Wang, R., 2024+)**

*Let $\mu_\theta^\eta, \mu_\theta \in \mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ for each $\eta > 0$ and $\theta \in \Theta$. Suppose that for any sequence $(\theta_n)_{n\geq 1}$, there exist a sub-sequence $(\theta_{n_k})_{k\geq 1}$ and $\theta^* \in \Theta$ s.t.*

$$\lim_{k\to\infty} \mu_{\theta_{n_k}}(f) = \mu_{\theta^*}(f) \qquad \forall f \in \mathcal{C}(\mathbb{S}\setminus\mathbb{C}).$$

*Then the following statements are equivalent:*

- *$\mu_\theta^\eta \to \mu_\theta$ in $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$ uniformly in $\theta$ on $\Theta$ as $\eta \downarrow 0$;*
- *For all $\varepsilon > 0$, $F, G$ bounded away from $\mathbb{C}$,*

$$\liminf_{\eta\downarrow 0} \inf_{\theta\in\Theta} \left(\mu_\theta^\eta(G) - \mu_\theta(G_\varepsilon)\right) \geq 0$$

$$\limsup_{\eta\downarrow 0} \sup_{\theta\in\Theta} \left(\mu_\theta^\eta(F) - \mu_\theta(F^\varepsilon)\right) \leq 0$$

*Furthermore, they both imply*

- *For all open $G$ and closed $F$ that are bounded away from $\mathbb{C}$,*

$$\inf_{\theta\in\Theta} \mu_\theta(G) \leq \liminf_{\eta\downarrow 0} \inf_{\theta\in\Theta} \mu_\theta^\eta(G)$$

$$\limsup_{\eta\downarrow 0} \sup_{\theta\in\Theta} \mu_\theta^\eta(F) \leq \sup_{\theta\in\Theta} \mu_\theta(F).$$

# Asymptotic Atom $\{A(\varepsilon) \subseteq \mathbb{S} : \varepsilon > 0\}$ of Markov Chain $\{V_j^\eta(x) : j \geq 0\}$

For measurable $B \subseteq \mathbb{S}$, there exist $\delta_B : (0,\infty) \times (0,\infty) \to (0,\infty)$, $\varepsilon_B > 0$, and $T_B > 0$ s.t.

$$
\begin{aligned}
C(B^\circ) - \delta_B(\varepsilon, T) &\leq \liminf_{\eta \downarrow 0} \frac{\inf_{x \in A(\varepsilon)} \mathbf{P}\big(\tau^\eta_{I(\varepsilon)^\complement}(x) \leq T/\eta;\ V^\eta_{\tau_\varepsilon}(x) \in B\big)}{\gamma(\eta)T/\eta} \\
&\leq \limsup_{\eta \downarrow 0} \frac{\sup_{x \in A(\varepsilon)} \mathbf{P}\big(\tau^\eta_{I(\varepsilon)^\complement}(x) \leq T/\eta;\ V^\eta_{\tau_\varepsilon}(x) \in B\big)}{\gamma(\eta)T/\eta} \leq C(B^-) + \delta_B(\varepsilon, T) \\
&\limsup_{\eta \downarrow 0} \frac{\sup_{x \in I(\varepsilon)} \mathbf{P}\big(\tau^\eta_{(I(\varepsilon)\setminus A(\varepsilon))^\complement}(x) > T/\eta\big)}{\gamma(\eta)T/\eta} = 0 \\
&\liminf_{\eta \downarrow 0} \inf_{x \in I(\varepsilon)} \mathbf{P}\big(\tau^\eta_{A(\varepsilon)}(x) \leq T/\eta\big) = 1 \quad (\{I(\varepsilon) \subseteq I : \varepsilon > 0\}\text{: covering of } I)
\end{aligned}
$$

for any $\varepsilon \leq \varepsilon_B$ and $T \geq T_B$, where $\gamma(\eta)/\eta \to 0$ as $\eta \downarrow 0$ and $\delta_B$'s are such that

$$\lim_{\varepsilon \to 0} \lim_{T \to \infty} \delta_B(\varepsilon, T) = 0.$$

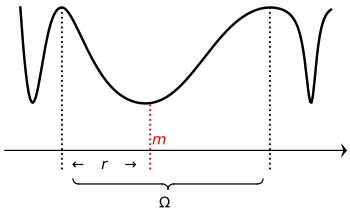
# Exit Time and Location under the Presence of Asymptotic Atom

