# Capitalizing Catastrophe Principle in Training DNN and Other Related Topics

Chang-Han Rhee

Northwestern University

INI Satellite Programme on Heavy Tails in Machine Learning, The Alan Turing Institute

April 5, 2024

Based on the joint works with

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

# Outline

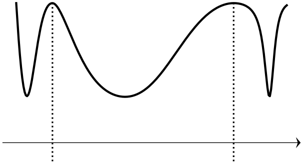
- Lecture 1: Heavy-Tailed Large Deviations Approach to Global Dynamics of SGD
  - Motivation, Phenomena, and Insights
- Lecture 2: Conspiracy vs Catastrophe Principle via Sample Path Large Deviations
  - Sample-Path Large Deviations: Light-Tailed vs Heavy-Tailed Frameworks
  - Fundamental Limitations of the Classical LDP Framework in Heavy-Tailed Systems
  - M-convergence and Related Machinery
  - Topological Considerations
- Lecture 3: Analysis of Local Stability and Global Dynamics of Heavy-Tailed SGD
  - Uniform M-convergence
  - Exit Time Analysis via Asymptotic Atom
  - Scaling Limit of Sample Paths
- Lecture 4: Capitalizing Catastrophe Principle in Training DNN & Other Related Topics
  - Tail Inflation Truncation Strategy
  - Large Deviations in M1' topology
  - Local Instability and its Connection to Edge of Stability

# Previously in this Lecture Series

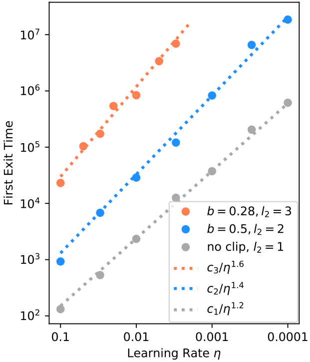
# Generalization and Sharp/Flat Local Minima

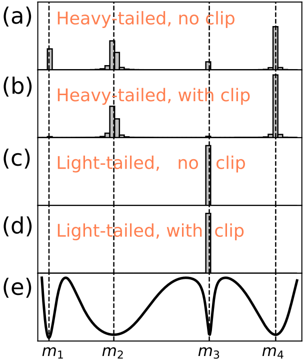
Training Set

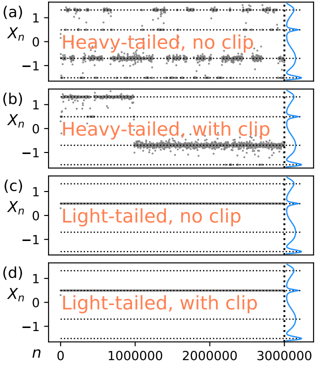
Test Set image source



# Catastrophe vs Conspiracy in SGD



# 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

**Systemwide rare events**

**arise because**

**EVERYTHING goes wrong.**

**(Conspiracy Principle)**

**Heavy-Tailed Distributions**

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

**Structural difference in the way systemwide rare events arise.**

# Rare Events depend on "Tail Behaviors"

**Light-Tailed Distributions**

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

**Systemwide rare events**

**arise because**

**EVERYTHING goes wrong.**

**(Conspiracy Principle)**

**Heavy-Tailed Distributions**

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

**Systemwide rare events**

**arise because of**

**A FEW Catastrophes.**

**(Catastrophe Principle)**

**Structural difference in the way systemwide rare events arise.**

# 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

Poisson Arrival

Initial Capital

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

Premium

i.i.d. Claim Size

# 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

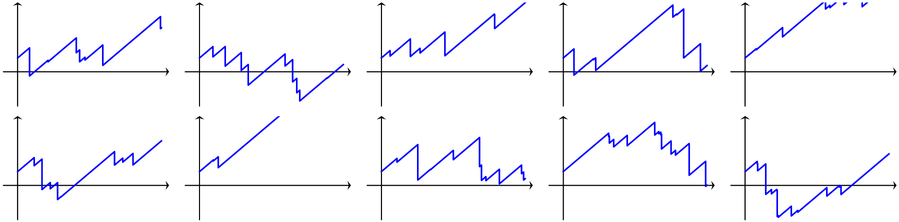
# Typical Scenario

# Typical Scenario

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

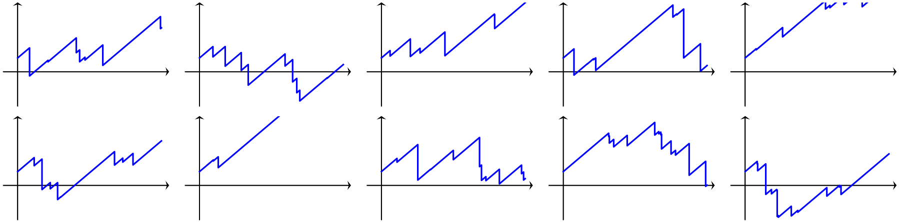
# Typical Scenario

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



# Typical Scenario

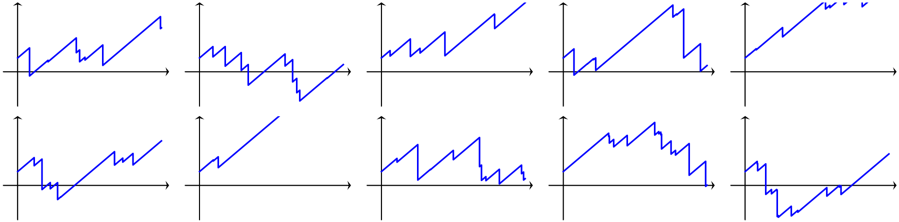
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**



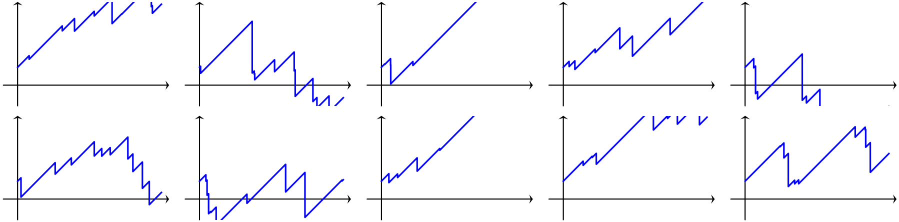
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

# Typical Scenario

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

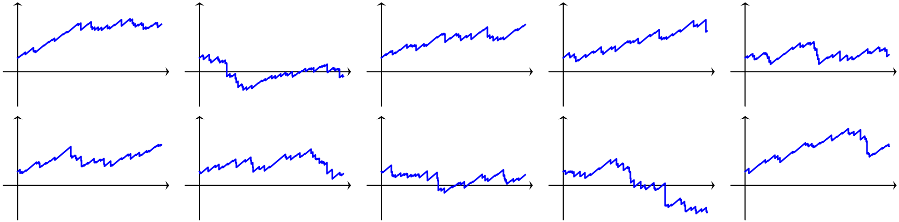


Sample paths of $\bar{Y}_n$:  $n$=10 & claim sizes are **heavy-tailed**

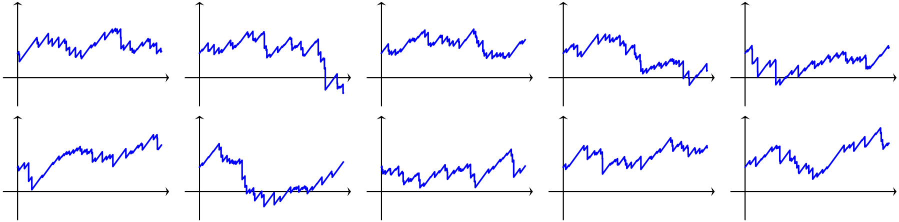


# Typical Scenario

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

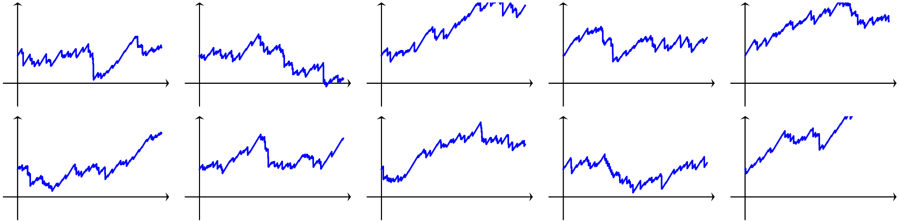


Sample paths of $\bar{Y}_n$: n=50 & claim sizes are **heavy-tailed**

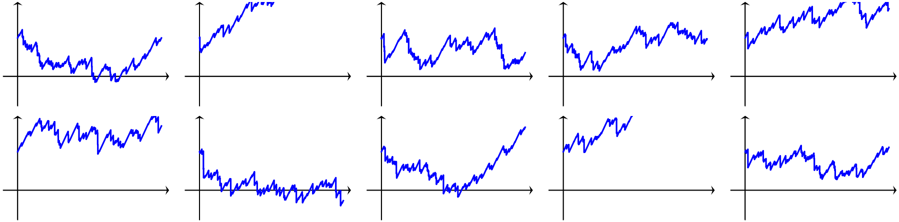


# Typical Scenario

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

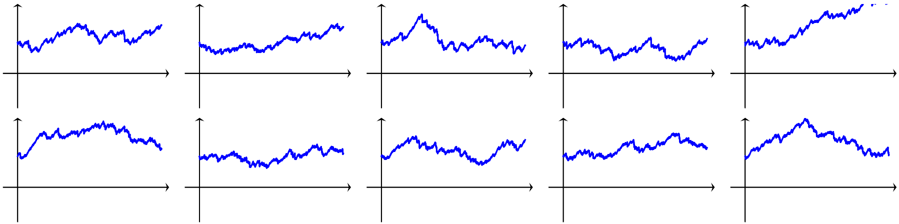


Sample paths of $\bar{Y}_n$: n=100 & claim sizes are **heavy-tailed**

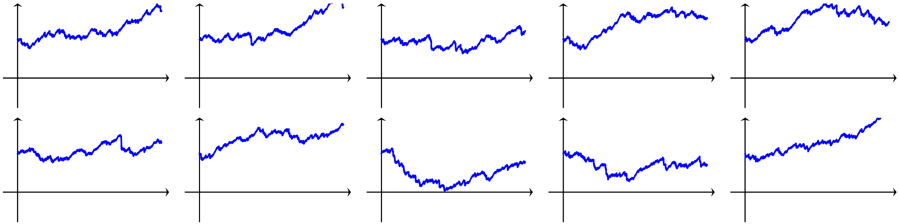


# Typical Scenario

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

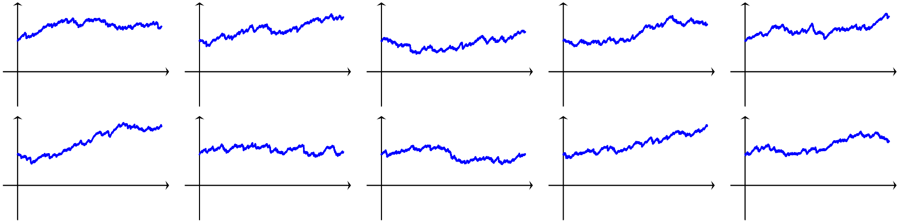


Sample paths of $\bar{Y}_n$: n=500 & claim sizes are **heavy-tailed**

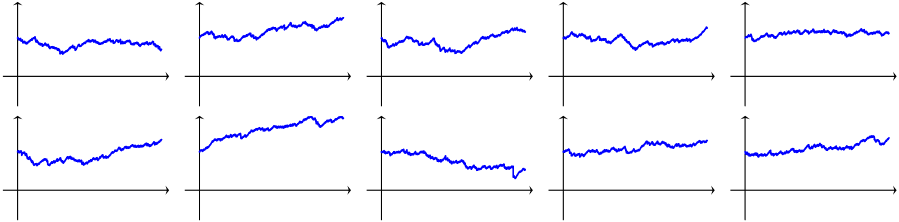


# Typical Scenario

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

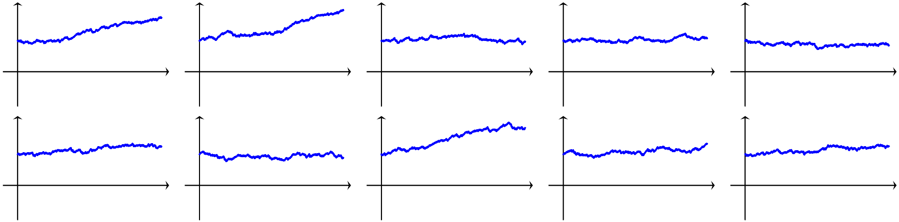


Sample paths of $\bar{Y}_n$: n=1000 & claim sizes are **heavy-tailed**

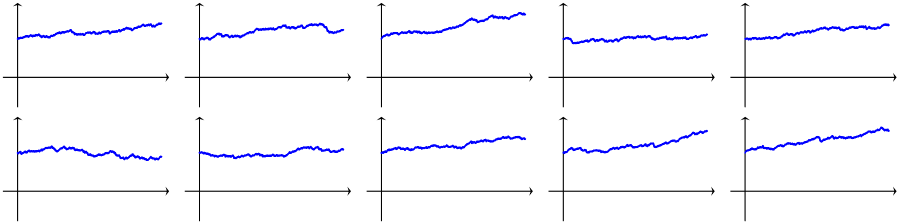


# Typical Scenario

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

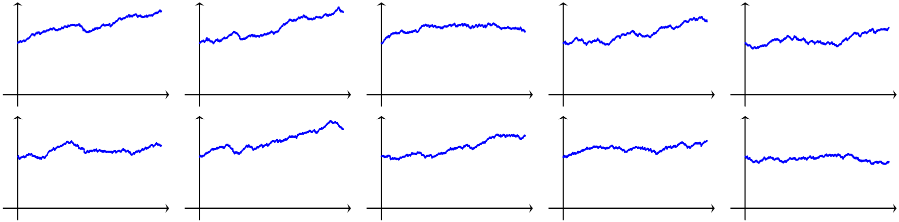


Sample paths of $\bar{Y}_n$: n=2500 & claim sizes are **heavy-tailed**

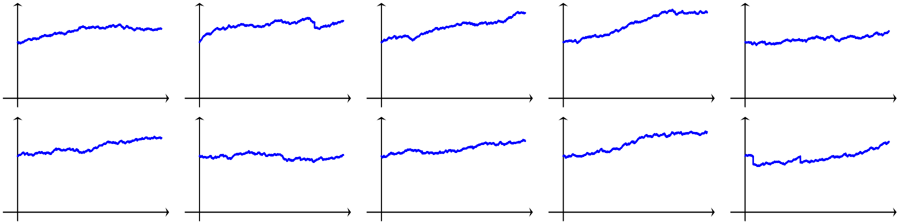


# Typical Scenario

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**

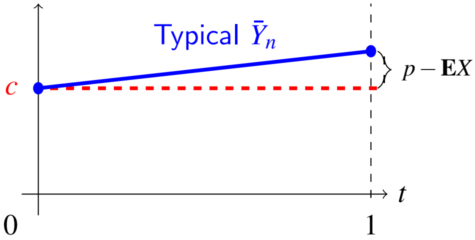


# Typical Scenario

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

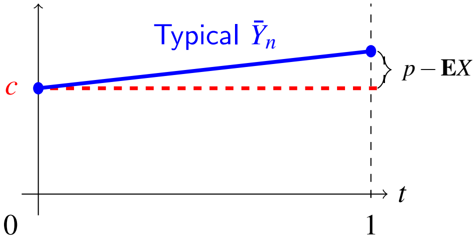
# Typical Scenario

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



# Typical Scenario

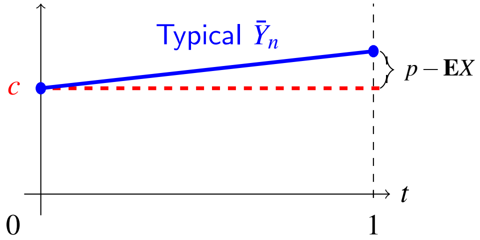
That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.




**Typically, your business will flourish**

# Typical Scenario

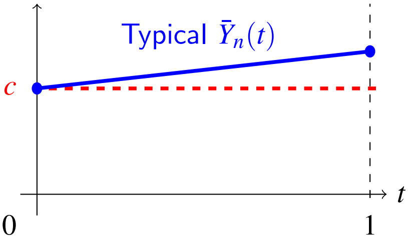
That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large $n$.




**Typically, your business will flourish**

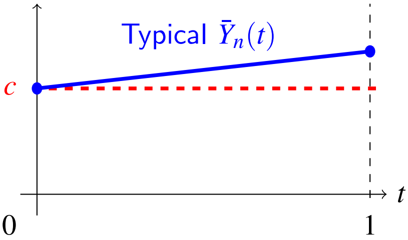
**regardless of the tail distributions.**

# A Rare Event: Bankruptcy



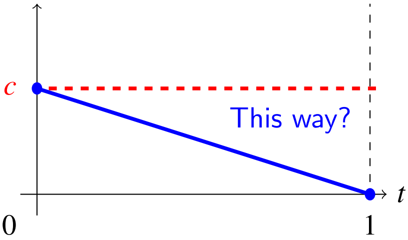
# A Rare Event: Bankruptcy

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



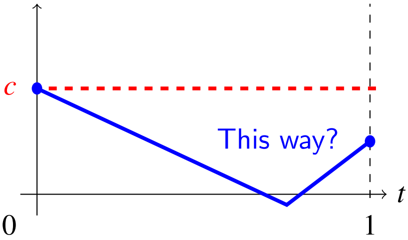
# A Rare Event: Bankruptcy

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



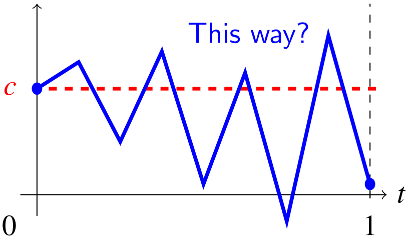
# A Rare Event: Bankruptcy

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



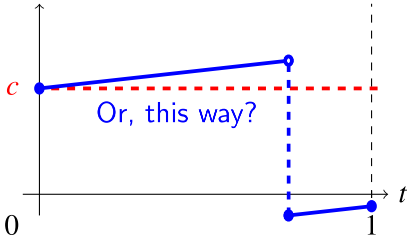
# A Rare Event: Bankruptcy

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



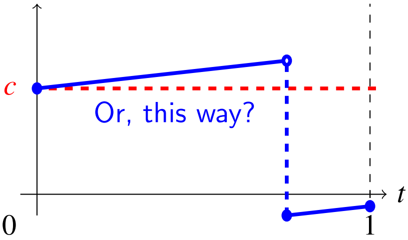
# A Rare Event: Bankruptcy

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



# A Rare Event: Bankruptcy

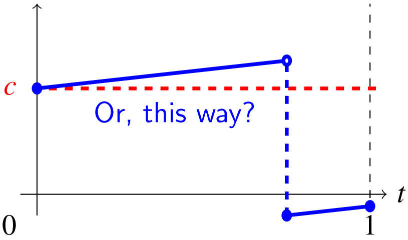
Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)




**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)

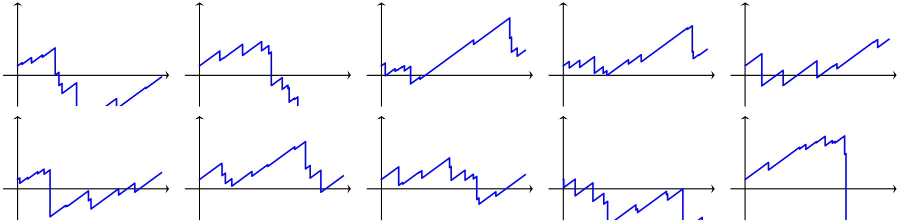



**Are we going to see clear patterns?**

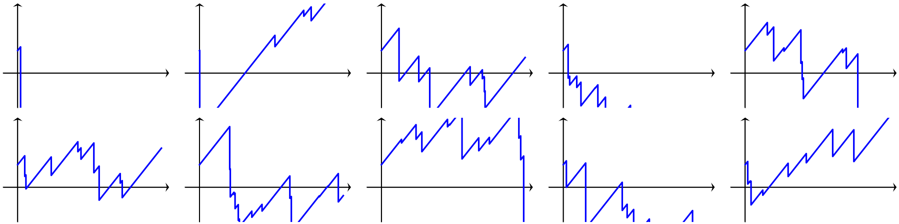
**Do they depend on the tail distributions?**

# A Rare Event: Bankruptcy

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



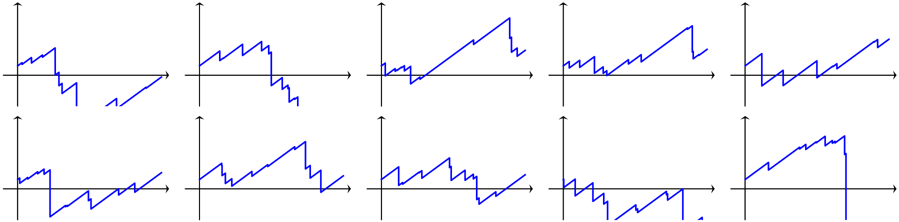
Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **heavy-tailed** claims:



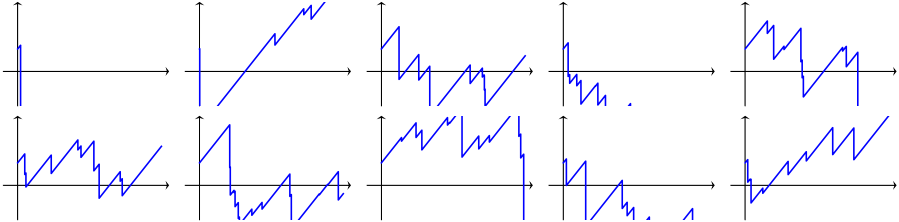
# A Rare Event: Bankruptcy

Bankruptcy

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



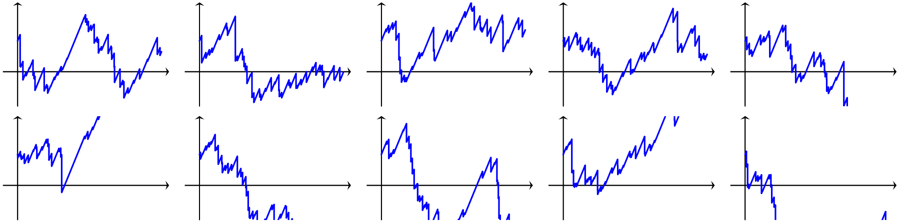
Sample paths of $\bar{Y}_{10}$ conditional on $B$ for **heavy-tailed** claims:



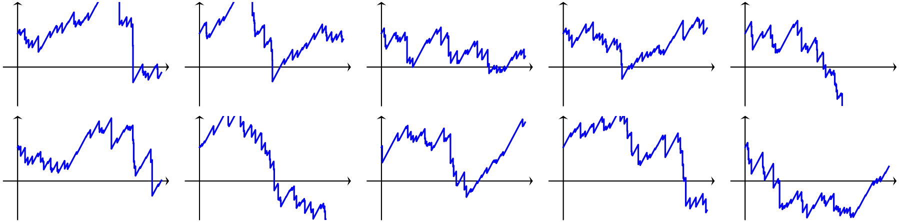
# A Rare Event: Bankruptcy

Bankruptcy

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



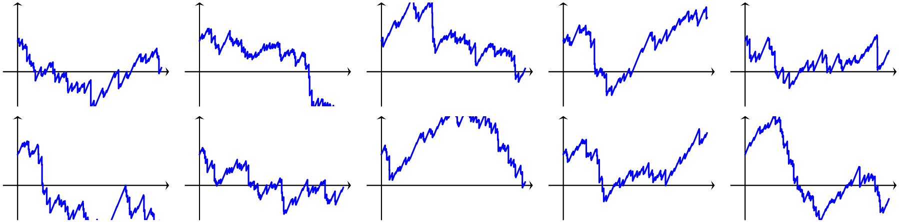
Sample paths of $\bar{Y}_{50}$ conditional on $B$ for **heavy-tailed** claims:



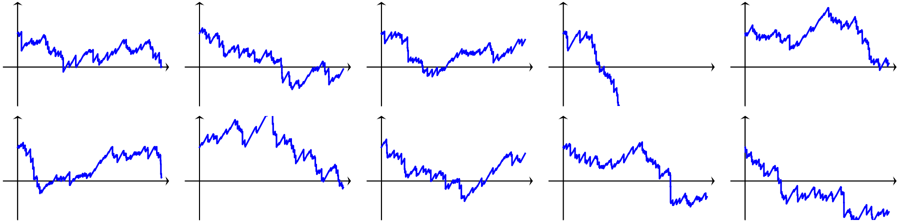
# A Rare Event: Bankruptcy

Bankruptcy

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



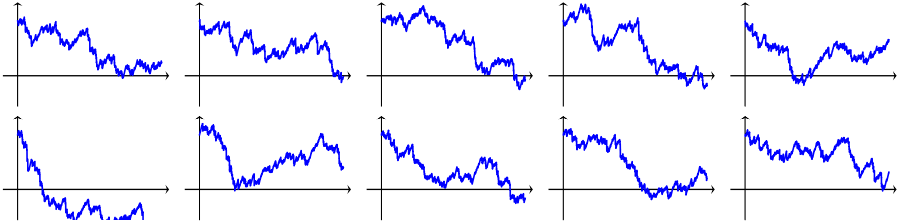
Sample paths of $\bar{Y}_{100}$ conditional on $B$ for **heavy-tailed** claims:



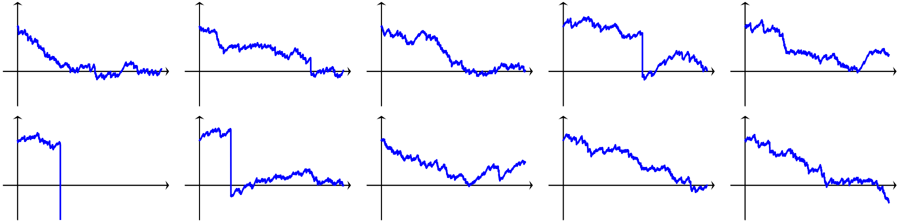
# A Rare Event: Bankruptcy

Bankruptcy

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



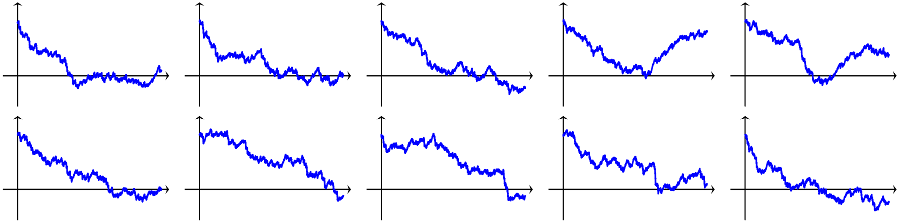
Sample paths of $\bar{Y}_{500}$ conditional on $B$ for **heavy-tailed** claims:



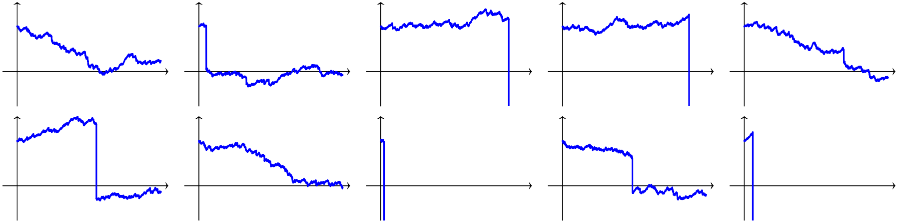
# A Rare Event: Bankruptcy

Bankruptcy

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



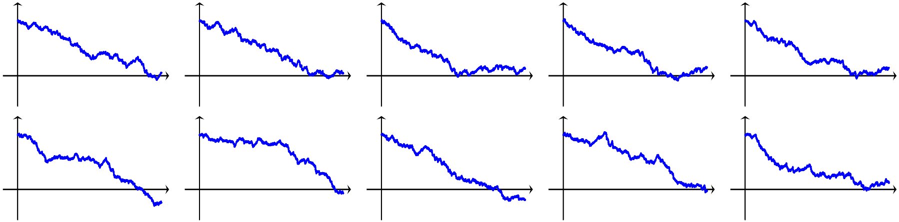
Sample paths of $\bar{Y}_{1000}$ conditional on $B$ for **heavy-tailed** claims:



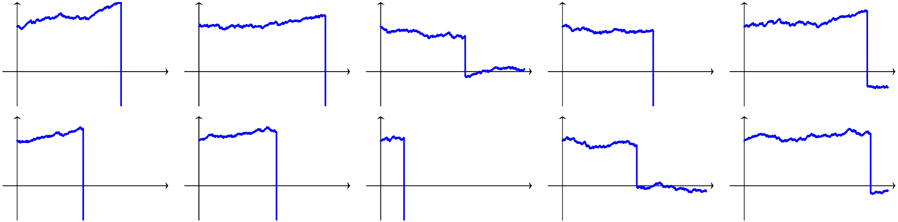
# A Rare Event: Bankruptcy

Bankruptcy

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



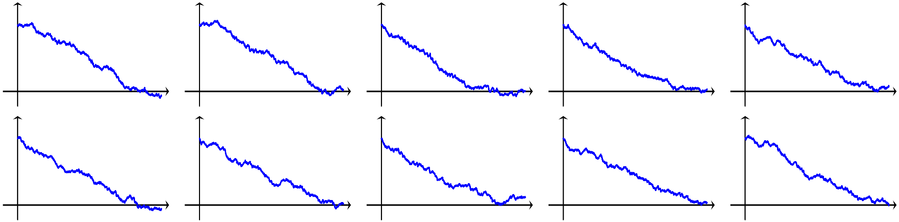
Sample paths of $\bar{Y}_{2500}$ conditional on $B$ for **heavy-tailed** claims:



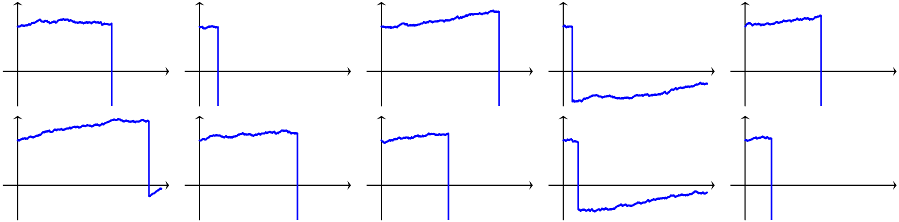
# A Rare Event: Bankruptcy

Bankruptcy

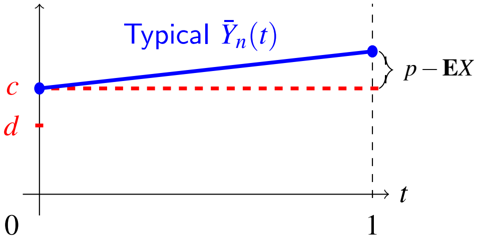
Sample paths of $\bar{Y}_{5000}$ conditional on $B$ for **light-tailed** claims:



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

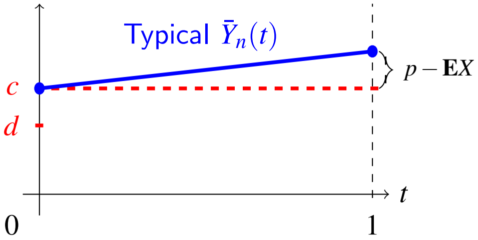


# Bankruptcy Despite of Reinsurance



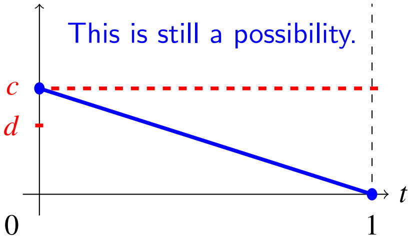
# Bankruptcy Despite of Reinsurance

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



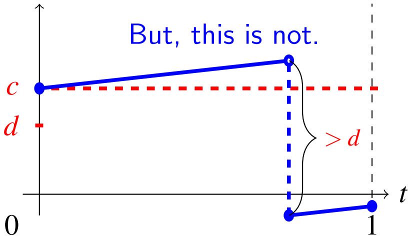
# Bankruptcy Despite of Reinsurance

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



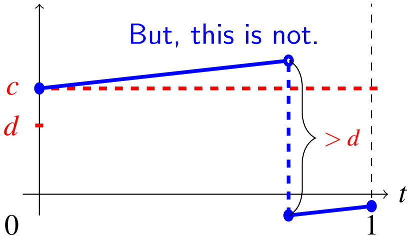
# Bankruptcy Despite of Reinsurance

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



# Bankruptcy Despite of Reinsurance

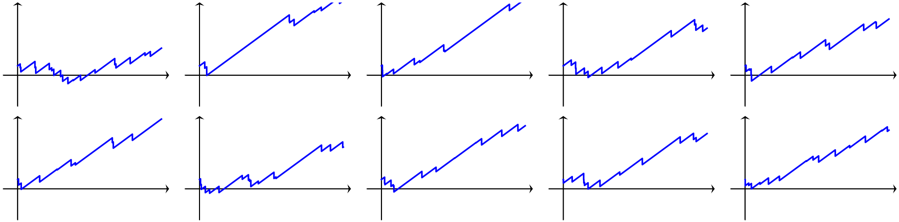
Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



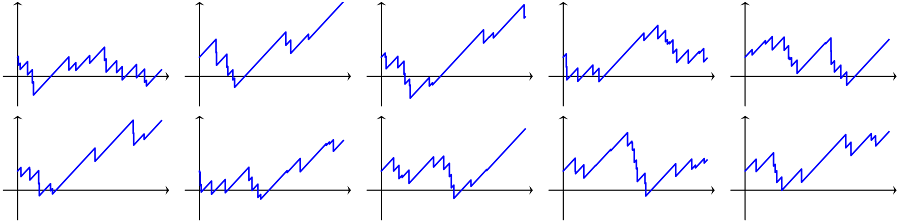

**How does the pattern change in this case?**

# Bankruptcy Despite of Reinsurance

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

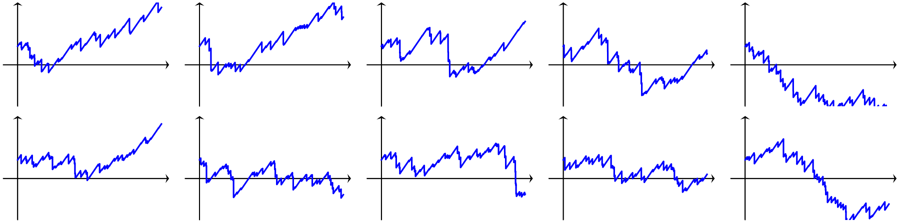


Sample paths of $\bar{Y}_{10}$ conditional on $R$ for **heavy-tailed** claims:

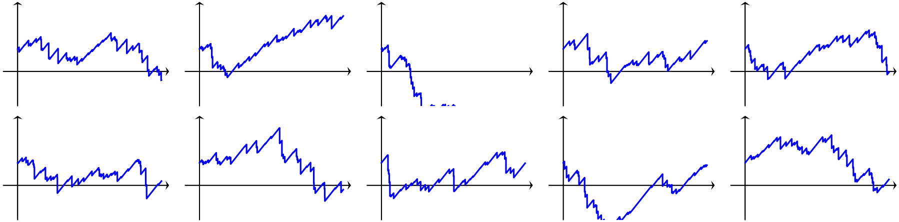


# Bankruptcy Despite of Reinsurance

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

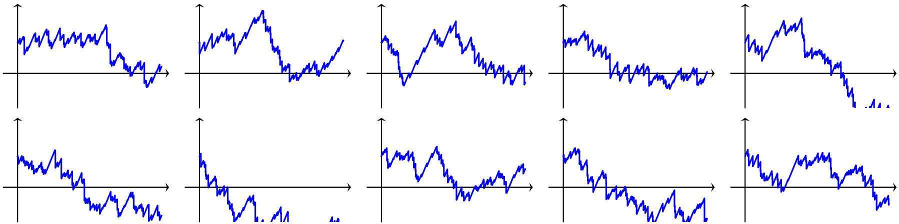


Sample paths of $\bar{Y}_{50}$ conditional on $R$ for **heavy-tailed** claims:

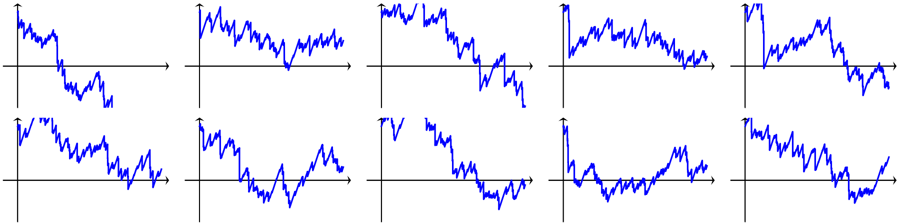


# Bankruptcy Despite of Reinsurance

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

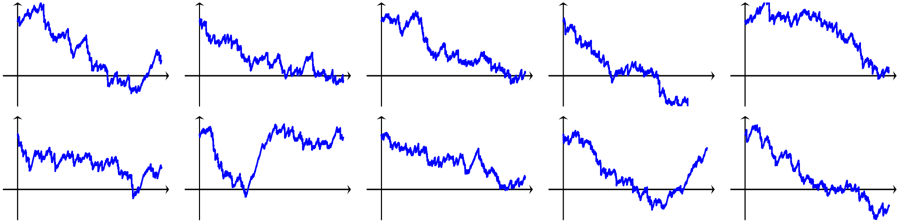


Sample paths of $\bar{Y}_{100}$ conditional on $R$ for **heavy-tailed** claims:

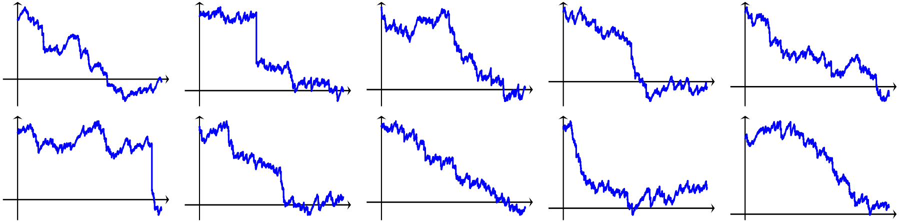


# Bankruptcy Despite of Reinsurance

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

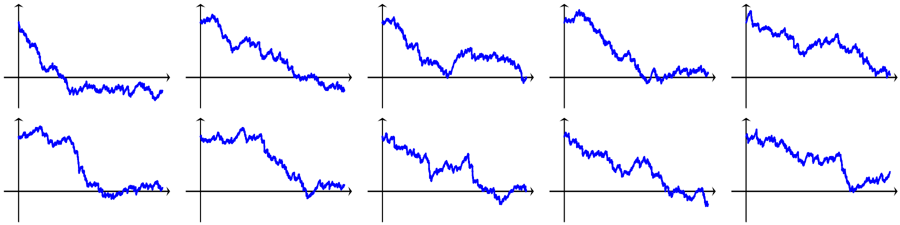


Sample paths of $\bar{Y}_{500}$ conditional on $R$ for **heavy-tailed** claims:

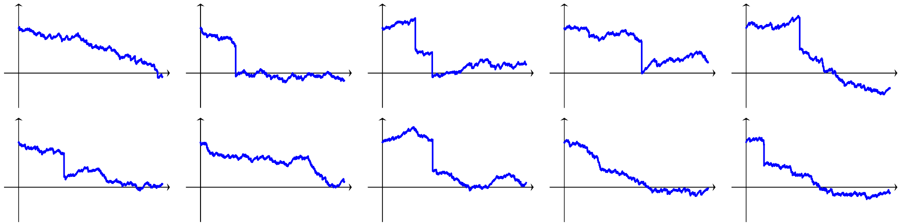


# Bankruptcy Despite of Reinsurance

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

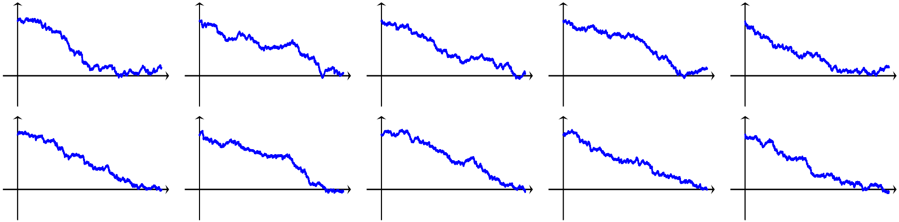


Sample paths of $\bar{Y}_{1000}$ conditional on $R$ for **heavy-tailed** claims:

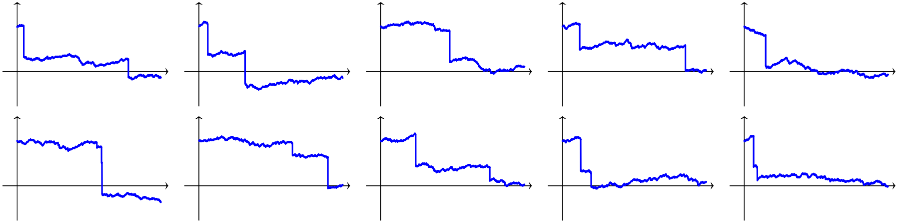


# Bankruptcy Despite of Reinsurance

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

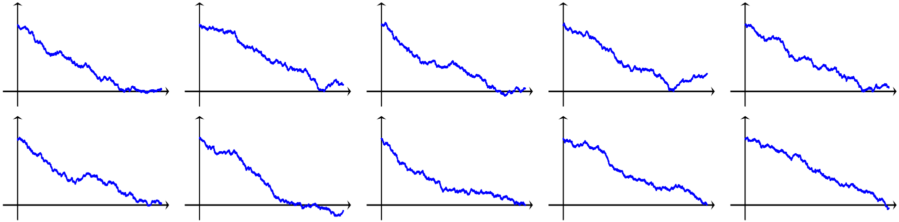


Sample paths of $\bar{Y}_{2500}$ conditional on $R$ for **heavy-tailed** claims:

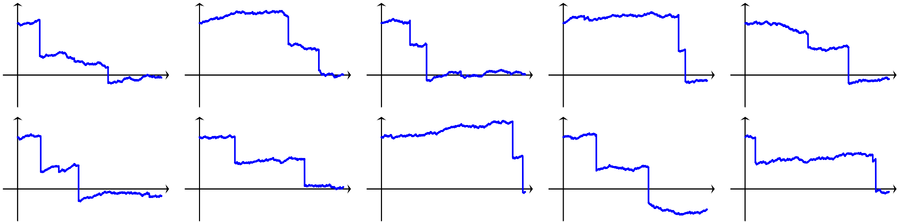


# Bankruptcy Despite of Reinsurance

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

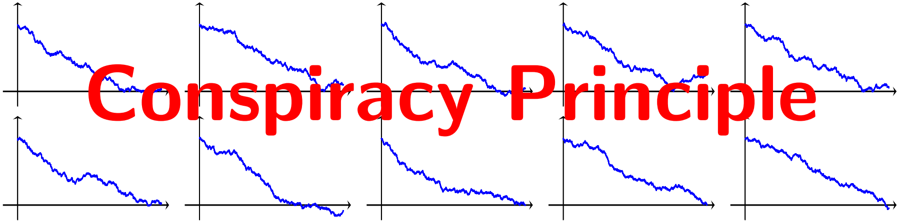


Sample paths of $\bar{Y}_{5000}$ conditional on $R$ for **heavy-tailed** claims:



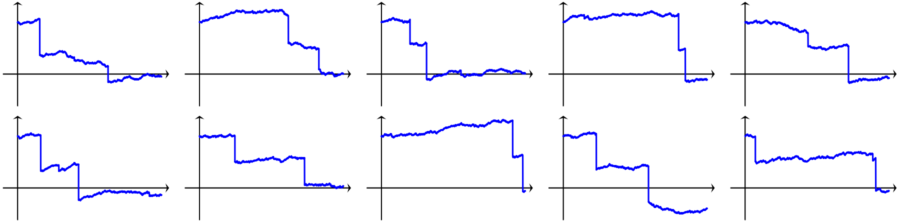
# Bankruptcy Despite of Reinsurance

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



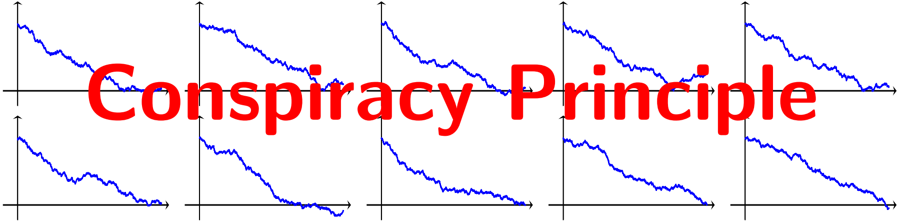
**Conspiracy Principle**

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

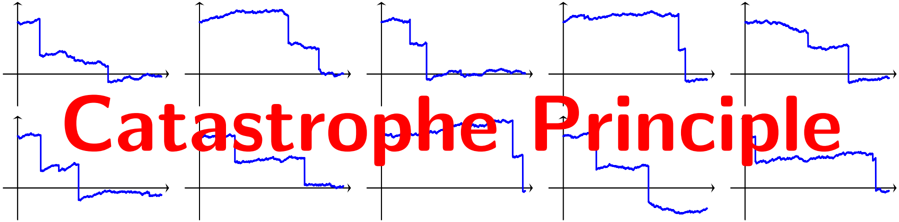


# Bankruptcy Despite of Reinsurance

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:



# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i,$$

$X_i$: centered iid r.v. with $\mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}$, $\alpha > 0$

# Heavy-Tailed Large Deviations

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i,$$

$X_i$: centered iid r.v. with $\mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}$, $\alpha > 0$

**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

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i,$$

$X_i$: 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

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i,$$

$X_i$: centered iid r.v. with $\mathbf{P}(X_i \geq x) = x^{-(\alpha+1)}$, $\alpha > 0$

**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* $A$

# Rigorous Characterization of the Catastrophe Principle

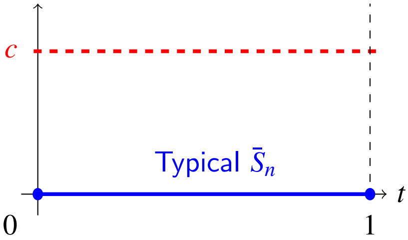
Under certain regularity conditions on $A$,

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

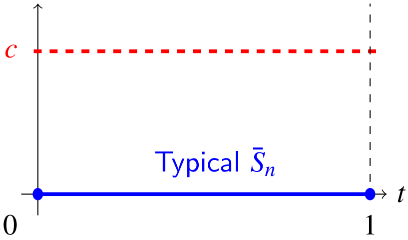
$\bar{S}_{|A}$: a (random) piecewise-constant function with $\mathcal{J}(A)$ jumps.