**Theorem (Wang, R., 2024+)**

*If Markov chain* $\{V_j^\eta(x) : j \geq 0\}$ *possesses an asymptotic atom* $\{A(\varepsilon) \subseteq \mathbb{S} : \varepsilon > 0\}$, *then*

$$\begin{aligned} C(B^\circ) \cdot e^{-t} &\leq \liminf_{\eta \downarrow 0} \inf_{x \in I(\varepsilon)} \mathbf{P}\big(\gamma(\eta)\tau_{I^\complement}^\eta(x) > t, V_\tau^\eta(x) \in B\big) \\ &\leq \limsup_{\eta \downarrow 0} \sup_{x \in I(\varepsilon)} \mathbf{P}\big(\gamma(\eta)\tau_{I^\complement}^\eta(x) > t, V_\tau^\eta(x) \in B\big) \leq C(B^-) \cdot e^{-t}. \end{aligned}$$

# First Exit Time Analysis for SGD




gradient noise

$l = \lceil r/c \rceil$: "width" of the attraction fields,

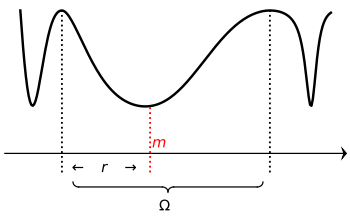
$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, Oh, R., 2022)**

*Let* $\sigma(\eta) = \min\{j \geq 0 : W_j^\eta \notin \Omega\}$ *and* $\lambda(\eta) \sim \eta^{1+\alpha \cdot l}$

$$\sigma(n)\lambda(n) \Rightarrow \mathrm{Exp}(1)$$

# First Exit Time Analysis for SGD




gradient noise

$l = \lceil r/c \rceil$: "width" of the attraction fields,

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

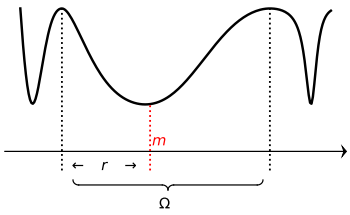
**Theorem (Wang, Oh, R., 2022)**

*Let* $\sigma(\eta) = \min\{j \geq 0 : W_j^\eta \notin \Omega\}$ *and* $\lambda(\eta) \sim \eta^{1+\alpha \cdot l}$

First Exit Time

$$\sigma(n)\lambda(n) \Rightarrow \mathrm{Exp}(1)$$

# First Exit Time Analysis for SGD




gradient noise

$l = \lceil r/c \rceil$: "width" of the attraction fields,

$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

**Theorem (Wang, Oh, R., 2022)**

*Let* $\sigma(\eta) = \min\{j \geq 0 : W_j^\eta \notin \Omega\}$ *and* $\lambda(\eta) \sim \eta^{1+\alpha \cdot l}$

First Exit Time

$$\sigma(n)\lambda(n) \Rightarrow \mathrm{Exp}(1)$$

$\sim (1/\eta)^{1+\alpha \cdot l}$

# Eliminating Sharp Local Minima with Truncated Heavy-Tails

gradient noise

$l^*$: "width" of the widest attraction fields, $\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$

**Theorem (Wang, Oh, R., 2022)**

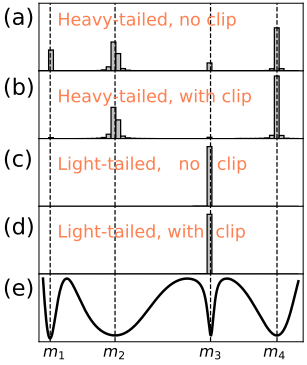
*Under certain structural conditions, for any $t > 0$ and $\beta > 1 + \alpha \cdot l^*$,*

$$\frac{1}{t/\eta^\beta} \int_0^{\lfloor t/\eta^\beta \rfloor} \mathbb{I}\{W^\eta_{\lfloor u \rfloor} \in \textit{sharp minima}\} du \xrightarrow{p} 0$$