# Conspiracy vs Catastrophe

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



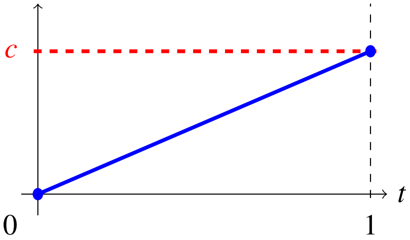

**Light-Tailed Claim Size**



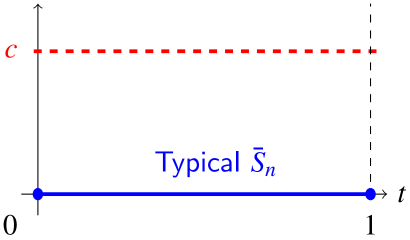

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

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

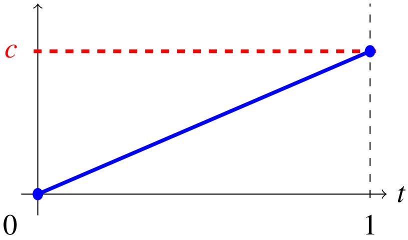



**Light-Tailed Claim Size**

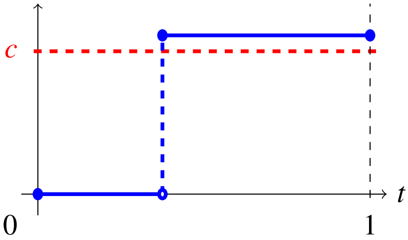



**Heavy-Tailed Claim Size**

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



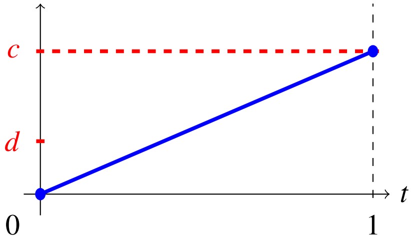

**Light-Tailed Claim Size**



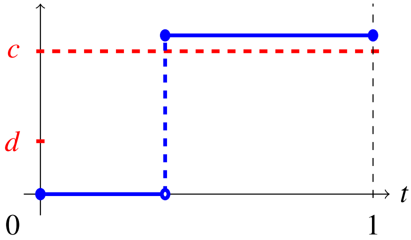

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

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



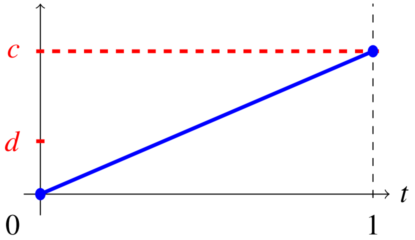

**Light-Tailed Claim Size**



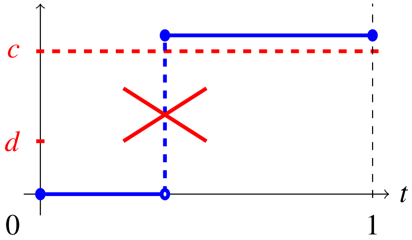

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

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



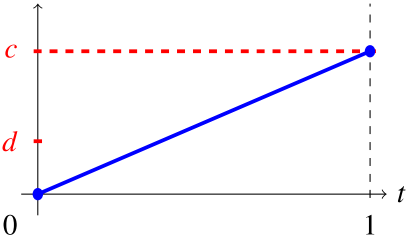

**Light-Tailed Claim Size**



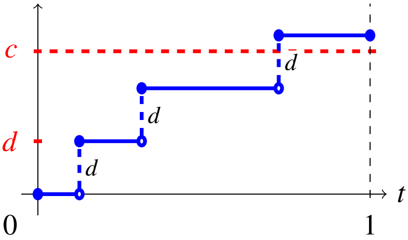

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

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



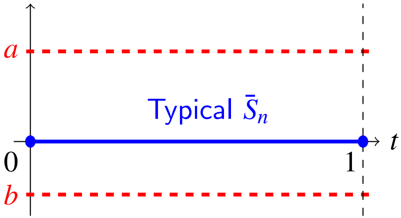

**Light-Tailed Claim Size**



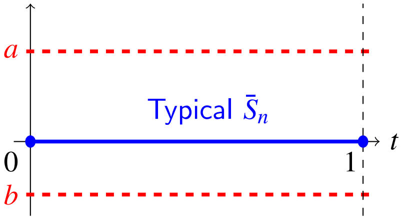

**Heavy-Tailed Claim Size**

# Conspiracy vs Catastrophe

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



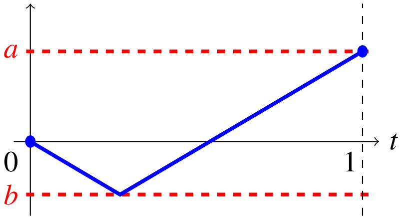

**Light-Tailed Increments**



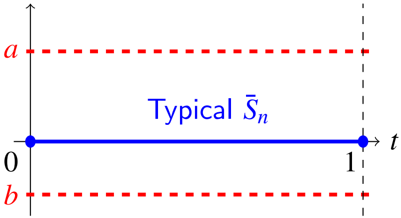

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

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



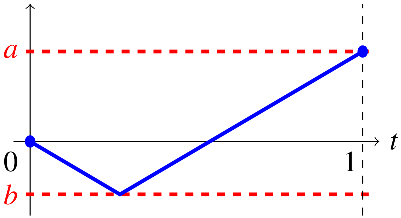

**Light-Tailed Increments**



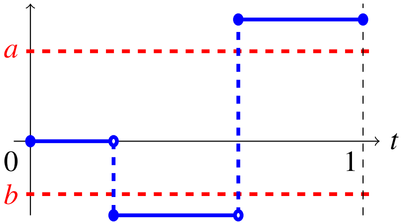

**Heavy-Tailed Increments**

## Conspiracy vs Catastrophe

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



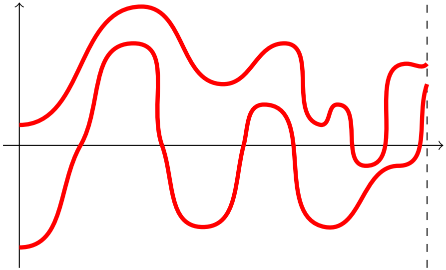

**Light-Tailed Increments**



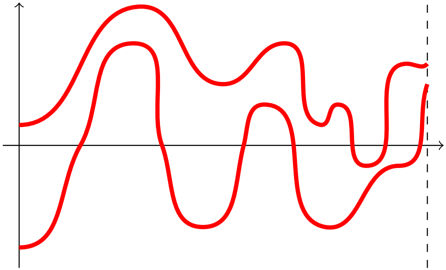

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

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



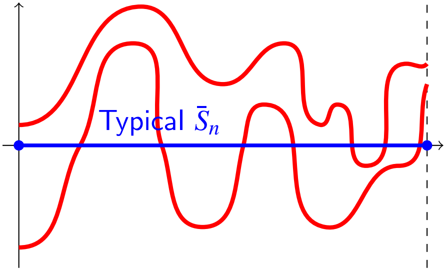
**Light-Tailed Increments**



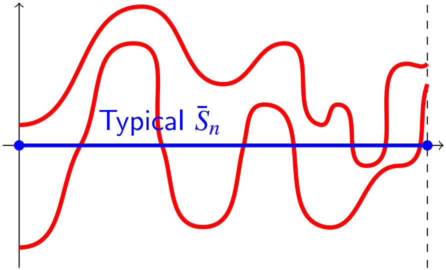
**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

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



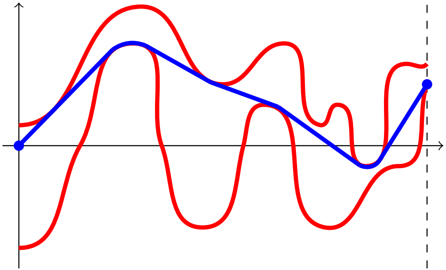

**Light-Tailed Increments**



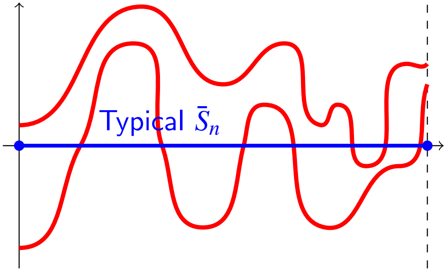

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

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



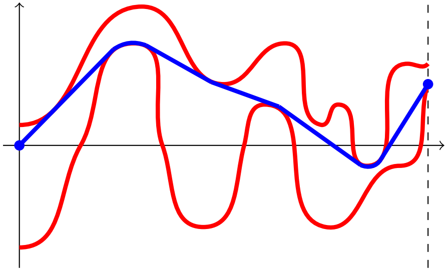
**Light-Tailed Increments**



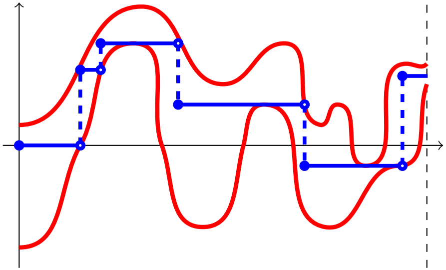

**Heavy-Tailed Increments**

# Conspiracy vs Catastrophe

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

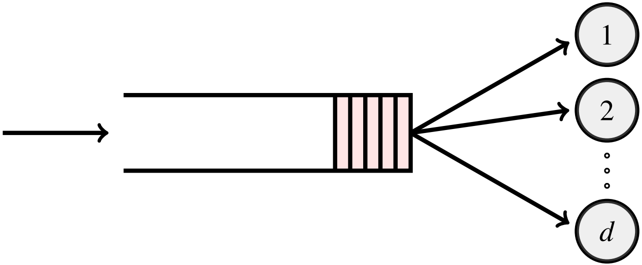


**Light-Tailed Increments**

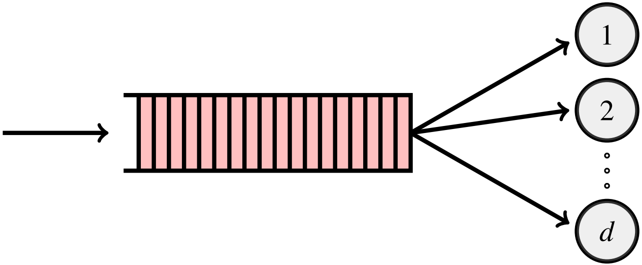


**Heavy-Tailed Increments**

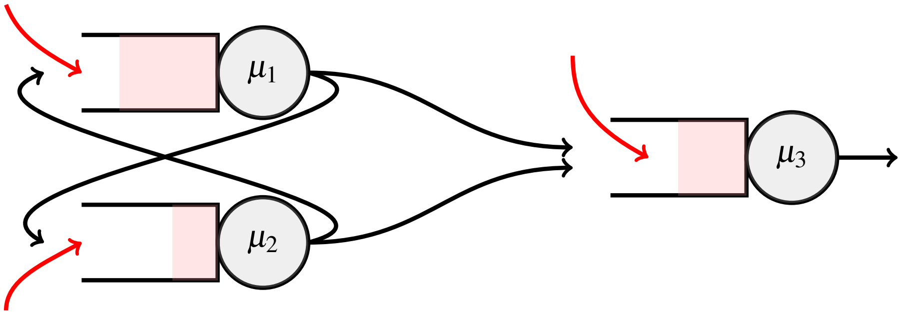
Congestion of Multiple Server Queue:



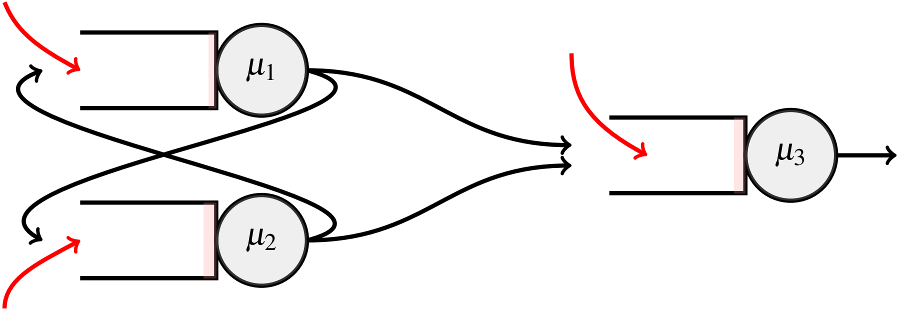
Congestion of Multiple Server Queue:



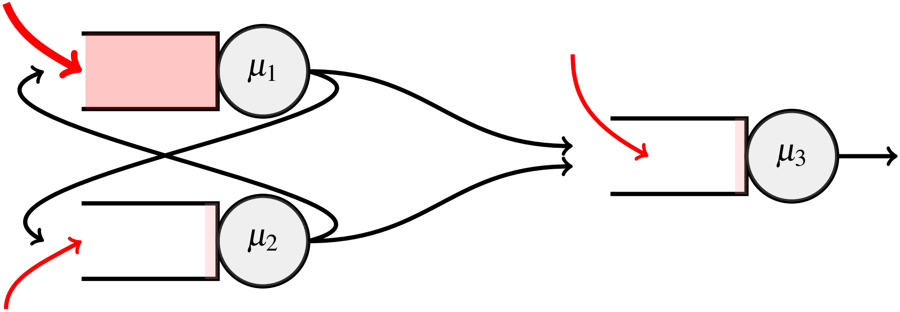
Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:

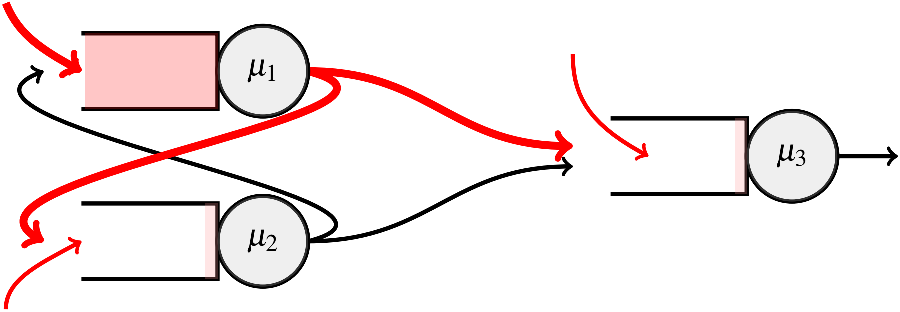


Congestion of Stochastic Fluid Networks:

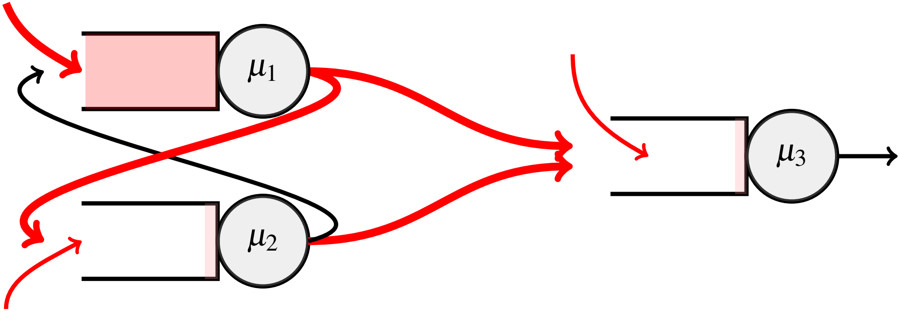


# Conspiracy vs Catastrophe

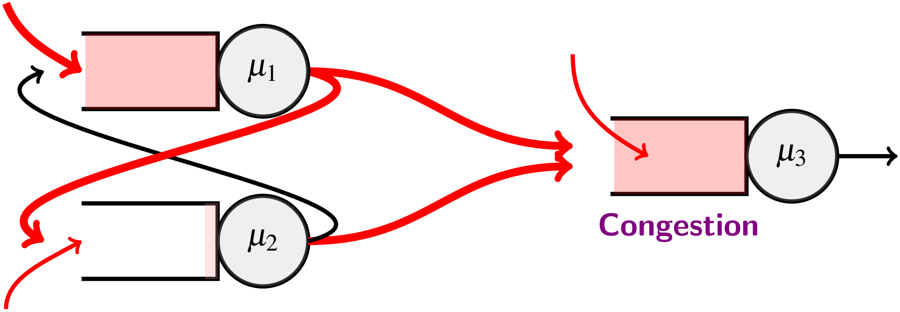
Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



# Lecture 2: Mathematical Machinery for Catastrophe Principle

- $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
  - Formulation
  - Continuous Mapping Theorem
  - Asymptotic Equivalence
  - Product of $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
  - Strengthening $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence by combining multiple $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
- Heavy-Tailed Large Deviations Formulation
- Fundamental Limitation of Classical LDP Formulation
  - No Regularly Varying Lévy Process Satisfies classical LDP for $J_1$ topology

# Lecture 2: Mathematical Machinery for Catastrophe Principle

- $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
  - Formulation (LRR '14)
  - Continuous Mapping Theorem (LRR '14)
  - Asymptotic Equivalence
  - Product of $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
  - Strengthening $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence by combining multiple $\mathbb{M}(\mathbb{S} \setminus \mathbb{C})$-convergence
- Heavy-Tailed Large Deviations Formulation
- Fundamental Limitation of Classical LDP Formulation
  - No Regularly Varying Lévy Process Satisfies classical LDP for $J_1$ topology

# Lecture 2: Mathematical Machinery for Catastrophe Principle

- $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-convergence
  - Formulation (LRR '14)
  - Continuous Mapping Theorem (LRR '14)
  - Asymptotic Equivalence (**R**BZ '19)
  - Product of $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-convergence (**R**BZ '19)
  - Strengthening $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-convergence by combining multiple $\mathbb{M}(\mathbb{S}\setminus\mathbb{C})$-convergence (**R**BZ '19)
- Heavy-Tailed Large Deviations Formulation
- Fundamental Limitation of Classical LDP Formulation
  - No Regularly Varying Lévy Process Satisfies classical LDP for $J_1$ topology

# 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 $f$:

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

# 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

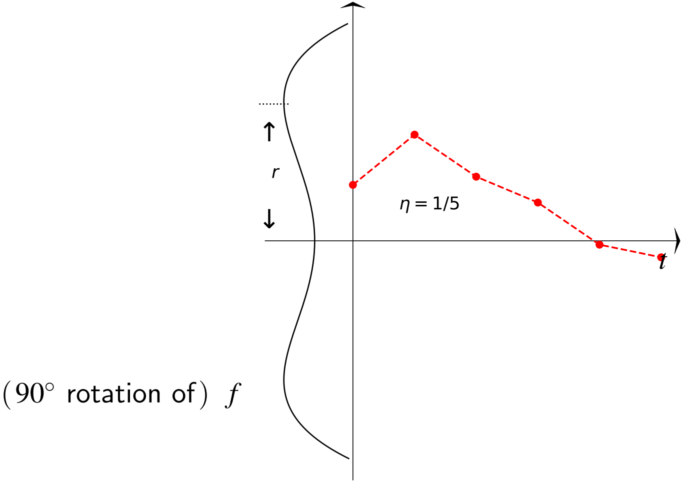
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

Gradient Flow

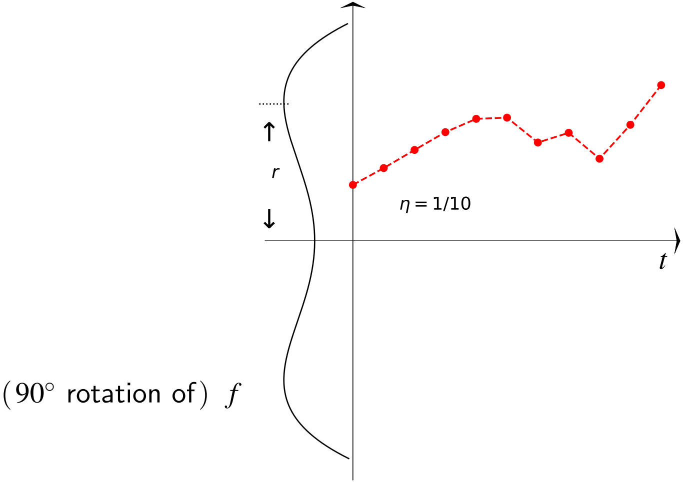
where

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

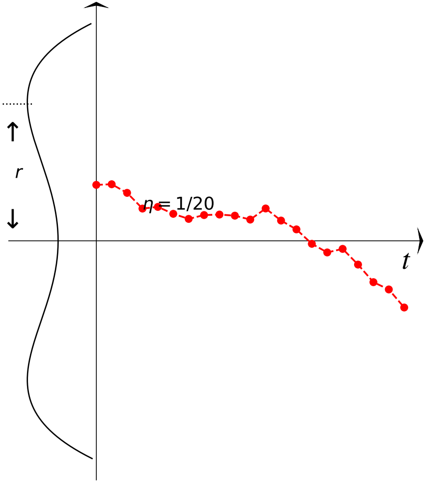
# Gradient Flow: Law of Large Numbers



# Gradient Flow: Law of Large Numbers

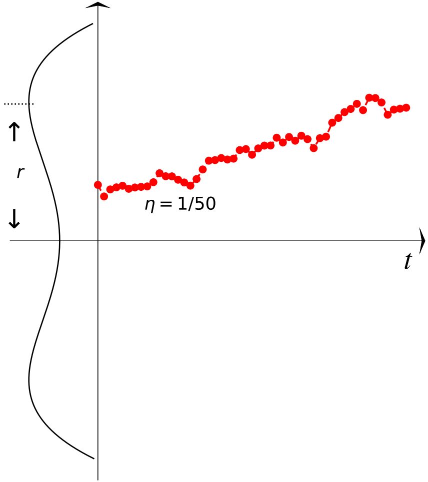


# Gradient Flow: Law of Large Numbers



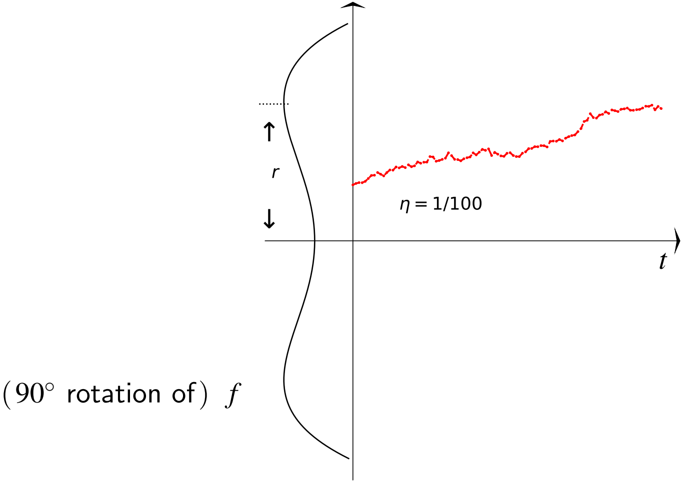

$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers

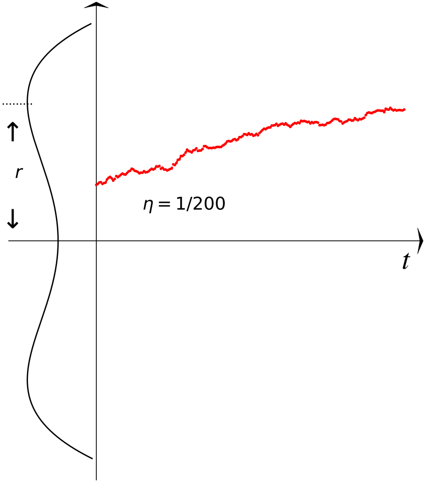



$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers

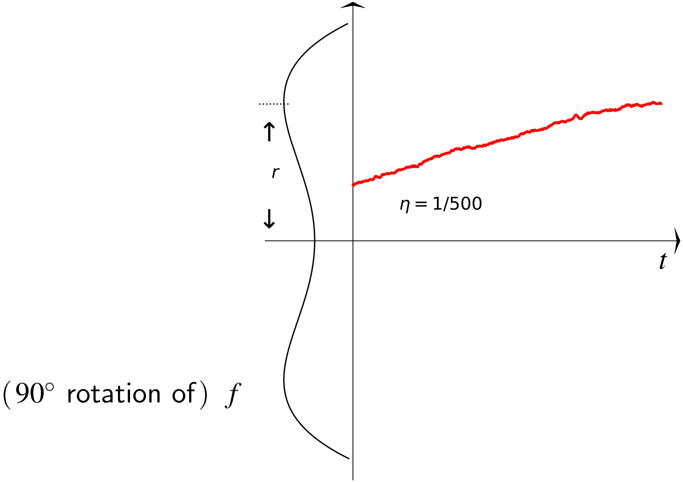


# Gradient Flow: Law of Large Numbers

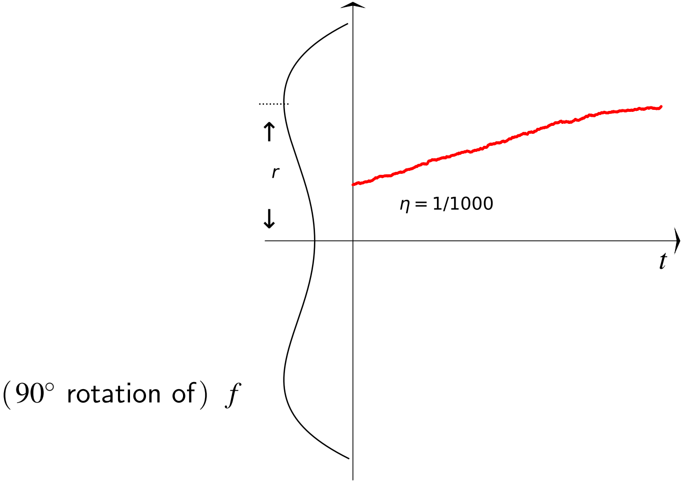



$(90^\circ$ rotation of$)$ $f$

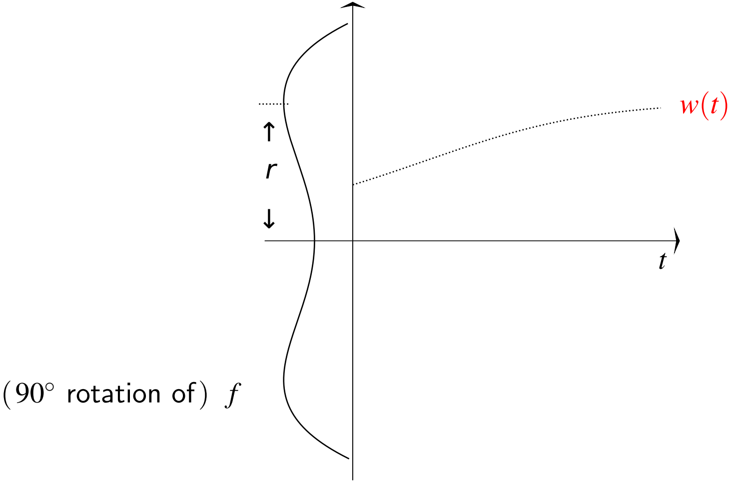
# Gradient Flow: Law of Large Numbers



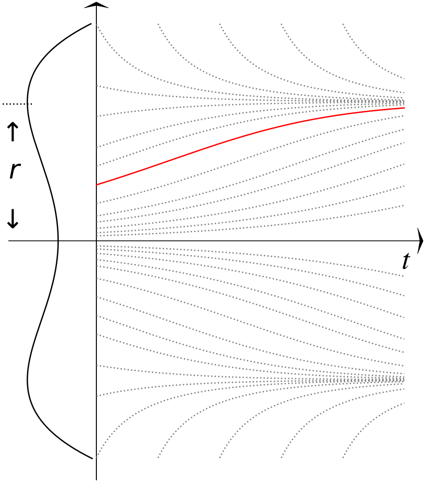
# Gradient Flow: Law of Large Numbers



# Gradient Flow: Law of Large Numbers

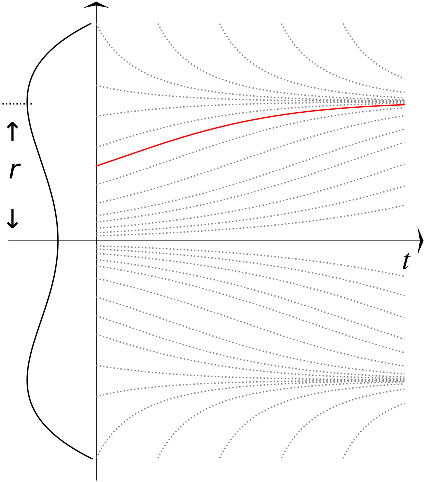


# Gradient Flow: Law of Large Numbers



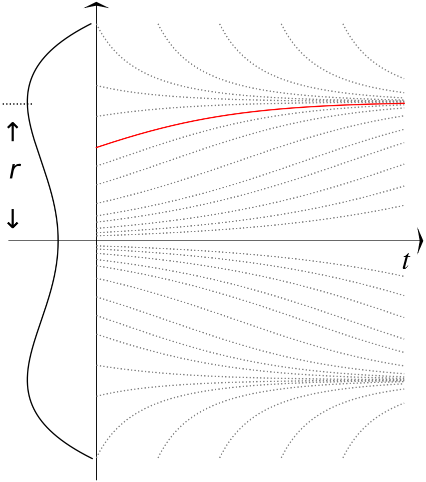

$($90° rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers



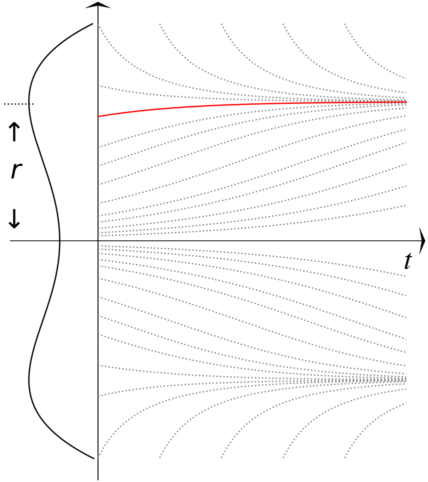

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



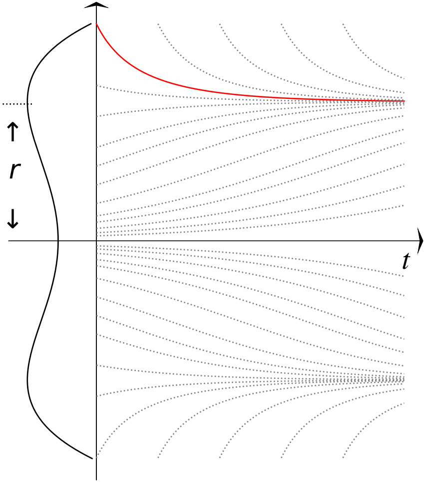

$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers



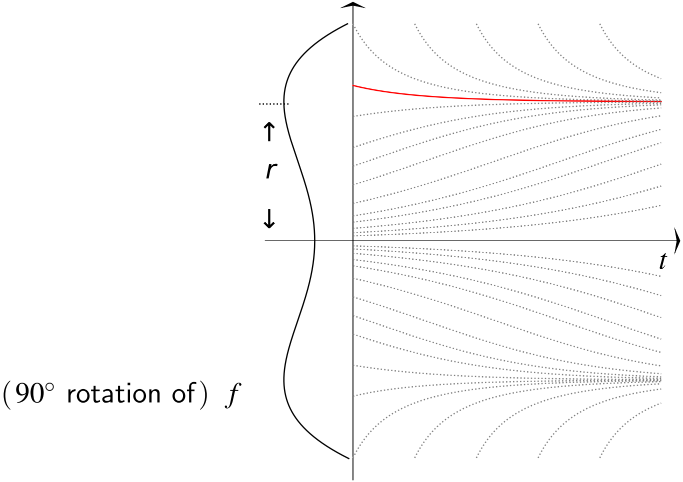

$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers

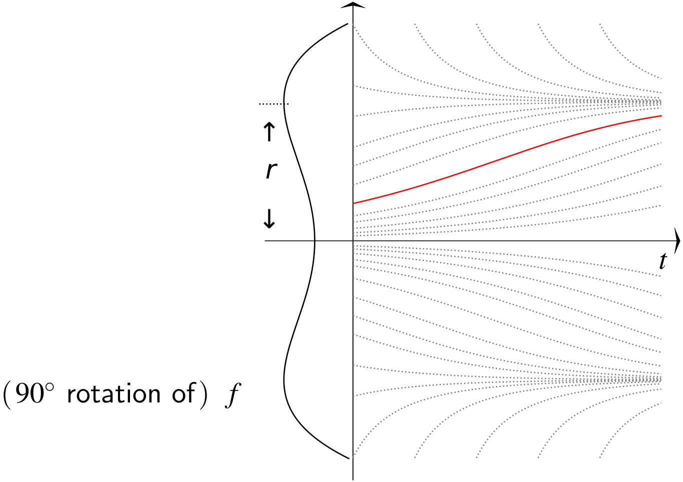



$(90^\circ \text{ rotation of})\ f$

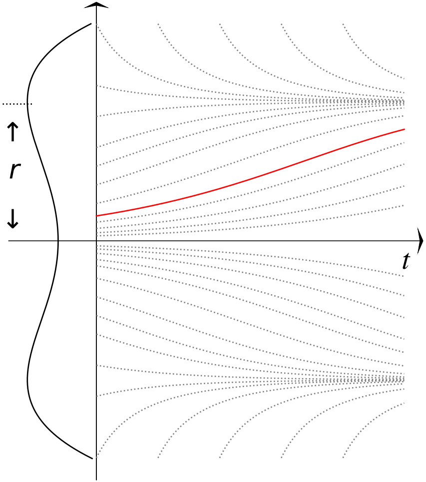
# Gradient Flow: Law of Large Numbers



# Gradient Flow: Law of Large Numbers

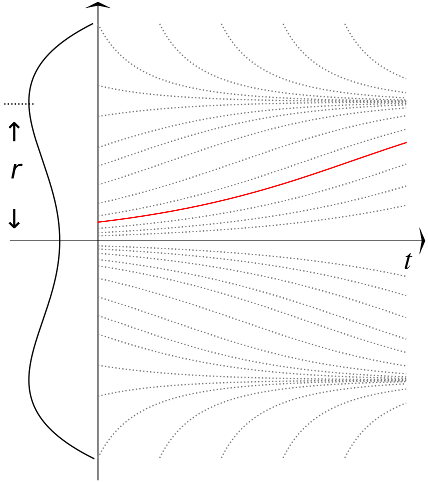


# Gradient Flow: Law of Large Numbers



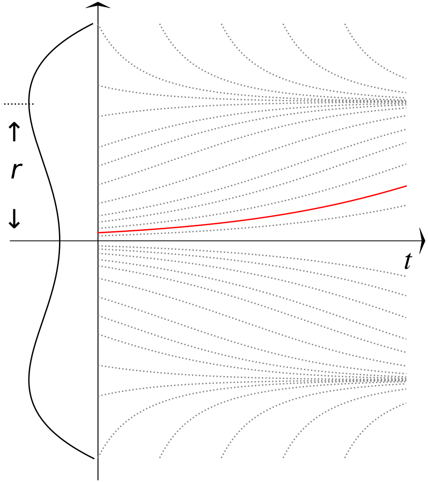

(90° rotation of) $f$

# Gradient Flow: Law of Large Numbers



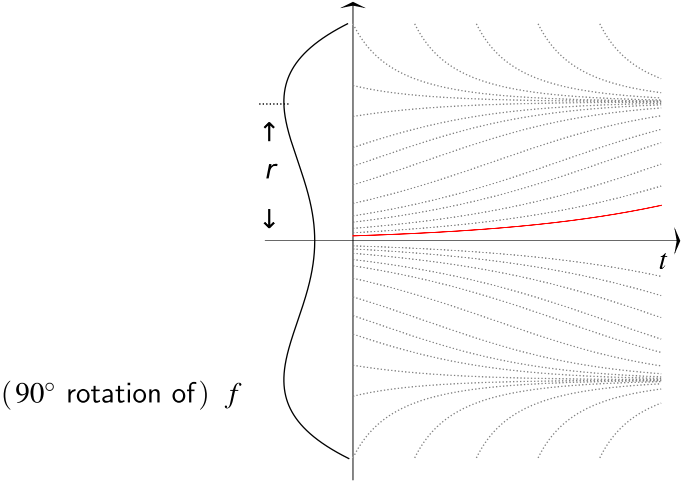

$($90° rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers

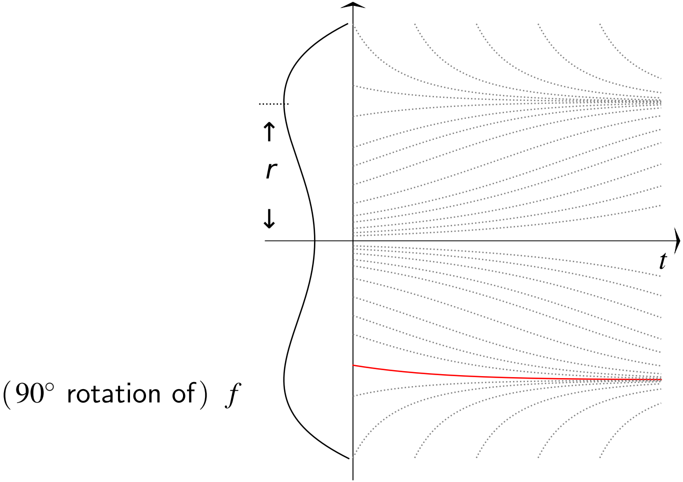



$($90$^\circ$ rotation of$)$ $f$

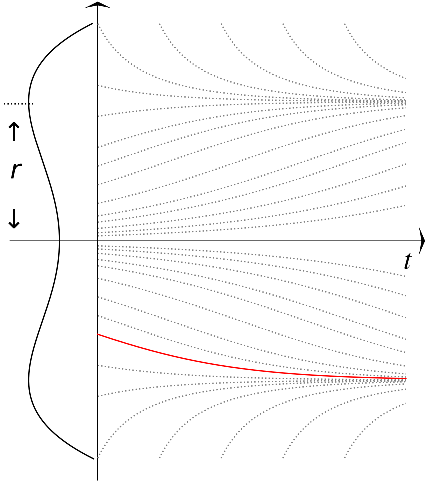
# Gradient Flow: Law of Large Numbers



# Gradient Flow: Law of Large Numbers

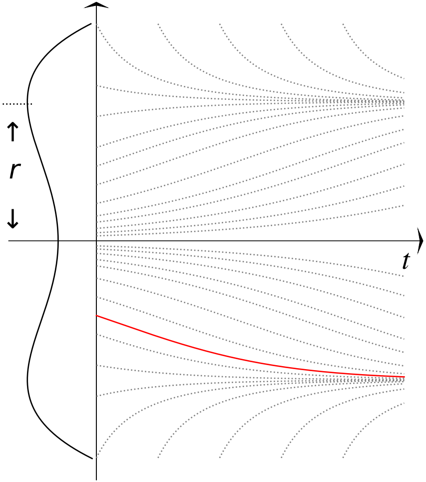


# Gradient Flow: Law of Large Numbers



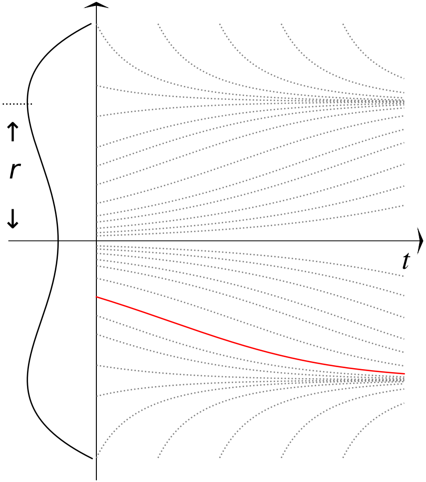

$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers



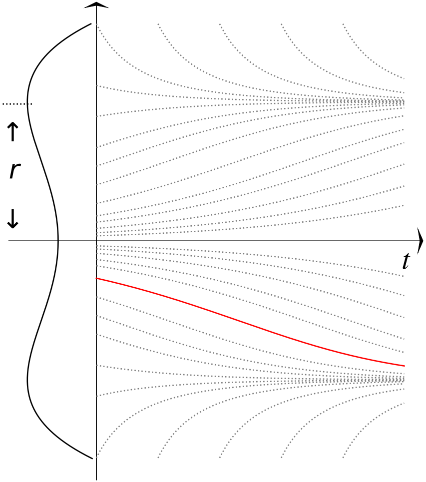

$(90^\circ$ rotation of$)$ $f$

# Gradient Flow: Law of Large Numbers




(90° rotation of) $f$

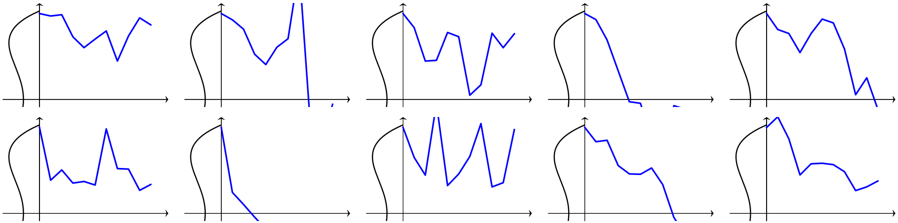
# Gradient Flow: Law of Large Numbers



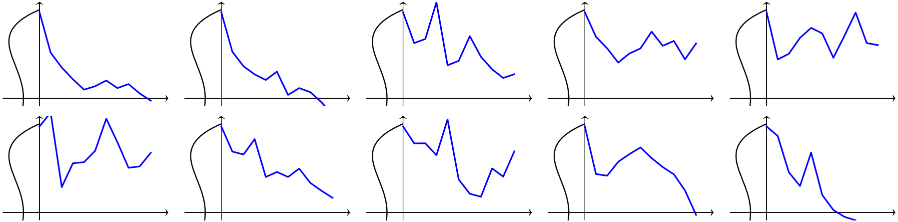

$(90^\circ$ rotation of$)$ $f$

# Typical Scenario

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

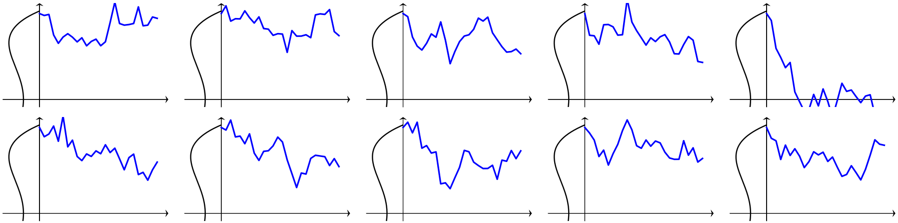


Trajectory of SGD $W^{\eta}$: $\eta = 1/10$ & noises are **heavy-tailed**

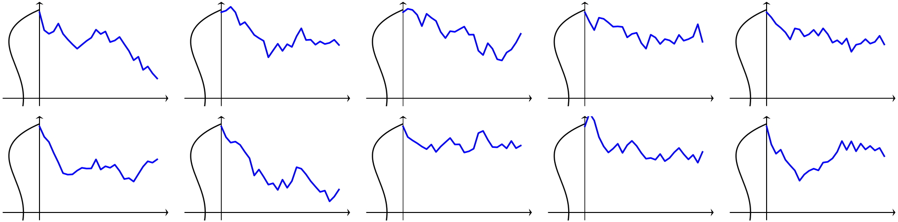


# Typical Scenario

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

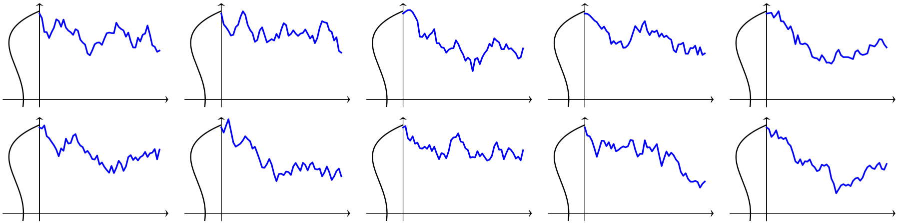


Trajectory of SGD $W^\eta$: $\eta = 1/25$ & noises are **heavy-tailed**

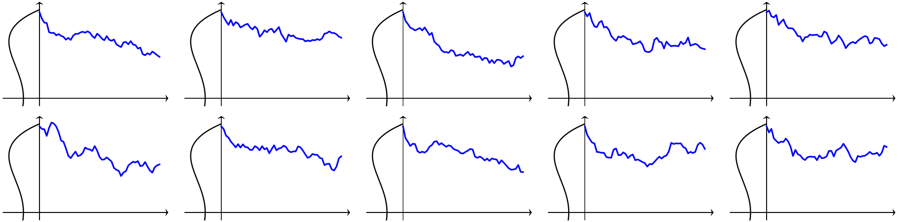


# Typical Scenario

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

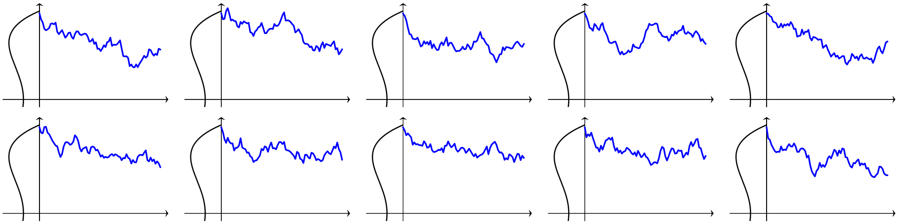


Trajectory of SGD $W^{\eta}$: $\eta = 1/50$ & noises are **heavy-tailed**

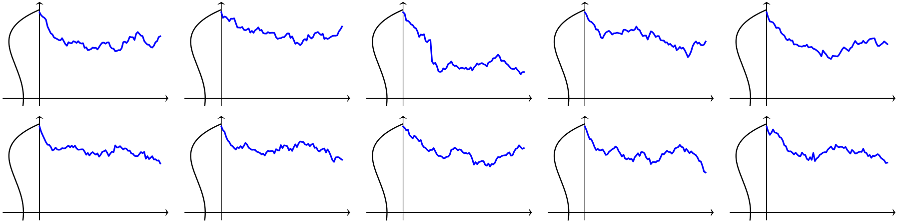


# Typical Scenario

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

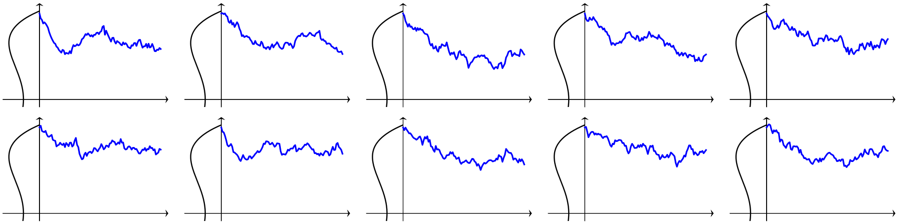


Trajectory of SGD $W^\eta$: $\eta = 1/75$ & noises are **heavy-tailed**

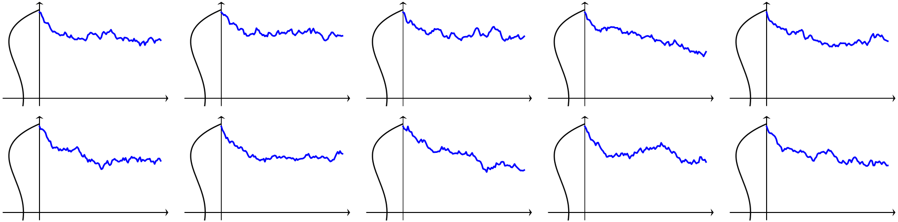


# Typical Scenario

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

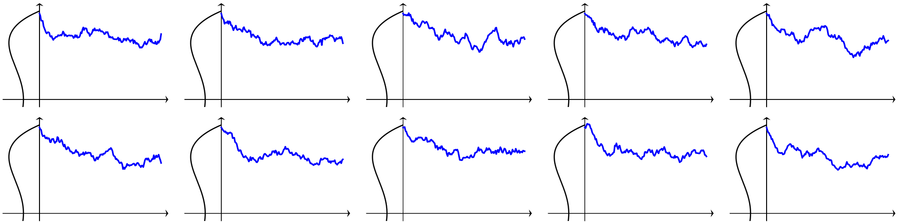


Trajectory of SGD $W^{\eta}$: $\eta = 1/100$ & noises are **heavy-tailed**

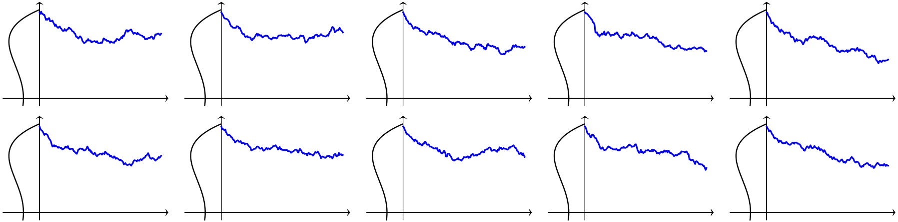


# Typical Scenario

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

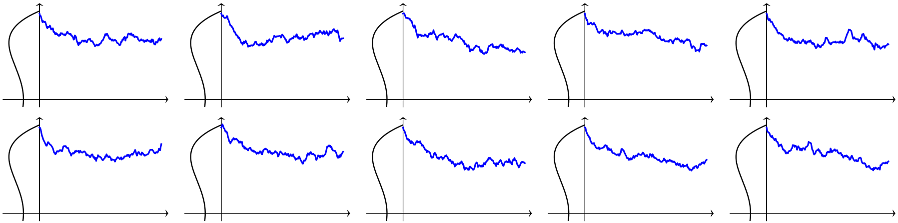


Trajectory of SGD $W^{\eta}$: $\eta = 1/150$ & noises are **heavy-tailed**

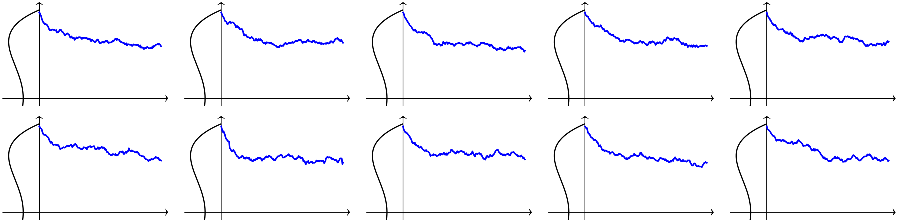


# Typical Scenario

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

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

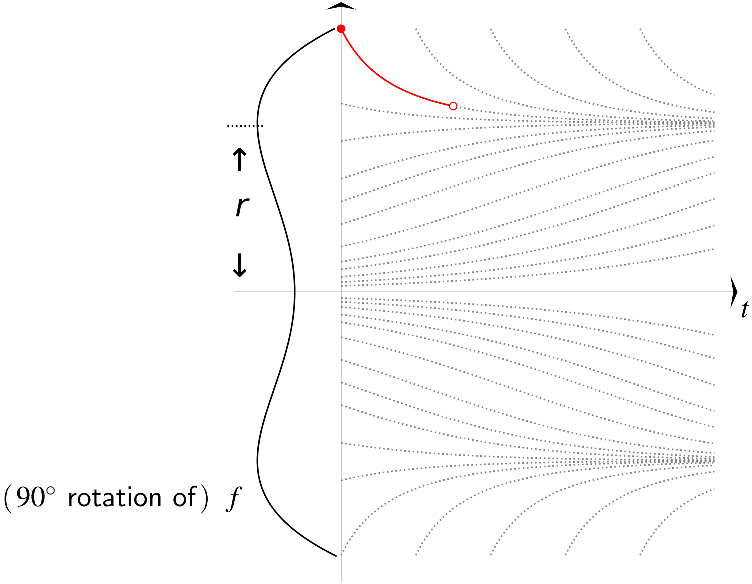
**Theorem (Wang, Su, R., 2022+)**

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

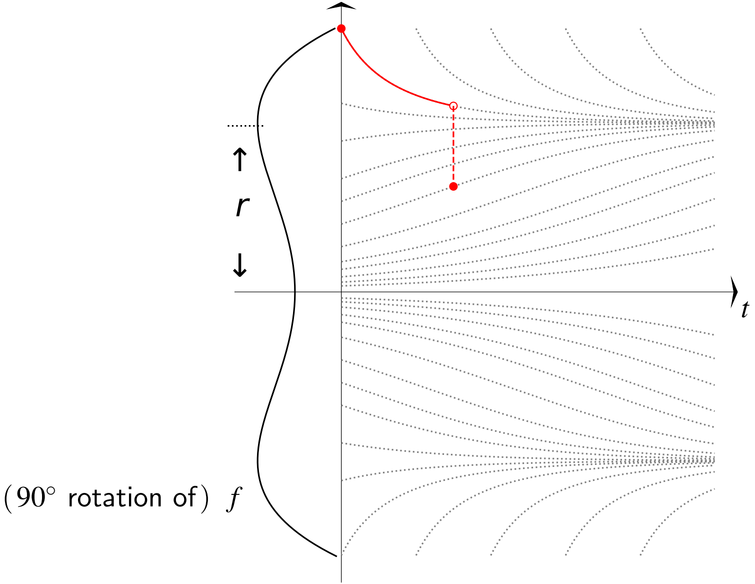
$$\mathbf{P}(W^\eta \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