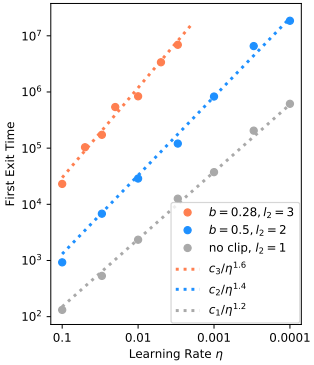
**In fact, $W^\eta_{\lfloor t/\eta^{1+\alpha \cdot l^*} \rfloor}$ converges to a Markov jump processes**

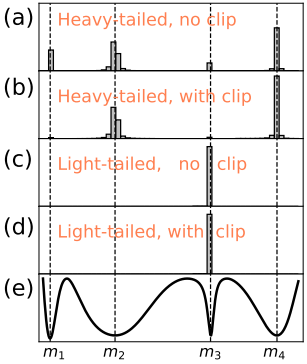
**whose state space consists of wide local minima only.**

# Eliminating Sharp Local Minima with Truncated Heavy-Tails

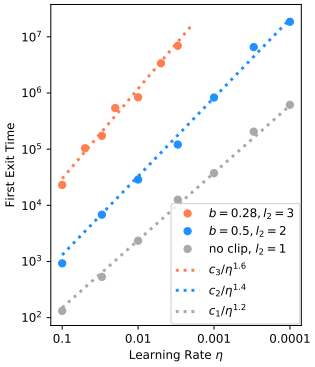


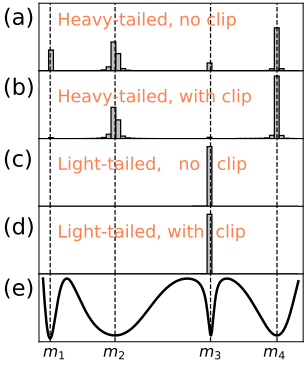
# Eliminating Sharp Local Minima with Truncated Heavy-Tails

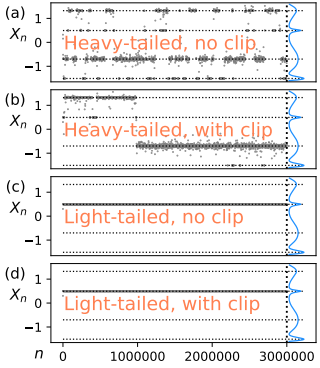


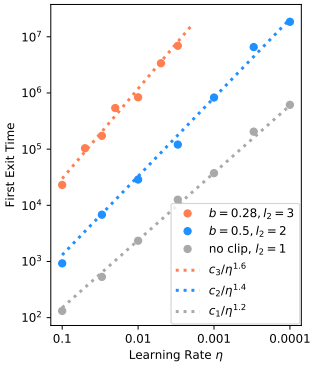
# Eliminating Sharp Local Minima with Truncated Heavy-Tails

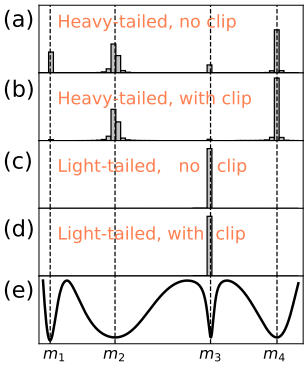


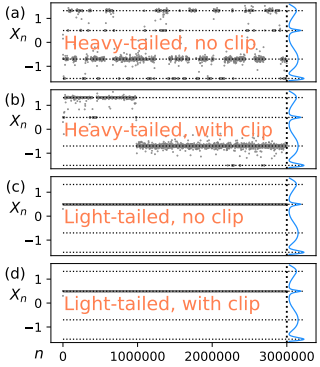



# Eliminating Sharp Local Minima with Truncated Heavy-Tails

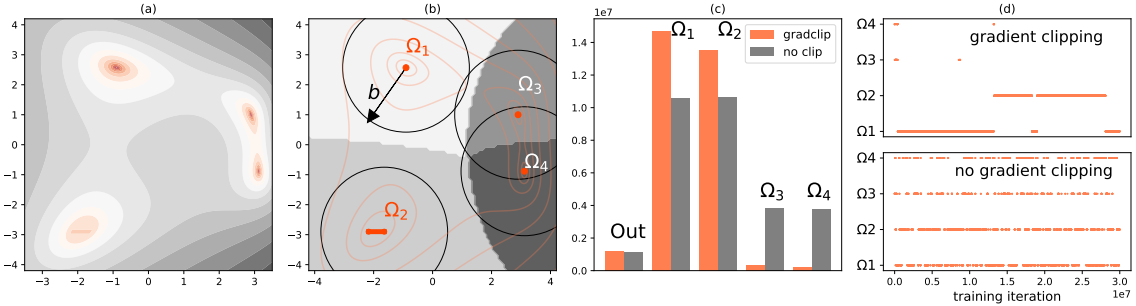


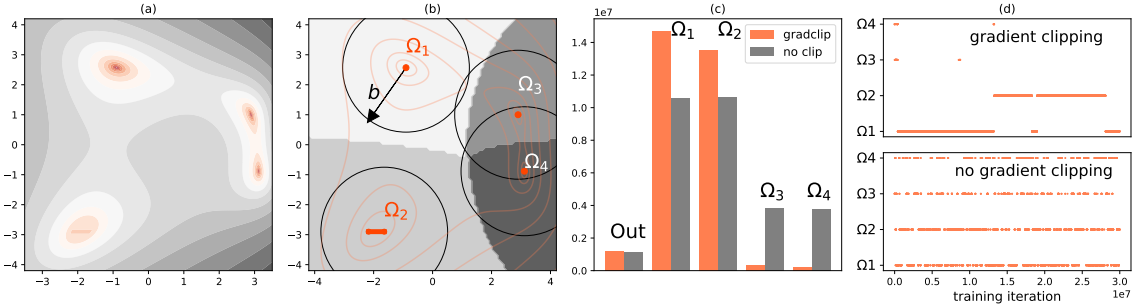




**Consistent with what theory predicts!**

# Same Phenomena in $\mathbb{R}^2$ with More General Geometry



# Same Phenomena in $\mathbb{R}^2$ with More General Geometry




**Again, consistent with what theory predicts!**

## New Training Strategy: Tail-INflation-Truncation

# Tail-Inflation-Truncation Scheme

$$\nabla\tilde{f} \quad = \nabla f_{\text{small batch}}$$

$$W_{k+1} = W_k - \varphi_b\big(\eta \cdot \nabla\tilde{f}(W_k)\big)$$

# Tail-Inflation-Truncation Scheme

Stochastic Gradient

$$\nabla\tilde{f} \quad = \nabla f_{\text{small batch}}$$

$$W_{k+1} = W_k - \varphi_b\big(\eta \cdot \nabla\tilde{f}(W_k)\big)$$

# Tail-Inflation-Truncation Scheme

Stochastic Gradient

$$\searrow$$

$$\nabla \tilde{f}_{\text{our}} = \nabla f_{\text{small batch}} + R \cdot (\nabla f_{\text{large batch}} - \nabla f_{\text{small batch}})$$

$$W_{k+1} = W_k - \varphi_b\big(\eta \cdot \nabla \tilde{f}(W_k)\big)$$

# Tail-Inflation-Truncation Scheme

Stochastic Gradient Pareto RV

$$\nabla \tilde{f}_{\text{our}} = \nabla f_{\text{small batch}} + R \cdot (\nabla f_{\text{large batch}} - \nabla f_{\text{small batch}})$$

$$W_{k+1} = W_k - \varphi_b\big(\eta \cdot \nabla \tilde{f}(W_k)\big)$$

# Tail-Inflation-Truncation Scheme

Heavy-Tailed Stochastic Gradient

Pareto RV

$$\nabla \tilde{f}_{\text{our}} = \nabla f_{\text{small batch}} + R \cdot (\nabla f_{\text{large batch}} - \nabla f_{\text{small batch}})$$