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

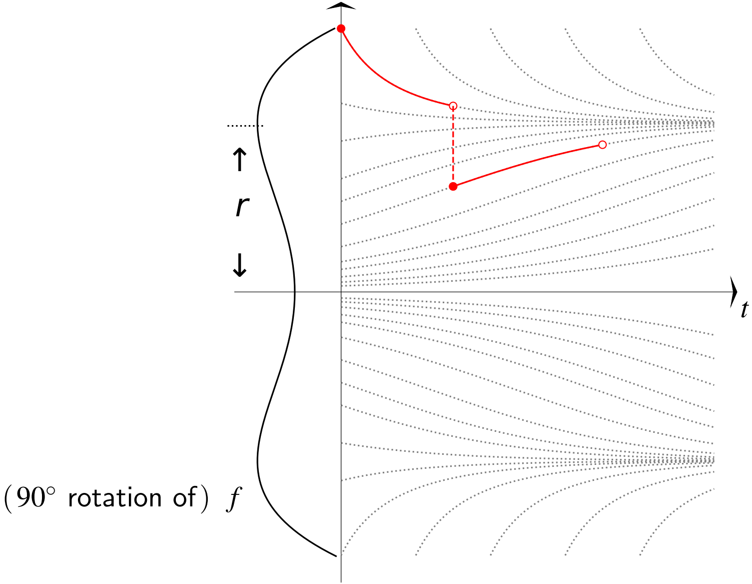
# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



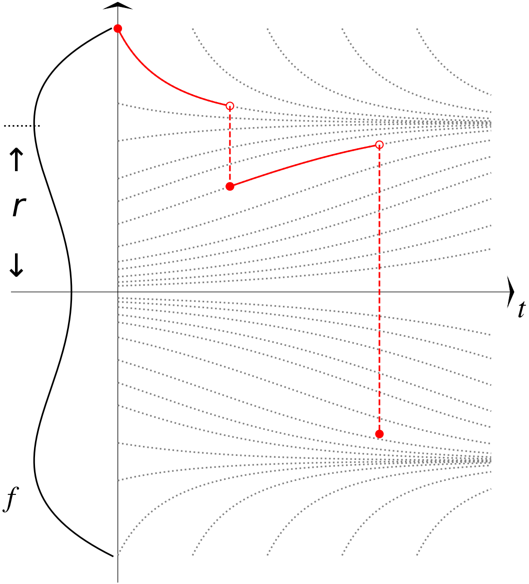
# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



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

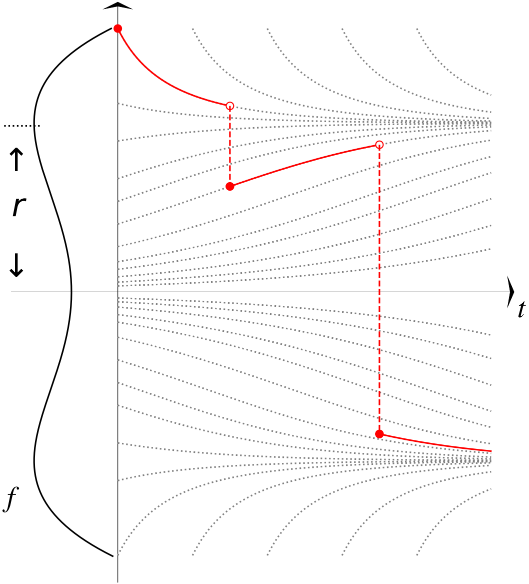


# Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



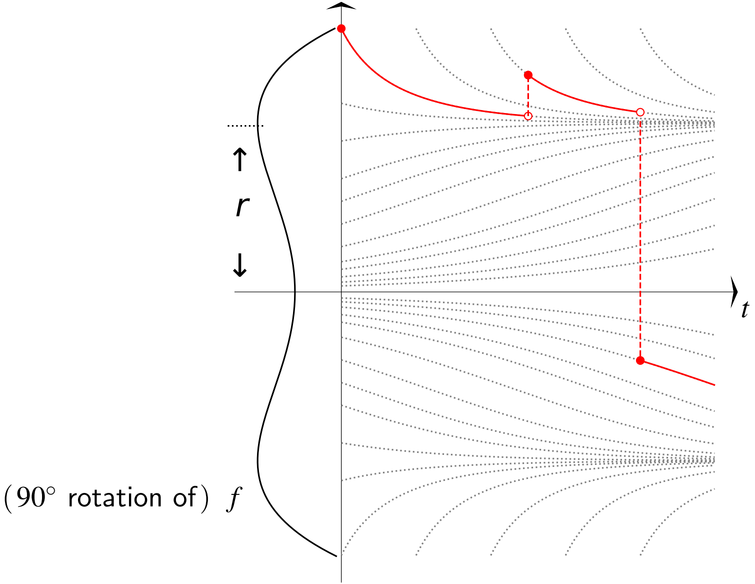

$(90^\circ$ rotation of$)$ $f$

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

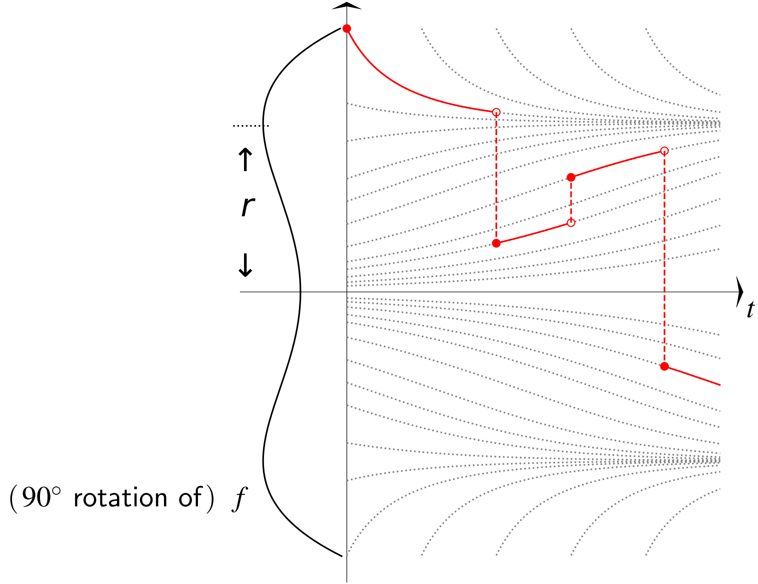



$(90^\circ$ rotation of$)$ $f$

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



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



# Recall: Heavy-Tailed Large Deviations for SGD

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

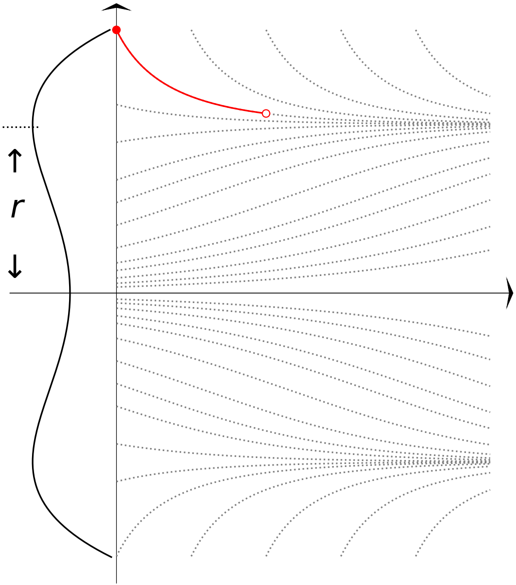
*For "general" $A \subseteq \mathbb{D}$*

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

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

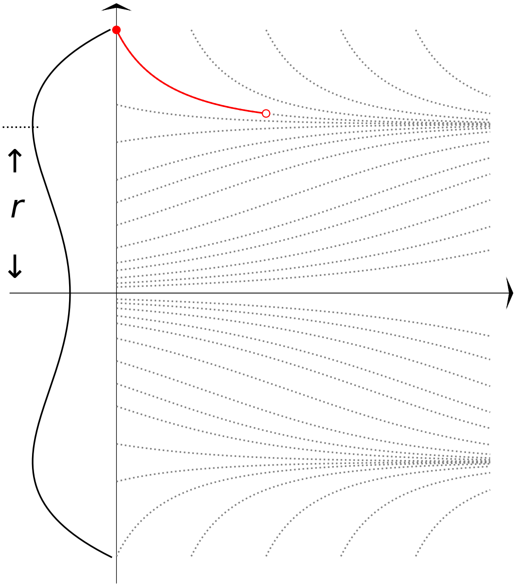
# Catastrophe Principle: Most Likely Escape Route

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



# Catastrophe Principle: Most Likely Escape Route

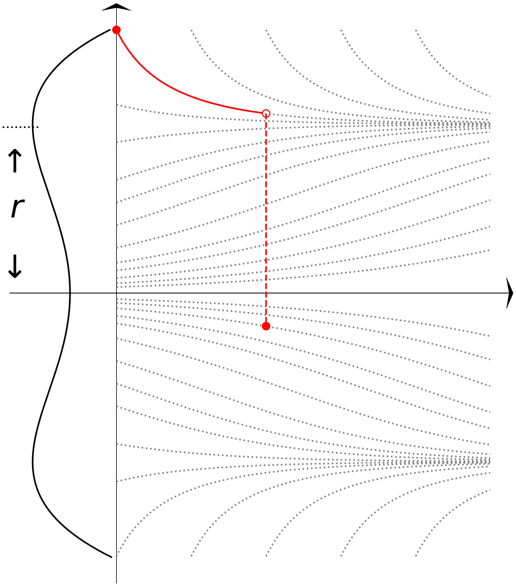
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$




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

# Catastrophe Principle: Most Likely Escape Route

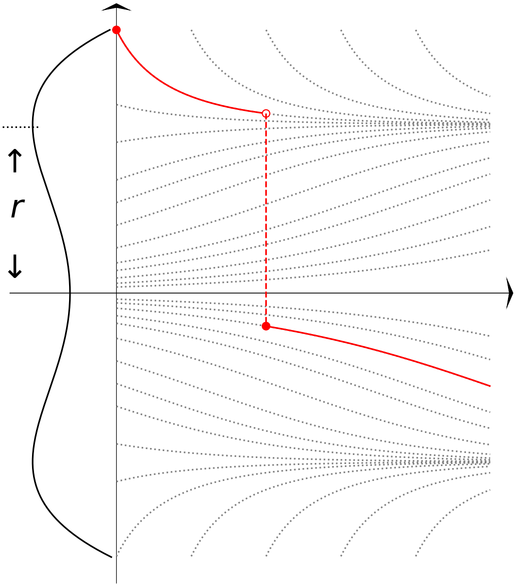
Most Likely Paths for $\{W^{\eta}$ escapes the local minimum$\}$




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

# Catastrophe Principle: Most Likely Escape Route

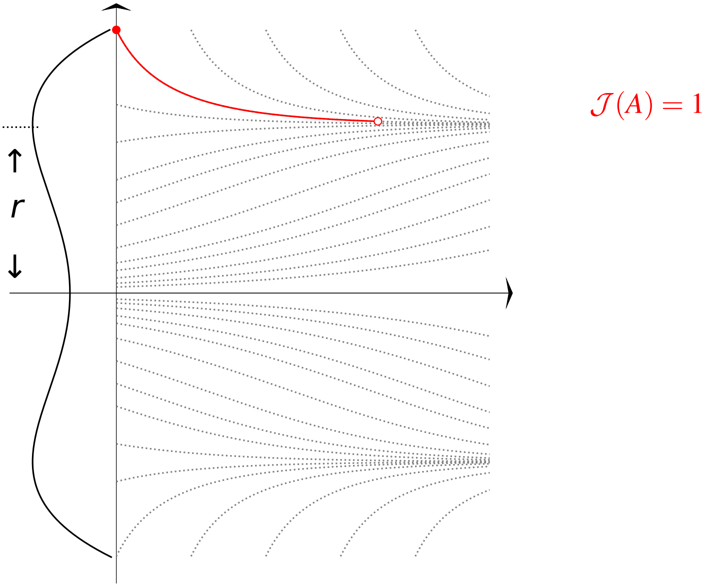
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$




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

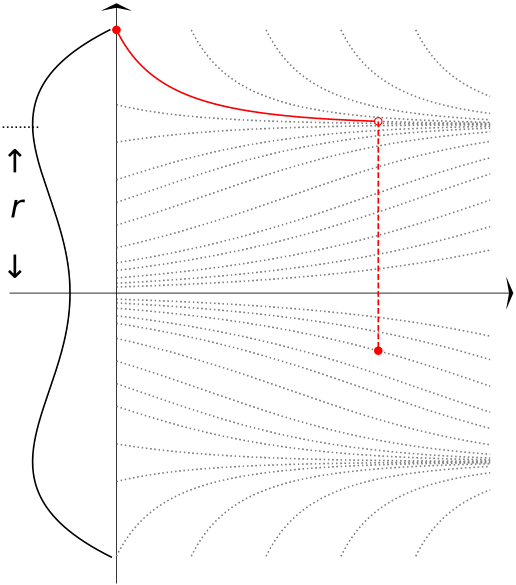
# Catastrophe Principle: Most Likely Escape Route

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



# Catastrophe Principle: Most Likely Escape Route

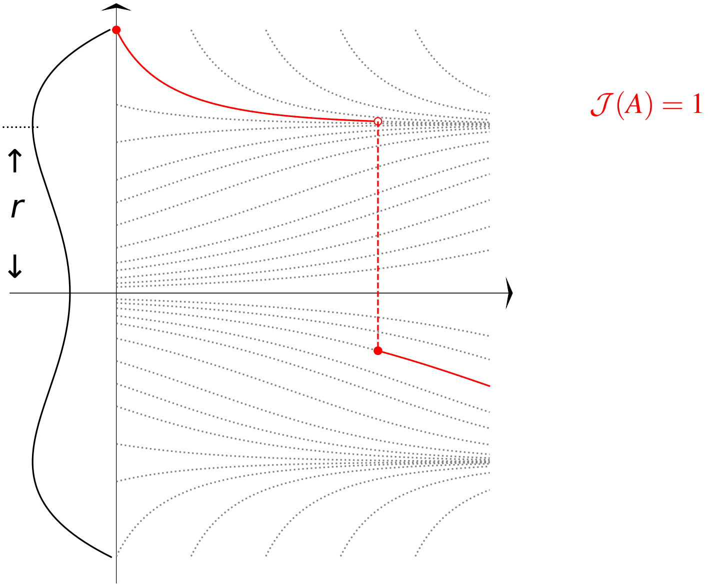
Most Likely Paths for $\{ W^\eta$ escapes the local minimum $\}$




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

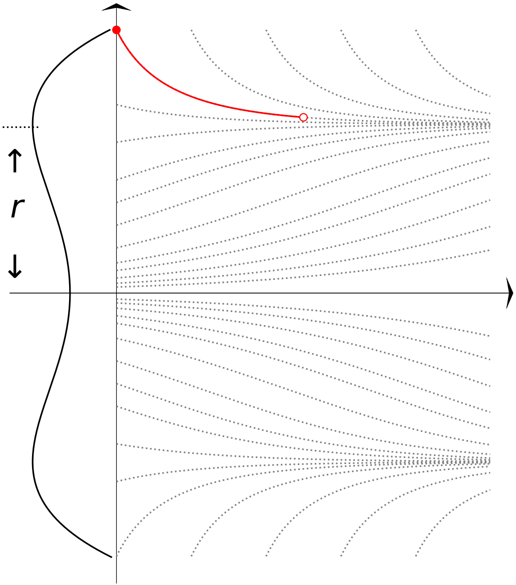
# Catastrophe Principle: Most Likely Escape Route

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



# Catastrophe Principle: Most Likely Escape Route

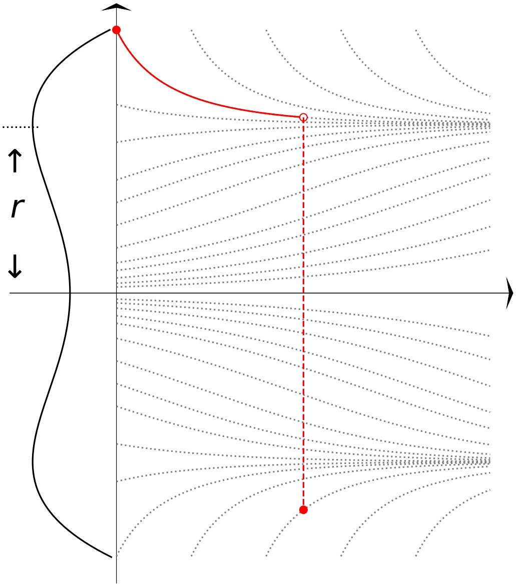
Most Likely Paths for $\{W^{\eta}$ escapes the local minimum$\}$




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

# Catastrophe Principle: Most Likely Escape Route

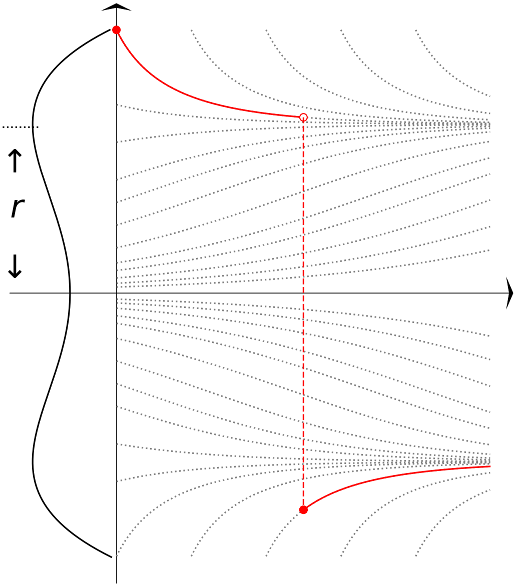
Most Likely Paths for $\{\, W^\eta$ escapes the local minimum $\}$




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

# Catastrophe Principle: Most Likely Escape Route

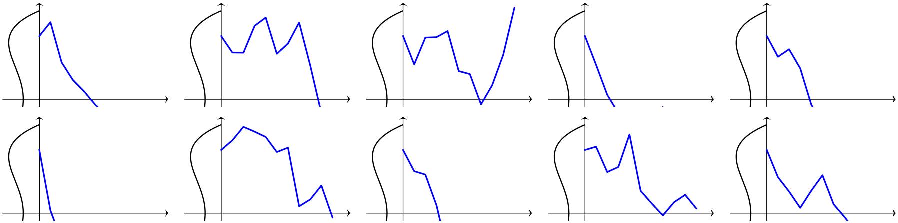
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



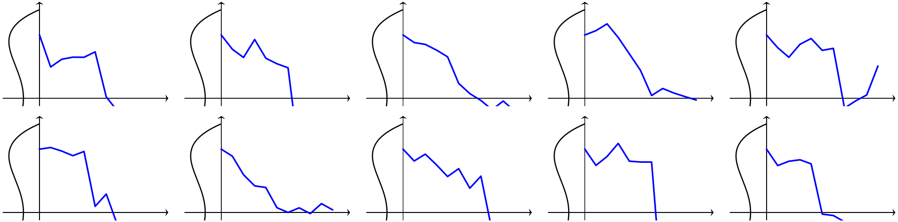

$\mathcal{J}(A) = 1$

# Catastrophe Principle Dictates SGD's Escape Route

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

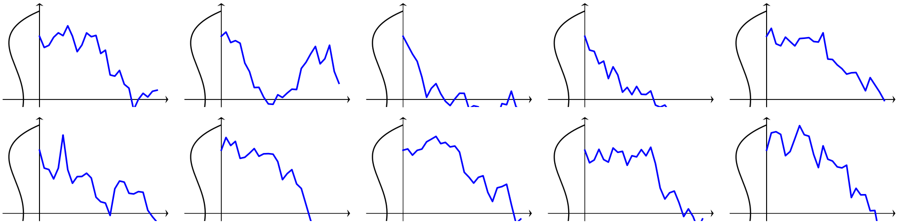


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

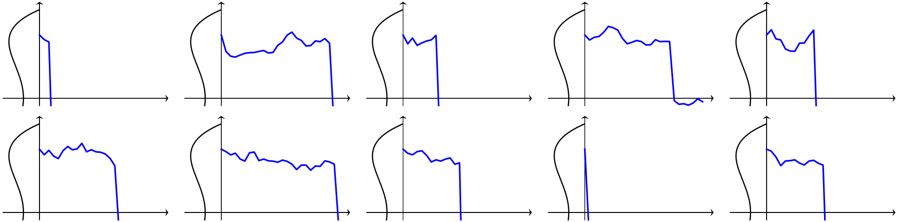


# Catastrophe Principle Dictates SGD's Escape Route

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

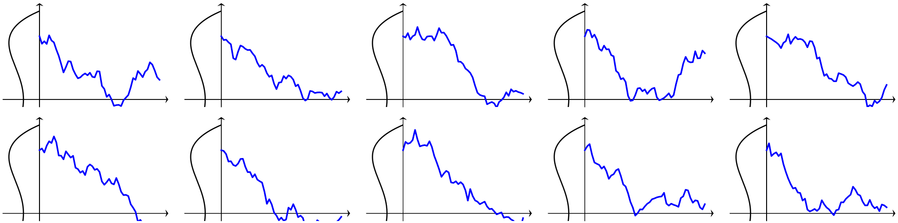


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

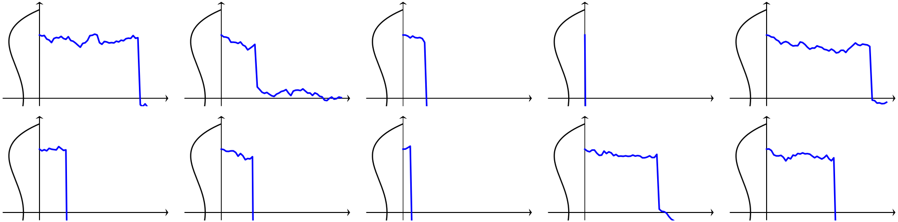


# Catastrophe Principle Dictates SGD's Escape Route

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

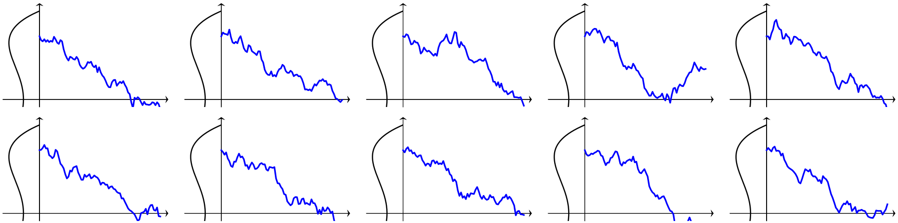


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/50$

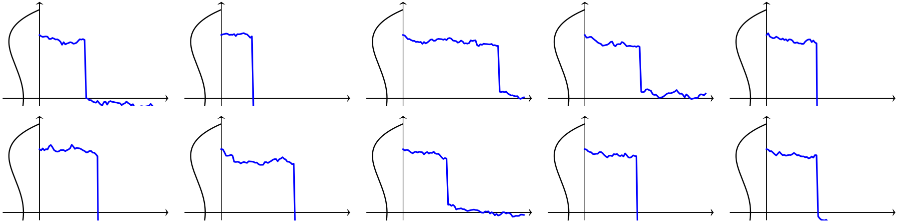


# Catastrophe Principle Dictates SGD's Escape Route

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

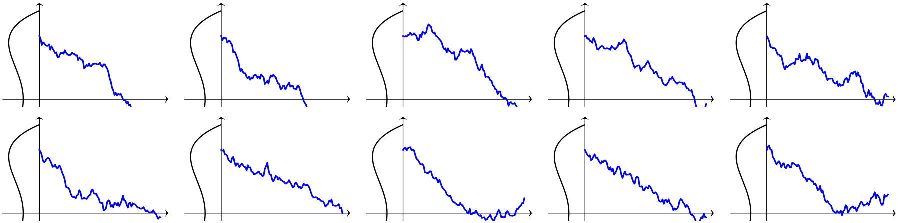


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

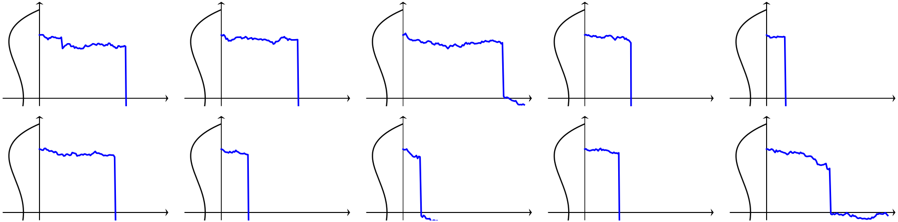


# Catastrophe Principle Dictates SGD's Escape Route

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

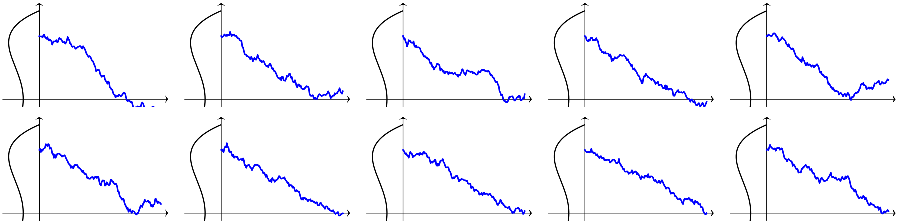


Trajectory of SGD $X^{\eta}$ conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

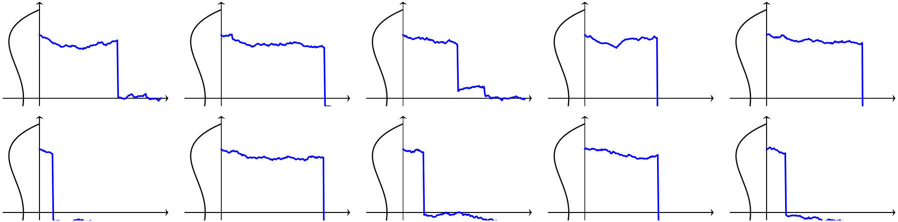


# Catastrophe Principle Dictates SGD's Escape Route

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

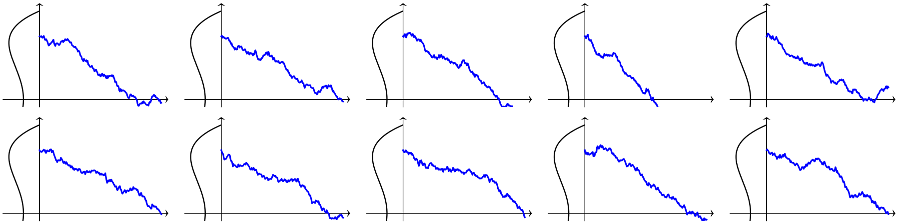


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

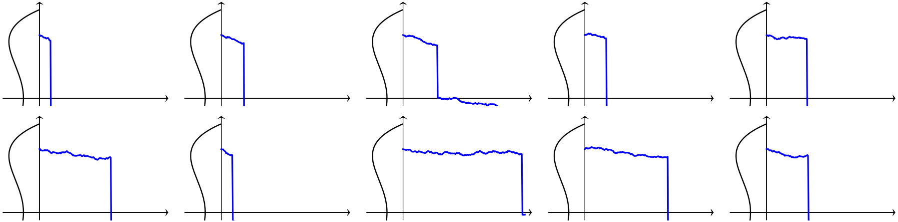


# Catastrophe Principle Dictates SGD's Escape Route

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

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

# 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

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

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,

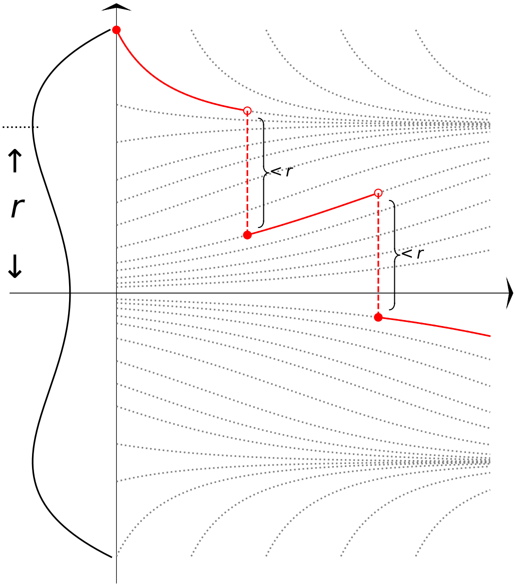
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

where

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

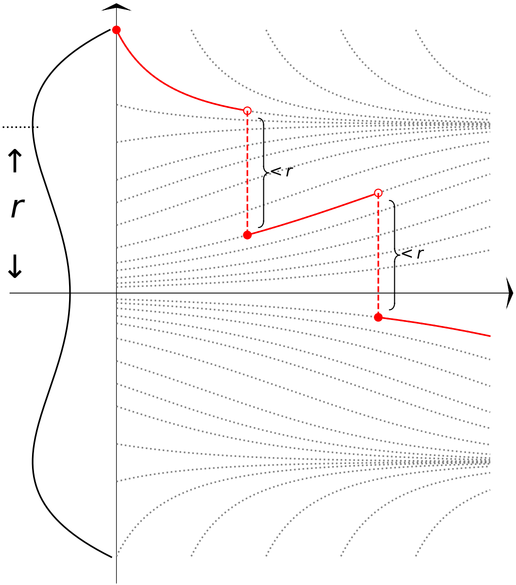
# How does $\mathcal{J}(A)$ change?

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



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

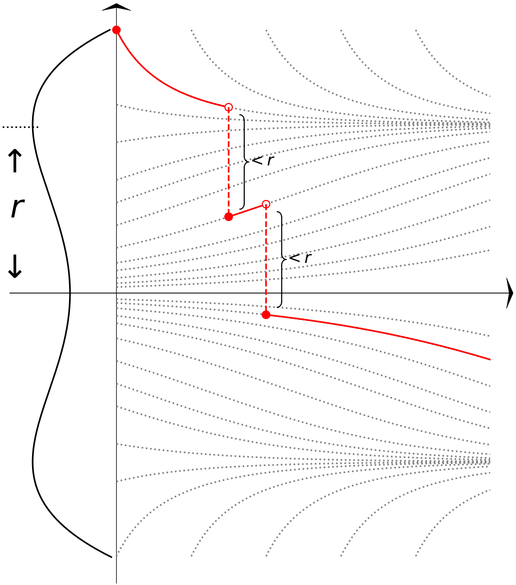
If $r \in (c, 2c)$




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

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

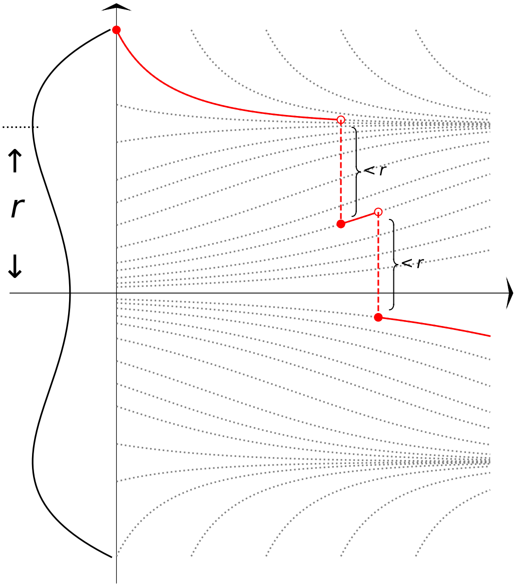
If $r \in (c, 2c)$




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

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

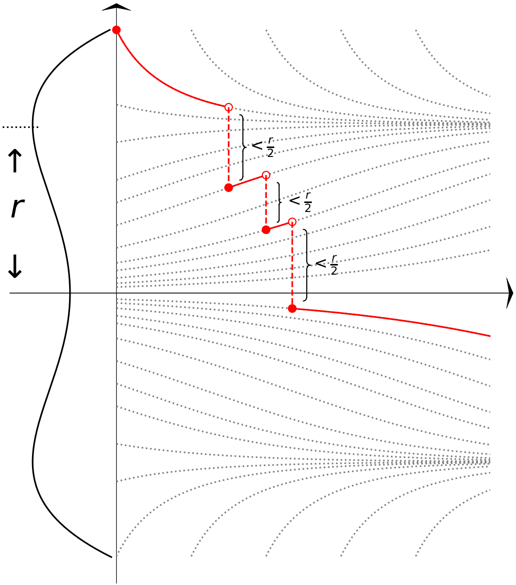
If $r \in (c, 2c)$




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

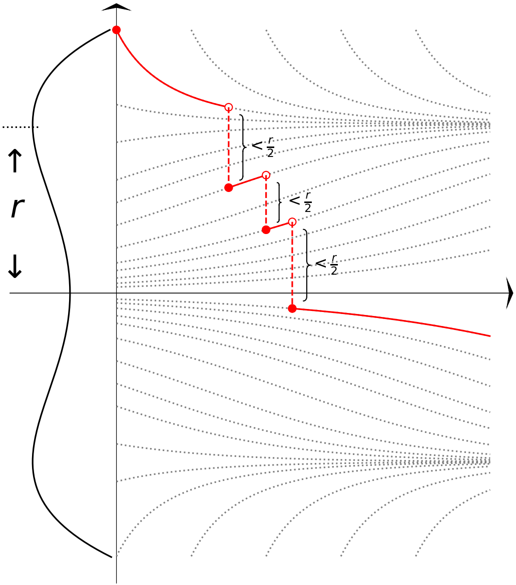
# How does $\mathcal{J}(A)$ change?

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



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

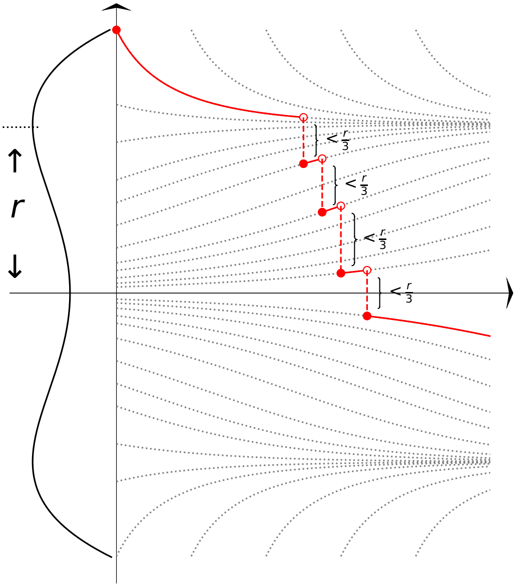
If $r \in (2c, 3c)$




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

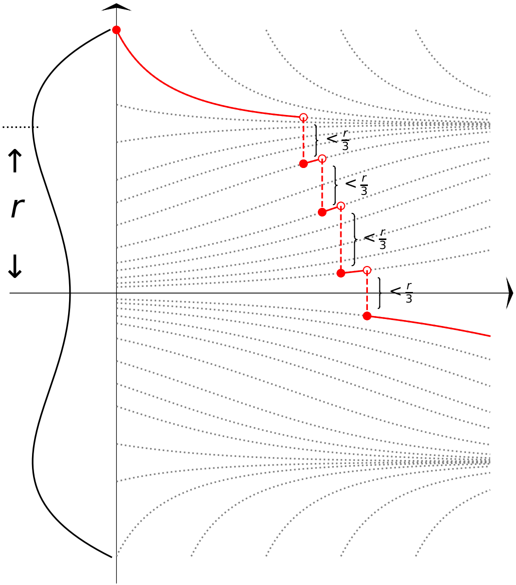
# How does $\mathcal{J}(A)$ change?

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



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

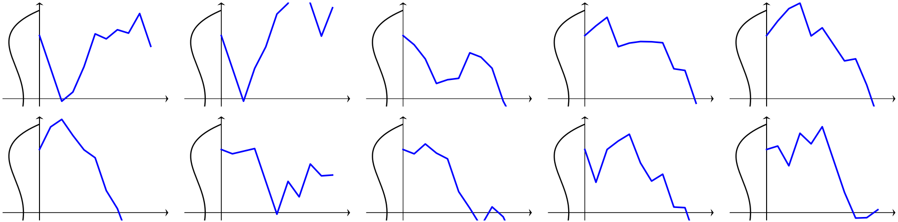
If $r \in (3c, 4c)$



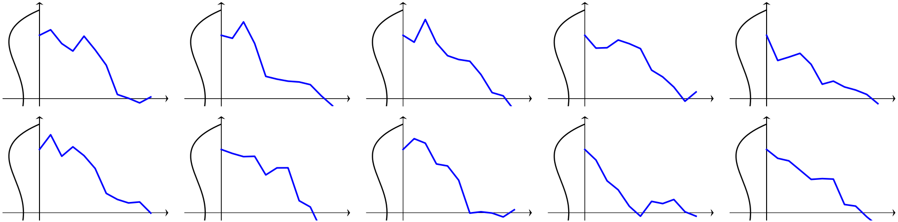

$\mathcal{J}(A) = 4$

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

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

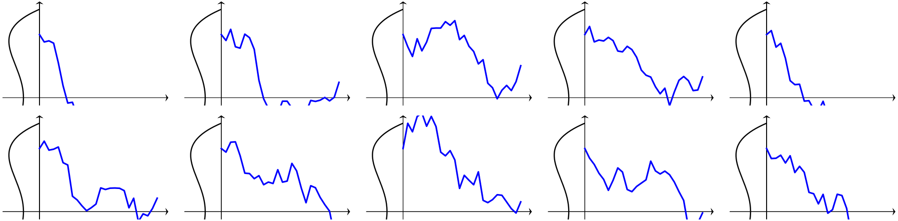


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

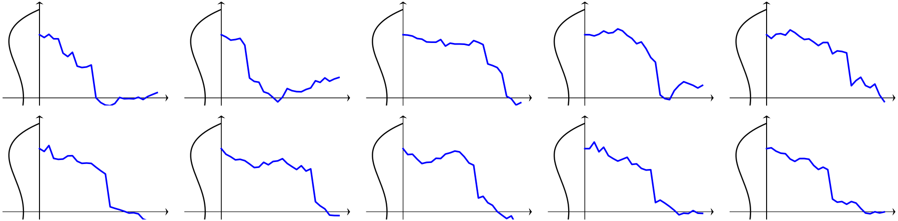


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

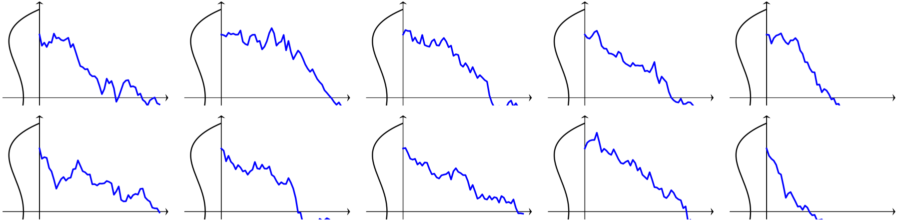


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

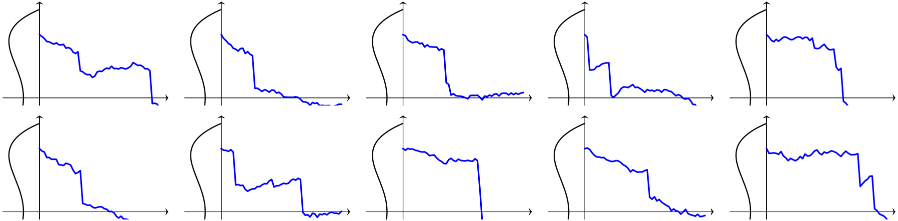


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

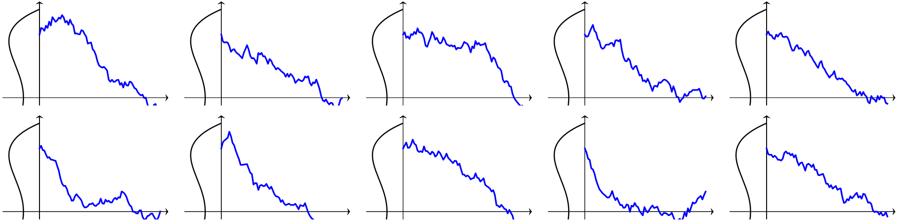


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

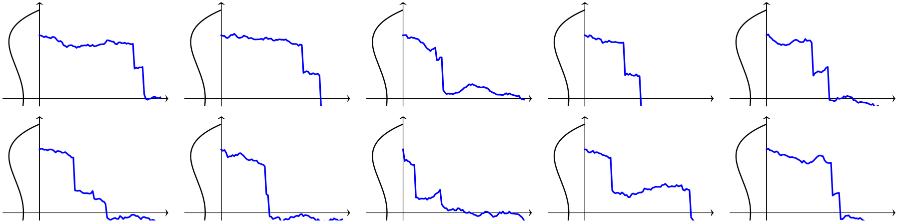


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

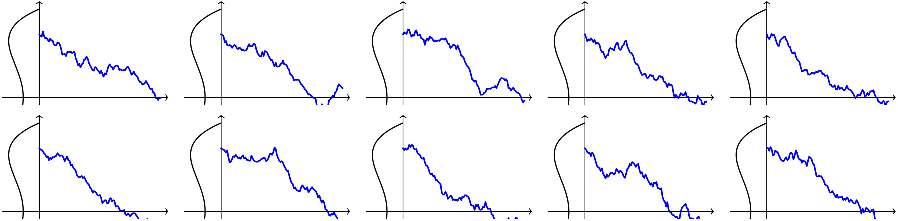


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

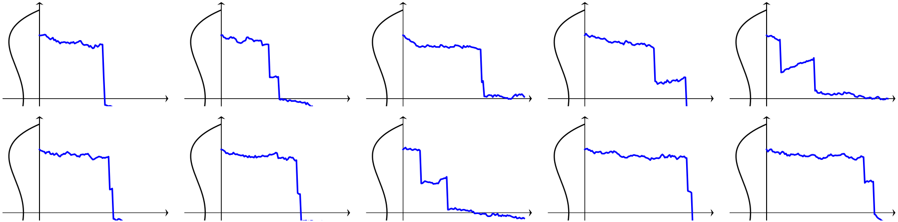


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

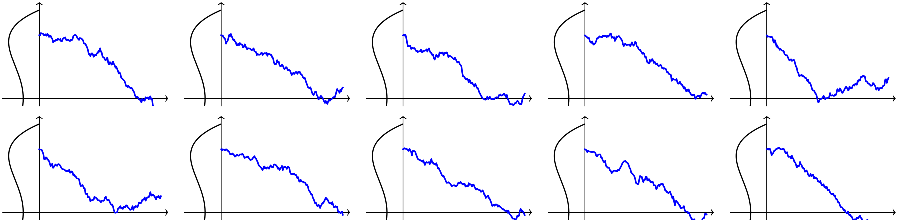


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

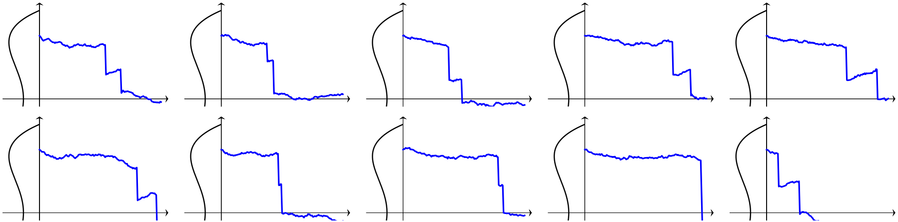


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

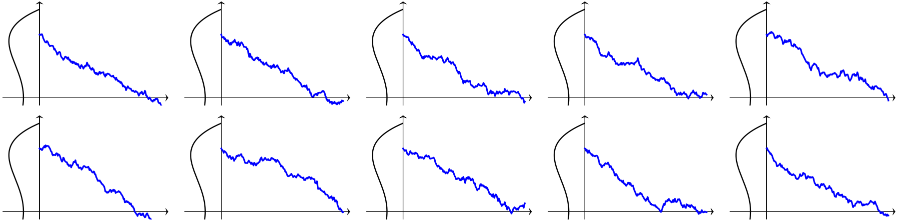


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

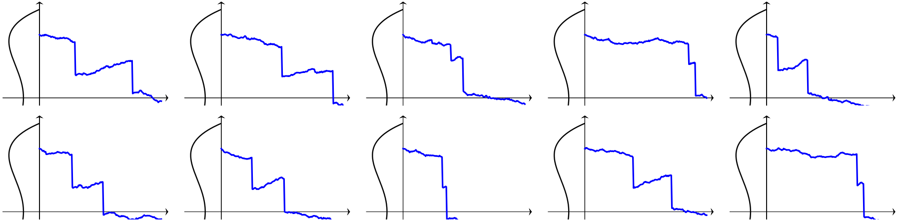


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

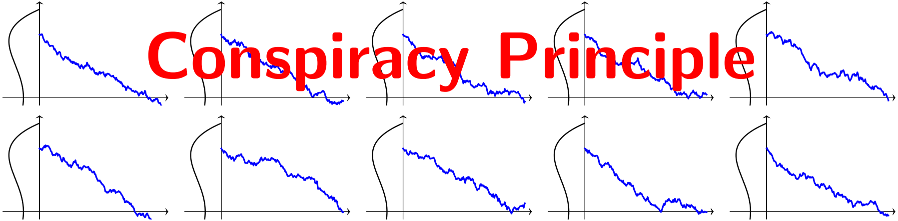


Trajectory of SGD $X^\eta$ conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



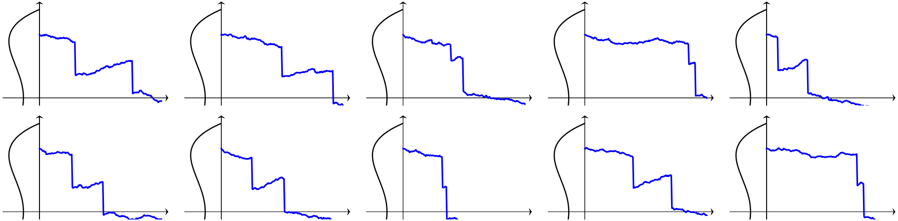
# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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



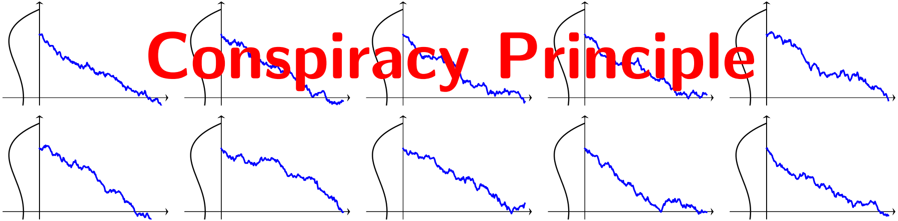
**Conspiracy Principle**

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

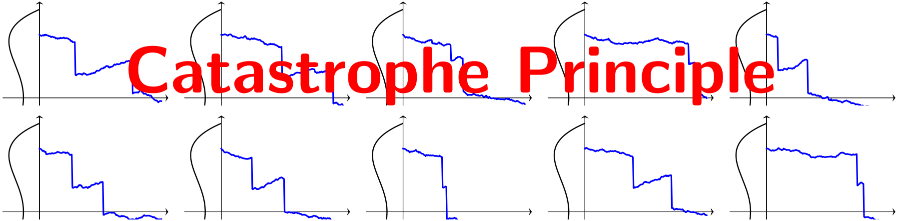


# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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., 2023+)**

*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., 2023+)**

*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., 2023+)**

*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., 2023+)**

*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

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

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

**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 $\mathbb{C}$

$\mathbb{M}(\mathbb{S}\setminus\mathbb{C}) = \{\nu(\cdot) : \nu(\mathbb{S}\setminus\mathbb{C}^{\varepsilon}) < \infty, \forall\varepsilon > 0;\quad \nu(\cdot)$ 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., 2023+)**

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., 2023+)

*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., 2023+)**

*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* $\mathbb{C}$,

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

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

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

**Theorem (Wang, R., 2023+)**

*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* $\mathbb{C}$,
$$\liminf_{\eta\downarrow 0} \inf_{\theta\in\Theta} \big(\mu_\theta^\eta(G) - \mu_\theta(G_\varepsilon)\big) \geq 0$$
$$\limsup_{\eta\downarrow 0} \sup_{\theta\in\Theta} \big(\mu_\theta^\eta(F) - \mu_\theta(F^\varepsilon)\big) \leq 0$$

*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.