$$W_{k+1} = W_k - \varphi_b\big(\eta \cdot \nabla \tilde{f}(W_k)\big)$$

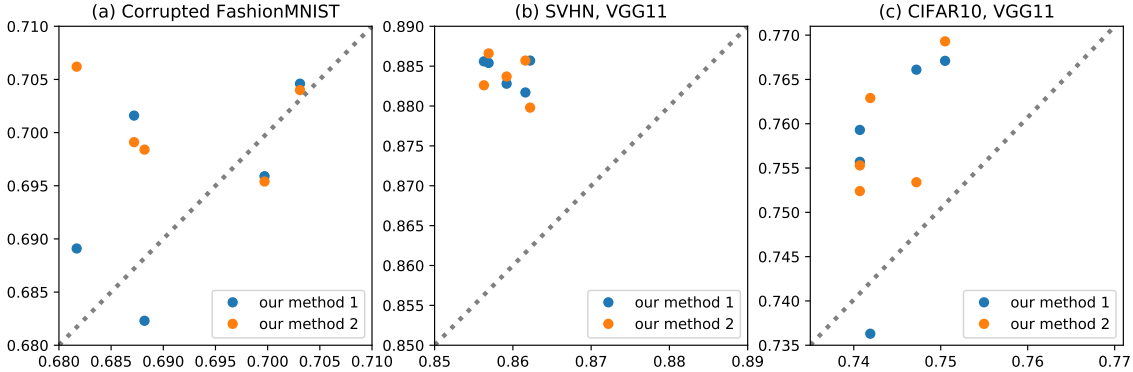
# Tail-Inflation-Truncation Scheme

Heavy-Tailed Stochastic Gradient ↘ Pareto RV ↙

$$\widetilde{\nabla f}_{\text{our}} = \nabla f_{\text{small batch}} + R \cdot (\nabla f_{\text{large batch}} - \nabla f_{\text{small batch}})$$

| Test accuracy | LB | SB | SB+Clip | SB+Noise | Our 1 | Our 2 |
|---|---|---|---|---|---|---|
| FashionMNIST, 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** |
| FashionMNIST, 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 |

# Improvement is Consistent



# Tail-Inflation-Truncation Scheme

| CIFAR10-VGG11 | SB + Clip | Our 1 | Our 2 |
|---|---|---|---|
| Test Accuracy | 89.54% | **90.67%** | 90.45% |
| Expected Sharpness | 0.167 | **0.085** | 0.096 |
| PAC-Bayes Sharpness | $1.31 \times 10^4$ | $\mathbf{9 \times 10^3}$ | $10^4$ |
| Maximal Sharpness | $1.66 \times 10^4$ | $1.29 \times 10^4$ | $\mathbf{1.22 \times 10^4}$ |
| 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 | $2.49 \times 10^4$ | $\mathbf{1.9 \times 10^4}$ | $1.98 \times 10^4$ |
| Maximal Sharpness | $2.75 \times 10^4$ | $\mathbf{2.12 \times 10^4}$ | $2.16 \times 10^4$ |

# Does This Actually Work with High-Volume Real-Life Data?

We will know soon:

- Moloco
  - top performing mobile ad tech company
  - 35B+ user impressions per month
- A project to improve Moloco's conversion-rate prediction accuracy

# Summary

# Summary

- Catastrophe Principle for various heavy-tailed stochastic processes

# Summary

- Catastrophe Principle for various heavy-tailed stochastic processes

- Solutions to open problems in queueing theory and rare-event simulation

# Summary

- Catastrophe Principle for various heavy-tailed stochastic processes
- Solutions to open problems in queueing theory and rare-event simulation
- Metastability analysis reveals the global dynamics of heavy-tailed dynamical systems

# Summary

- Catastrophe Principle for various heavy-tailed stochastic processes
- Solutions to open problems in queueing theory and rare-event simulation
- Metastability analysis reveals the global dynamics of heavy-tailed dynamical systems
- Elimination of sharp local minima from SGD with truncated heavy-tailed gradient noise