$$C(B^\circ) - \delta_B(\varepsilon, T) \leq \liminf_{\eta \downarrow 0} \frac{\inf_{x \in A(\varepsilon)} \mathbf{P}\big(\tau^\eta_{I(\varepsilon)^\complement}(x) \leq T/\eta;\ V^\eta_{\tau_\varepsilon}(x) \in B\big)}{\gamma(\eta) T/\eta}$$

$$\leq \limsup_{\eta \downarrow 0} \frac{\sup_{x \in A(\varepsilon)} \mathbf{P}\big(\tau^\eta_{I(\varepsilon)^\complement}(x) \leq T/\eta;\ V^\eta_{\tau_\varepsilon}(x) \in B\big)}{\gamma(\eta) T/\eta} \leq C(B^-) + \delta_B(\varepsilon, T)$$

$$\limsup_{\eta \downarrow 0} \frac{\sup_{x \in I(\varepsilon)} \mathbf{P}\big(\tau^\eta_{(I(\varepsilon)\setminus A(\varepsilon))^\complement}(x) > T/\eta\big)}{\gamma(\eta) T/\eta} = 0$$

$$\liminf_{\eta \downarrow 0} \inf_{x \in I(\varepsilon)} \mathbf{P}\big(\tau^\eta_{A(\varepsilon)}(x) \leq T/\eta\big) = 1$$

($\{I(\varepsilon) \subseteq I : \varepsilon > 0\}$: covering of $I$)

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

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

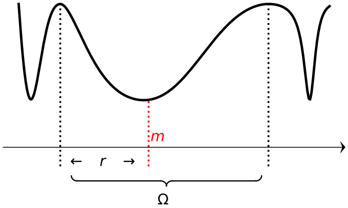
# Exit Time and Location under the Presence of Asymptotic Atom

**Theorem (Wang, R., 2023+)**

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

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

# First Exit Time Analysis for SGD




gradient noise

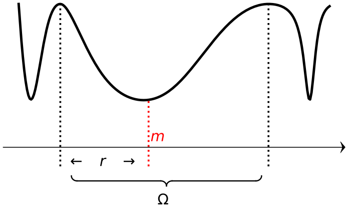
$l = \lceil r/c \rceil$: "width" of the attraction fields, $\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$

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

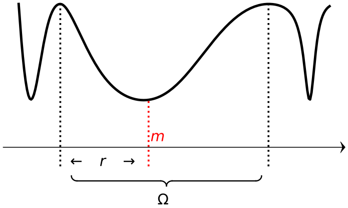
**Theorem (Wang, Oh, R., 2022)**

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

First Exit Time

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

# First Exit Time Analysis for SGD




gradient noise

$l = \lceil r/c \rceil$: "width" of the attraction fields, $\qquad \mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$

**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)**

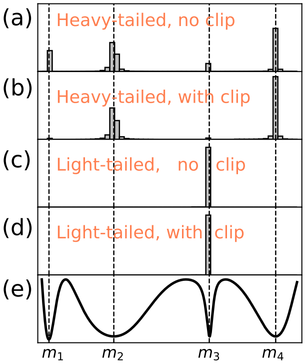
*Under certain structural conditions, for any $t > 0$ and $\beta > 1 + \alpha \cdot l^*$,*

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

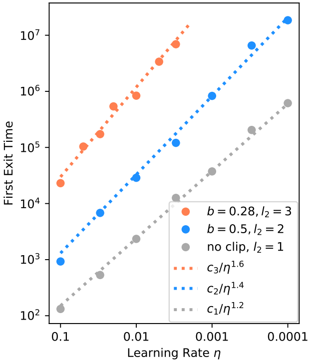
**In fact, $W^\eta_{\lfloor t/\eta^{1+\alpha \cdot l^*} \rfloor}$ converges to a Markov jump processes**

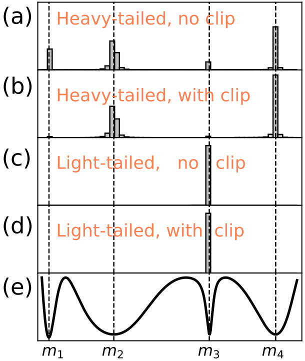
**whose state space consists of wide local minima only.**

# Eliminating Sharp Local Minima with Truncated Heavy-Tails

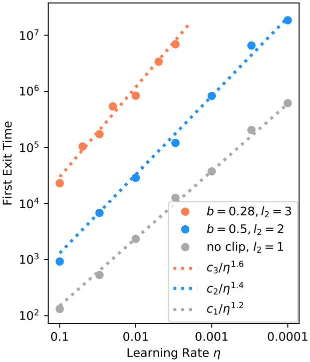


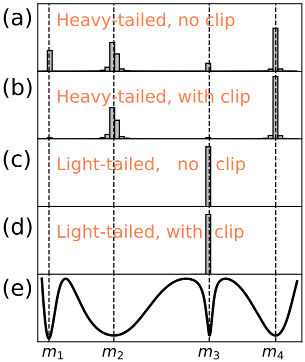
# Eliminating Sharp Local Minima with Truncated Heavy-Tails

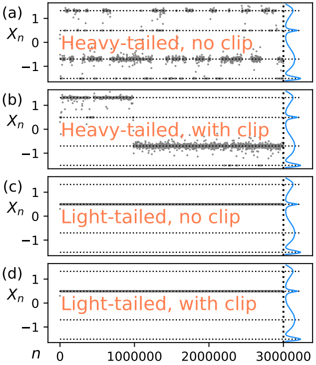


# Eliminating Sharp Local Minima with Truncated Heavy-Tails



# 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

$$\searrow$$

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

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

# Tail-Inflation-Truncation Scheme

Stochastic Gradient

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

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

# Tail-Inflation-Truncation Scheme

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

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

# 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

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

| 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 |

# 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 | $2.75 \times 10^4$ | $\mathbf{2.12 \times 10^4}$ | $2.16 \times 10^4$ |

# Choice of Clipping Threshold $b$

- Recall

$$\tau(\eta) \approx O(1/\eta^{1+\lceil r/b \rceil \alpha})$$

# Choice of Clipping Threshold $b$

- Recall

width of attraction field

$$\tau(\eta) \approx O(1/\eta^{1+\lceil r/b \rceil \alpha})$$

# Choice of Clipping Threshold $b$

- Recall

width of attraction field

$$\tau(\eta) \approx O(1/\eta^{1+\lceil r/b \rceil \alpha})$$

clipping threshold

# Choice of Clipping Threshold $b$

- Recall

width of attraction field

$$\tau(\eta) \approx O(1/\eta^{1+\lceil r/b \rceil \alpha})$$

clipping threshold

- Performance under different $b$

# Choice of Clipping Threshold $b$

- Recall

width of attraction field

$$\tau(\eta) \approx O(1/\eta^{1+\lceil r/b \rceil \alpha})$$

clipping threshold

- Performance under different $b$

| b | 0.005 | 0.025 | 0.1 | 0.25 | 0.5 | 1 | 2.5 | 5 |
|---|---|---|---|---|---|---|---|---|
| Test Accuracy (%) | 70.2 | 71.1 | 74.6 | 76.9 | 76.9 | 76.3 | 52.5 | 30.4 |

CIFAR10-VGG11. $\alpha = 1.4$. $\eta = 0.05$. Training iterations 30,000.

# Other Related Topic I: Heavy Tails from Multiplicative Dynamics

# Markov Modulated Process

We want sample path large deviations of

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=0}^{\lfloor nt \rfloor} X_i \qquad \text{where} \qquad X_{n+1} = A_n X_n + B_n$$

**Light-Tailed Components**

# Markov Modulated Process

We want sample path large deviations of

$$\bar{S}_n(t) = \frac{1}{n} \sum_{i=0}^{\lfloor nt \rfloor} X_i \qquad \text{where} \qquad X_{n+1} = A_n X_n + B_n$$

# Markov Modulated Process

We want sample path large deviations of

$$\bar{S}_n(t) = \frac{1}{n} \sum_{i=0}^{\lfloor nt \rfloor} X_i \qquad \text{where} \qquad X_{n+1} = A_n X_n + B_n$$

$$\mathbf{E} A_n^{\alpha} = 1$$

**for some** $\alpha > 1$

# Markov Modulated Process

We want sample path large deviations of

$\mathbf{E}|B_n|^{\alpha+\varepsilon} < \infty$

**for some** $\varepsilon > 0$

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=0}^{\lfloor nt \rfloor} X_i \qquad \text{where} \qquad X_{n+1} = A_n X_n + B_n$$

$\mathbf{E}A_n^{\alpha} = 1$

**for some** $\alpha > 1$

# Large Deviations for Markov Modulated Processes

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_{n+1} = A_n X_n + B_n, \qquad \mathbf{E} A_1^{\alpha} = 1$$

**Theorem (Chen, R., Zwart, 2020+)**

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

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

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

# Large Deviations for Markov Modulated Processes

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_{n+1} = A_n X_n + B_n, \qquad \mathbf{E} A_1^{\alpha} = 1$$

**Theorem (Chen, R., Zwart, 2020+)**

*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* $A$

# Large Deviations for Markov Modulated Processes

$$\bar{S}_n(t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad X_{n+1} = A_n X_n + B_n, \qquad \mathbf{E} A_1^{\alpha} = 1$$

**Theorem (Chen, R., Zwart, 2020+)**

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

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

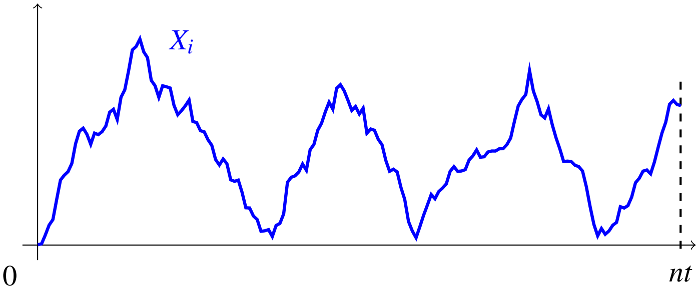
- $\mathcal{J}(A)$: *min #jumps for step functions to be inside* $A$

**Catastrophe Principle:** $\mathcal{L}(\bar{S}_n | \bar{S}_n \in A) \to \mathcal{L}(\text{step functions with } \mathcal{J}(A) \text{ jumps})$

# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (X_0 + X_1 + \cdots + X_{r_1-1} + X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}$

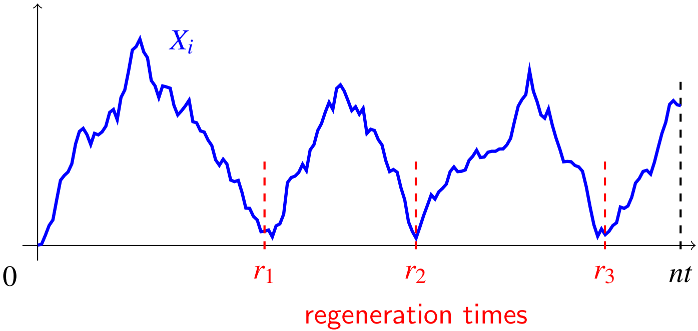
$$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$



# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (X_0 + X_1 + \cdots + X_{r_1 - 1} + X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}$

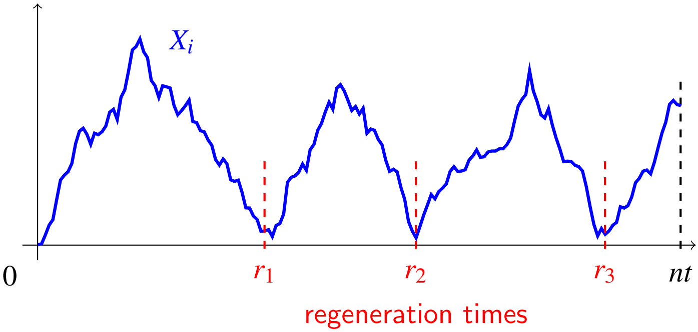
$$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$



# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (X_0 + X_1 + \cdots + X_{r_1-1} + X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}$

$$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$

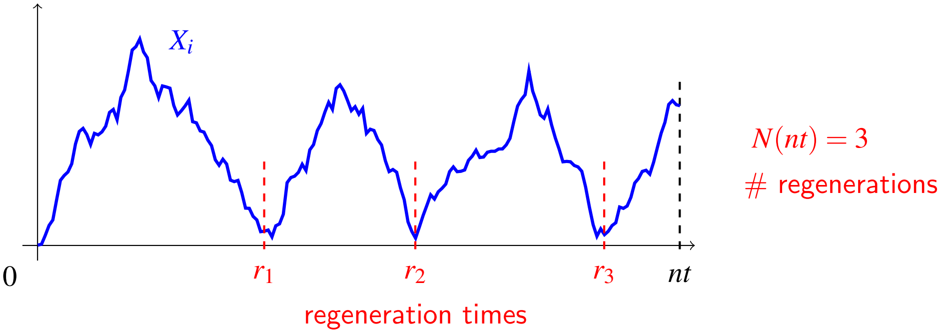



$N(nt)$
# regenerations

# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (X_0 + X_1 + \cdots + X_{r_1-1} + X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}$

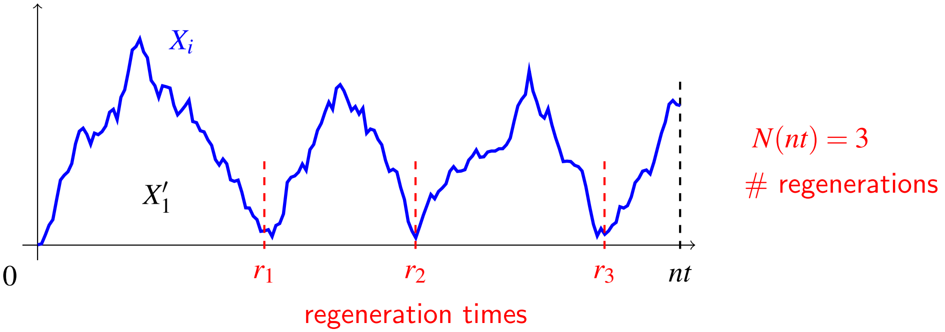
$$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$



# View Regeneration Cycle as Heavy-Tailed Increment

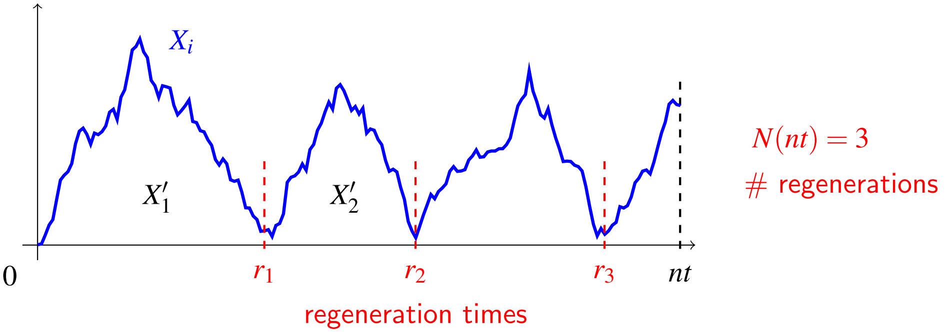
- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1 - 1}}_{\triangleq X_1'} + X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}$

$$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$



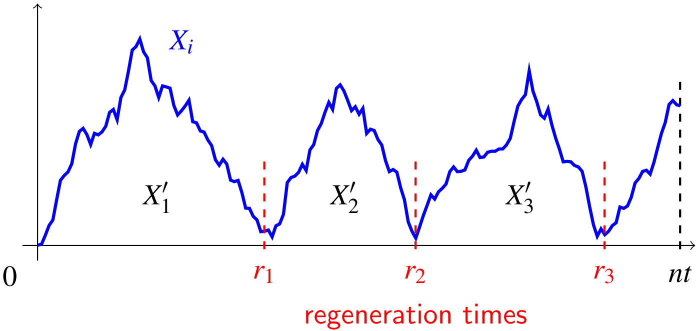
# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1-1}}_{\triangleq X'_1} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X'_2}$
$+ \cdots + X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$



# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1-1}}_{\triangleq X_1'} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X_2'}$
  $+ \cdots + \underbrace{X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1}}_{\triangleq X'_{N(nt)}} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$

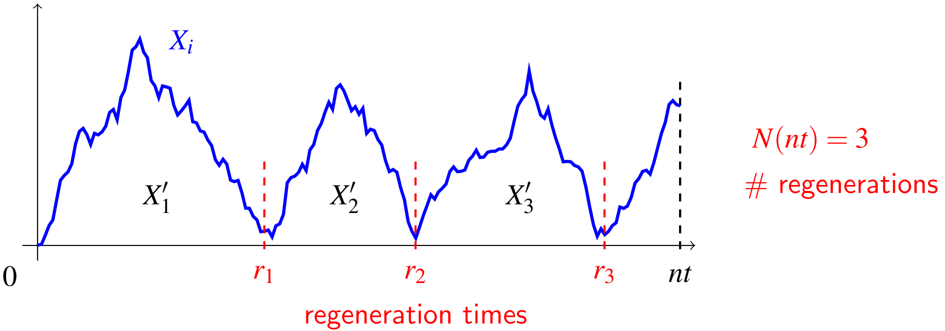



$N(nt) = 3$

# regenerations

# View Regeneration Cycle as Heavy-Tailed Increment

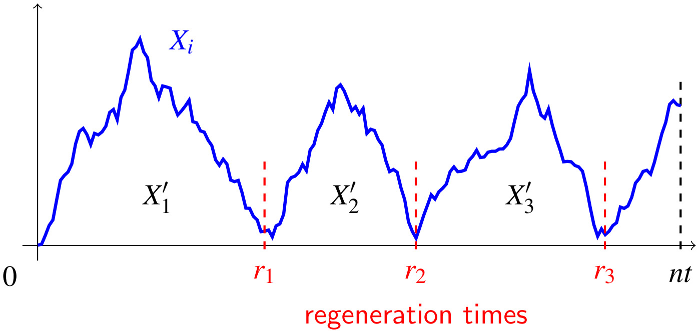
- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1-1}}_{\triangleq X_1'} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X_2'}$
  $+ \cdots + \underbrace{X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1}}_{\triangleq X'_{N(nt)}} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$



# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1-1}}_{\triangleq X'_1} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X'_2}$
$+ \cdots + \underbrace{X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1}}_{\triangleq X'_{N(nt)}} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$

$X'_i$'s are iid RV($\alpha$)

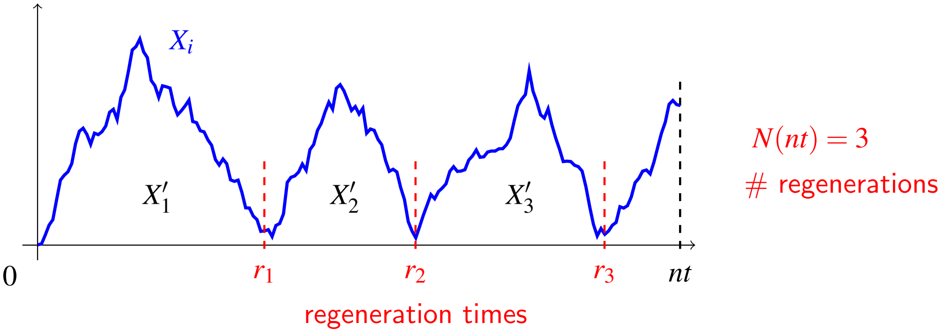



$N(nt) = 3$

\# regenerations

# View Regeneration Cycle as Heavy-Tailed Increment

- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1-1}}_{\triangleq X_1'} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X_2'}$
$$+ \cdots + \underbrace{X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1}}_{\triangleq X'_{N(nt)}} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$$
$X_i'$'s are iid RV($\alpha$)
- **A Key Step:** Sample Path LD for $\bar{S}_n'(t) = (X_1' + \cdots + X'_{N(nt)})/n$



# View Regeneration Cycle as Heavy-Tailed Increment

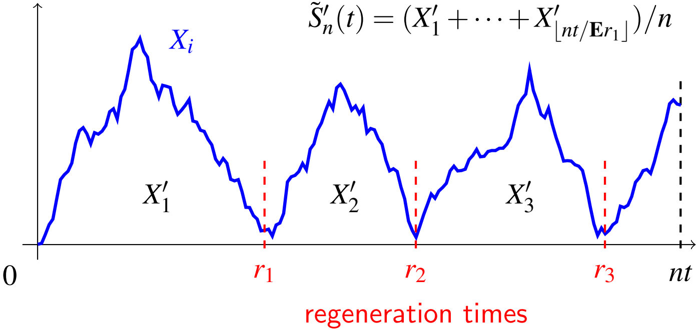
- $\bar{S}_n(t) = (\underbrace{X_0 + X_1 + \cdots + X_{r_1 - 1}}_{\triangleq X_1'} + \underbrace{X_{r_1} + X_{r_1+1} + \cdots + X_{r_2-1}}_{\triangleq X_2'}$
$+ \cdots + \underbrace{X_{r_{N(nt)-1}} + \cdots + X_{r_{N(nt)}-1}}_{\triangleq X_{N(nt)}'} + X_{r_{N(nt)}} + \cdots + X_{\lfloor nt \rfloor})/n$

$X_i'$'s are iid RV($\alpha$)

- **A Key Step:** Sample Path LD for $\bar{S}_n'(t) = (X_1' + \cdots + X_{N(nt)}')/n$

via "Asymptotic Equivalence" with

$\tilde{S}_n'(t) = (X_1' + \cdots + X_{\lfloor nt/\mathbf{E}r_1 \rfloor}')/n$




$N(nt) = 3$

\# regenerations

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- Our original definition of asymptotic equivalence:

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(d(\bar{S}'_n,\tilde{S}'_n) > \varepsilon\big) \;=\; 0, \quad \forall j \geq 0,$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Our original definition of asymptotic equivalence:

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) \;=\; 0, \quad \forall j \geq 0,$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

  and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Our original definition of asymptotic equivalence:

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) \;=\; 0, \quad \forall j \geq 0,$$

- No hope for this since

$$\mathbf{P}(X'_i > n) \sim n^{-\alpha}.$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Generalized asymptotic equivalence:

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\tilde{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n,\tilde{S}'_n) > \varepsilon\big) \ = \ 0, \quad \forall j \geq 0,$$

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\bar{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n,\tilde{S}'_n) > \varepsilon\big) \ = \ 0, \quad \forall j \geq 0,$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Generalized asymptotic equivalence:

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\tilde{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) \ = \ 0, \quad \forall j \geq 0,$$

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\bar{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) \ = \ 0, \quad \forall j \geq 0,$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Generalized asymptotic equivalence:

"$j$ big jumps"

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\Big(\tilde{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\Big) \;=\; 0, \quad \forall j \geq 0,$$

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\Big(\bar{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\Big) \;=\; 0, \quad \forall j \geq 0,$$

# Asymptotic Equivalence between $\bar{S}'_n$ and $\tilde{S}'_n$

- On the previous page:

$$\bar{S}'_n(t) = (X'_1 + \cdots + X'_{N(nt)})/n$$

  and

$$\tilde{S}'_n(t) = (X'_1 + \cdots + X'_{\lfloor nt/\mathbf{E}r_1 \rfloor})/n$$

- Generalized asymptotic equivalence:

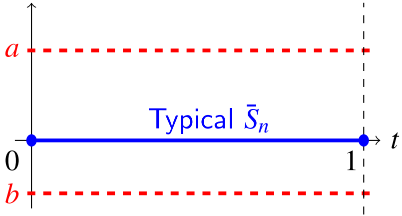
"$j$ big jumps"

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\tilde{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) = 0, \quad \forall j \geq 0,$$

$$\limsup_{n\to\infty} n^{(\alpha-1)j}\mathbf{P}\big(\bar{S}'_n \in (\mathbb{D}\setminus\mathbb{D}_{<j})^{-\gamma}, d(\bar{S}'_n, \tilde{S}'_n) > \varepsilon\big) = 0, \quad \forall j \geq 0,$$

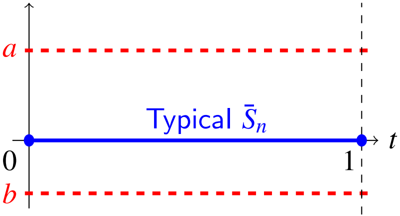
- Then $\tilde{S}'_n$ and $\bar{S}'_n$ are asymptotically equivalent

# Example: Barrier Option



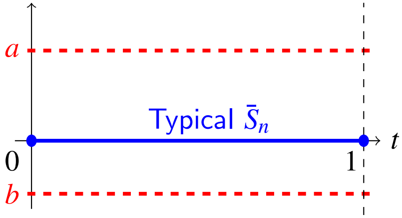
# Example: Barrier Option

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



# Example: Barrier Option

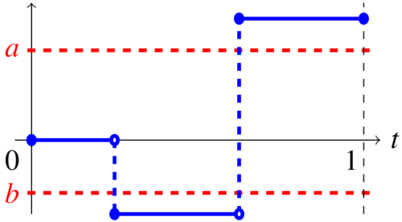
$$\bar{S}_n \in \{\, f \in \mathbb{D} : \ f \text{ hits below } b \text{ on } [0,1] \text{ and ends up above } a \,\} \triangleq E$$




**How do we expect this rare event arises?**

# Example: Barrier Option

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

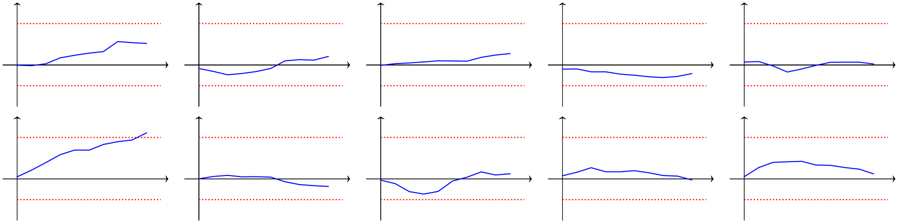



This way!

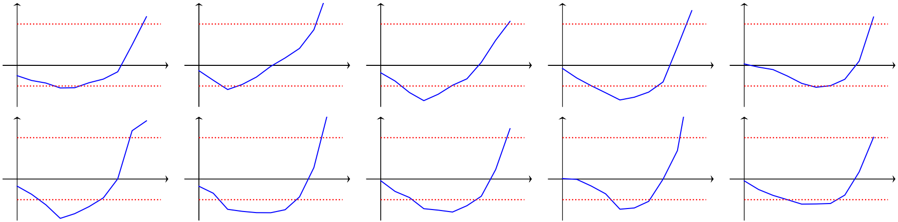
**How do we expect this rare event arises?**

# Example: Barrier Option

Typical sample paths of $\bar{S}_{10}$

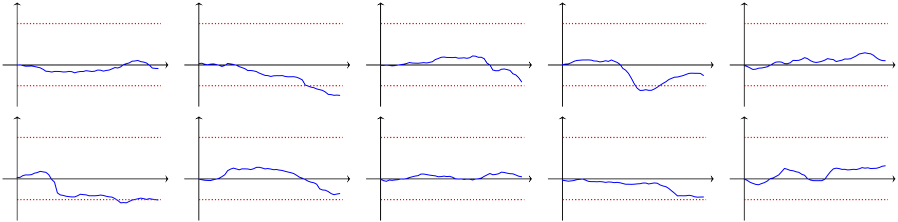


Sample paths of $\bar{S}_{10}$ conditional on $\bar{S}_n \in E$

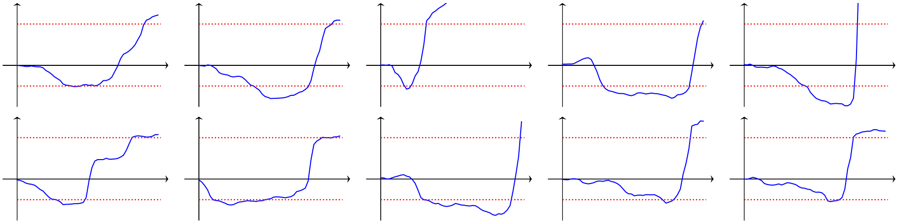


# Example: Barrier Option

Typical sample paths of $\bar{S}_{50}$

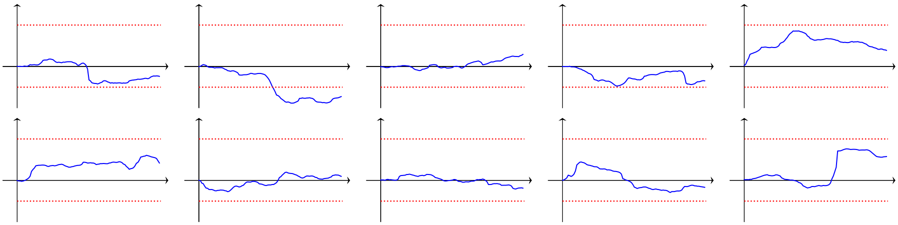


Sample paths of $\bar{S}_{50}$ conditional on $\bar{S}_n \in E$

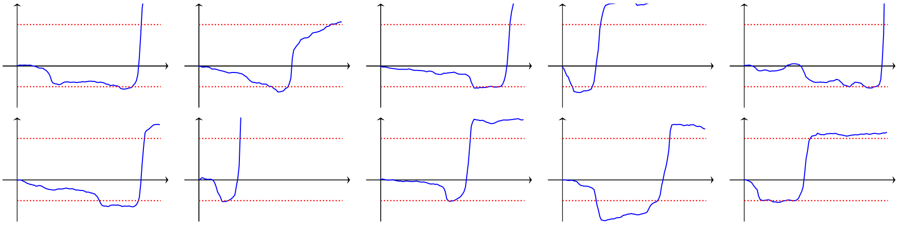


# Example: Barrier Option

Typical sample paths of $\bar{S}_{100}$

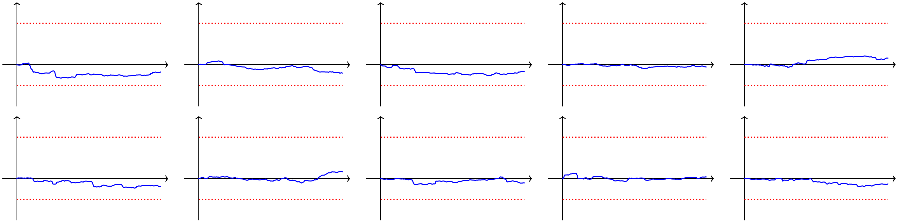


Sample paths of $\bar{S}_{100}$ conditional on $\bar{S}_n \in E$

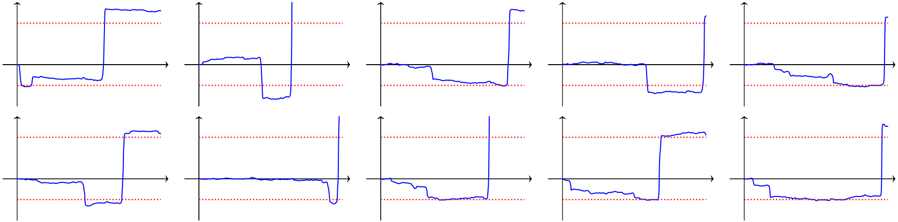


# Example: Barrier Option

Typical sample paths of $\bar{S}_{500}$

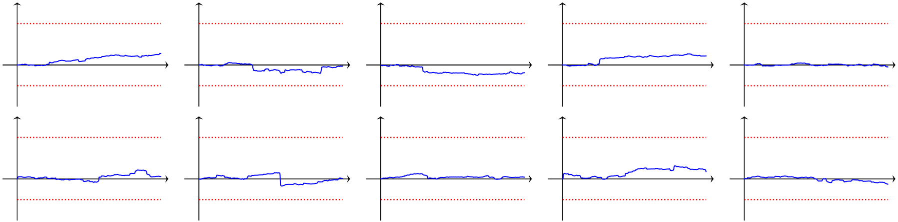


Sample paths of $\bar{S}_{500}$ conditional on $\bar{S}_n \in E$

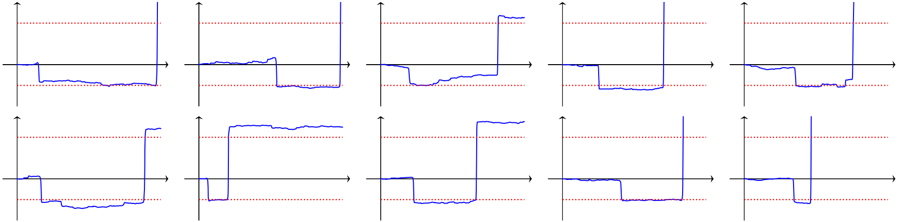


# Example: Barrier Option

Typical sample paths of $\bar{S}_{1000}$

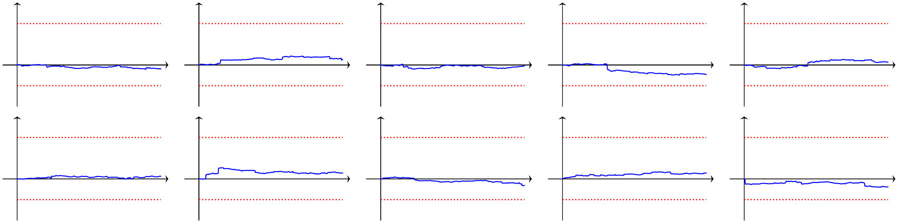


Sample paths of $\bar{S}_{1000}$ conditional on $\bar{S}_n \in E$

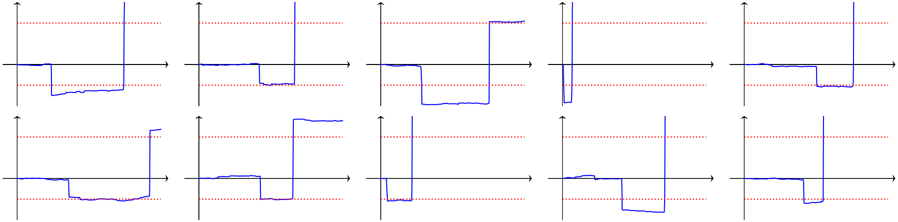


# Example: Barrier Option

Typical sample paths of $\bar{S}_{2500}$

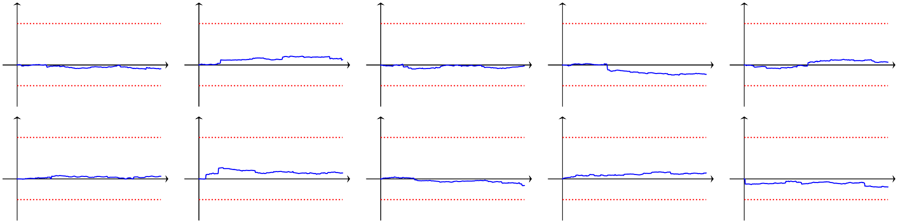


Sample paths of $\bar{S}_{2500}$ conditional on $\bar{S}_n \in E$

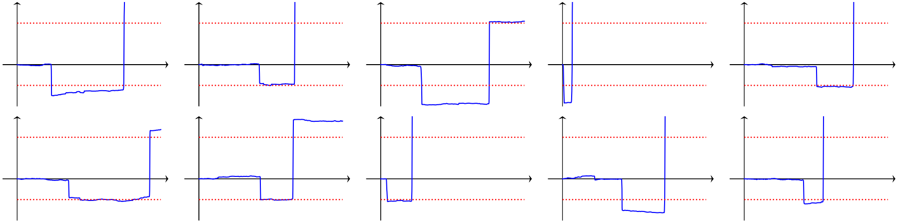


# Example: Barrier Option

Typical sample paths of $\bar{S}_{2500}$



Sample paths of $\bar{S}_{2500}$ conditional on $\bar{S}_n \in E$



# Other Topic II: Large Deviations for Weibull Tails

$$\mathbf{P}(X_i \geq x) = \exp(-x^{\alpha}),\ \ \alpha \in (0,1)$$

# What's already known: LDP w.r.t. $L_1$ topology

Nina Gantert (1998)

- $\limsup_{n\to\infty} \frac{1}{L(n)n^{\alpha}} \log \mathbf{P}\left(\bar{S}_n \in A\right) \leq -\inf_{\xi \in A^-} I(\xi)$, $A^-$: closure of $A$
- $\liminf_{n\to\infty} \frac{1}{L(n)n^{\alpha}} \log \mathbf{P}\left(\bar{S}_n \in A\right) \geq -\inf_{\xi \in A^\circ} I(\xi)$, $A^\circ$ : interior of $A$

# What's already known: LDP w.r.t. $L_1$ topology

Nina Gantert (1998)

- $\limsup_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(\bar{S}_n \in A\right) \le -\inf_{\xi\in A^-} I(\xi)$, $\qquad$ $A^-$: closure of $A$
- $\liminf_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(\bar{S}_n \in A\right) \ge -\inf_{\xi\in A^\circ} I(\xi)$, $\qquad$ $A^\circ$: interior of $A$
- $I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t-)\}} \left(\xi(t)-\xi(t-)\right)^\alpha & \text{if } \xi \text{ is a nondecreasing step function} \\ \infty & \text{otherwise} \end{cases}$

# What's already known: LDP w.r.t. $L_1$ topology

Nina Gantert (1998)

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha}I(A^{-})\right)$, $\quad I(A^{-}) = \inf_{\xi \in A^{-}} I(\xi)$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \gtrsim \exp\left(-L(n)n^{\alpha}I(A^{\circ})\right)$, $\quad I(A^{\circ}) = \inf_{\xi \in A^{\circ}} I(\xi)$
- $I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t-)\}} \left(\xi(t) - \xi(t-)\right)^{\alpha} & \text{if } \xi \text{ is a nondecreasing step function} \\ \infty & \text{otherwise} \end{cases}$

# What's already known: LDP w.r.t. $L_1$ topology

Nina Gantert (1998)

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^{-})\right)$, $\qquad I(A^{-}) = \inf_{\xi \in A^{-}} I(\xi)$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \gtrsim \exp\left(-L(n)n^{\alpha} I(A^{\circ})\right)$, $\qquad I(A^{\circ}) = \inf_{\xi \in A^{\circ}} I(\xi)$
- $I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t-)\}} \left(\xi(t) - \xi(t-)\right)^{\alpha} & \text{if } \xi \text{ is a nondecreasing step function} \\ \infty & \text{otherwise} \end{cases}$
- $I$ is a good rate function w.r.t. $L_1$ topology

  i.e., $d(\xi,\zeta) = \int_0^1 |\xi(s) - \zeta(s)| ds$

# What's already known: LDP w.r.t. $L_1$ topology

Nina Gantert (1998)

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha}I(A^-)\right)$, $\quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \gtrsim \exp\left(-L(n)n^{\alpha}I(A^\circ)\right)$, $\quad I(A^\circ) = \inf_{\xi \in A^\circ} I(\xi)$
- $I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t-)\}} \left(\xi(t)-\xi(t-)\right)^{\alpha} & \text{if } \xi \text{ is a nondecreasing step function} \\ \infty & \text{otherwise} \end{cases}$
- $I$ is a good rate function w.r.t. $L_1$ topology

  i.e., $d(\xi,\zeta) = \int_0^1 |\xi(s)-\zeta(s)|ds$

$\Rightarrow$ **Classical Tools in LD Theory Suffices**

# What's already known: LDP w.r.t. $L_1$ topology

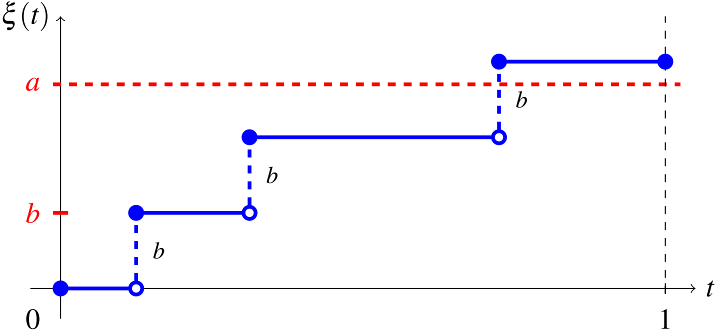
Nina Gantert (1998)

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha}I(A^{-})\right)$, $\quad I(A^{-}) = \inf_{\xi \in A^{-}} I(\xi)$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \gtrsim \exp\left(-L(n)n^{\alpha}I(A^{\circ})\right)$, $\quad I(A^{\circ}) = \inf_{\xi \in A^{\circ}} I(\xi)$
- $I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t-)\}} \left(\xi(t)-\xi(t-)\right)^{\alpha} & \text{if } \xi \text{ is a nondecreasing step function} \\ \infty & \text{otherwise} \end{cases}$
- $I$ is a good rate function w.r.t. $L_1$ topology

  i.e., $d(\xi,\zeta) = \int_0^1 |\xi(s)-\zeta(s)|ds$ $\quad\Rightarrow\quad$ **Classical Tools in LD Theory Suffices**

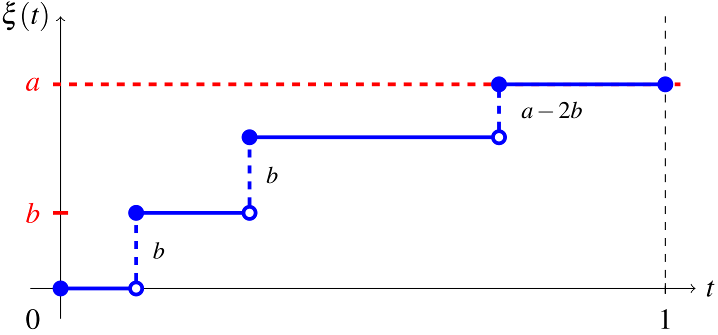
**BUT**

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + b^{\alpha}$

# $L_1$ topology is weak

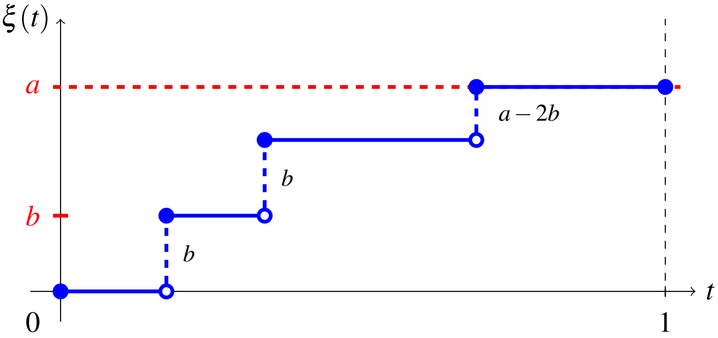



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

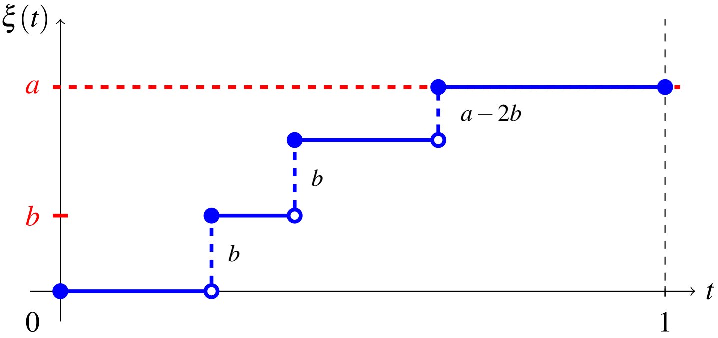



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$

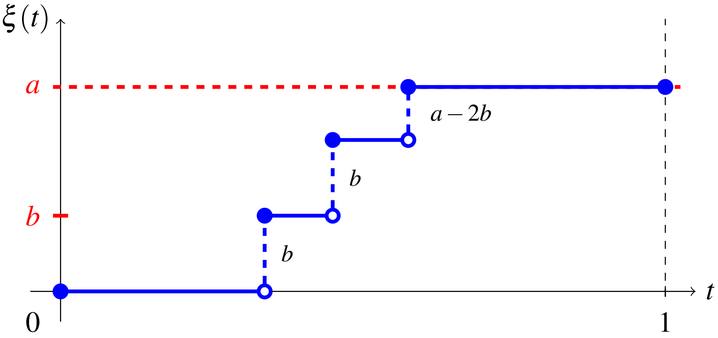
- $I(A^-) \overset{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

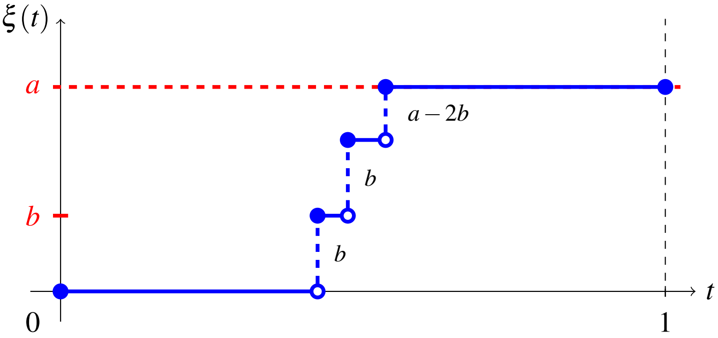



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

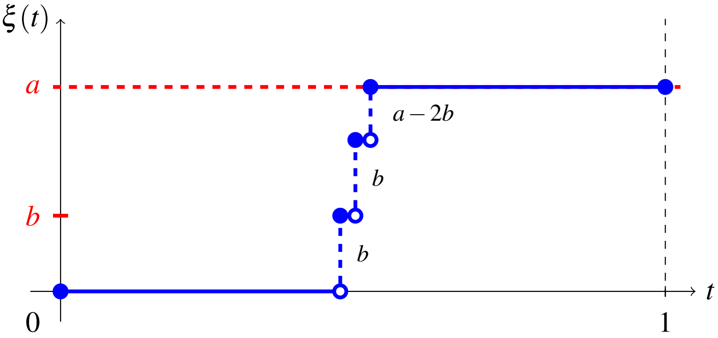



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

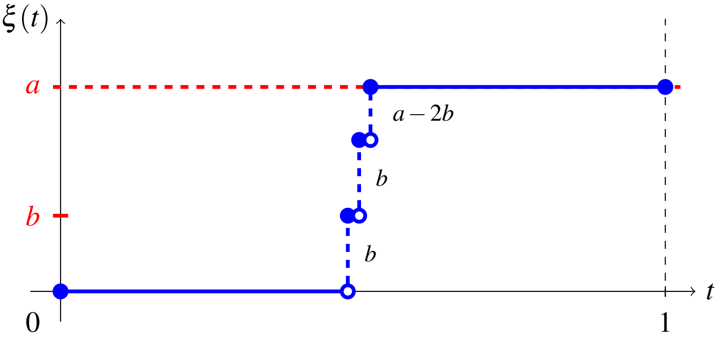
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

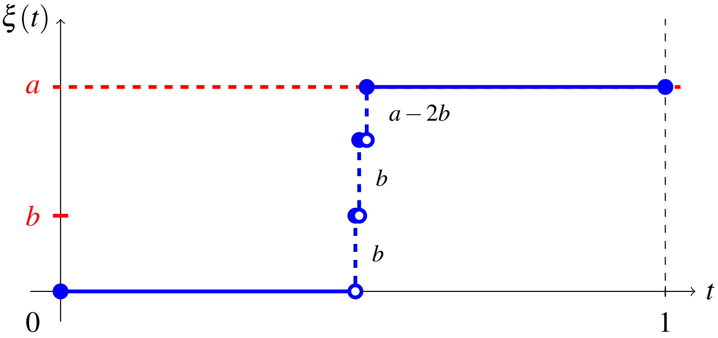



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

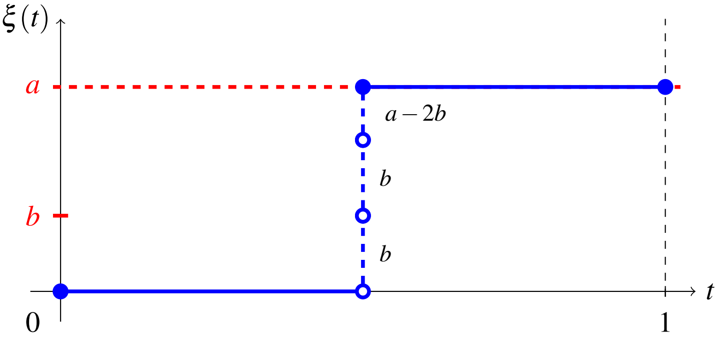
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

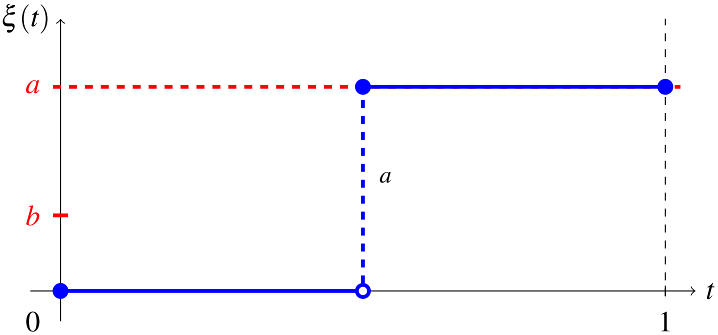



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^{-})\right),$ $\qquad I(A^{-}) = \inf_{\xi \in A^{-}} I(\xi)$

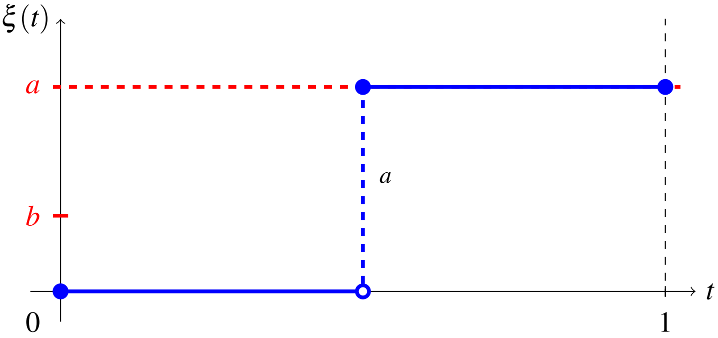
- $I(A^{-}) \overset{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak



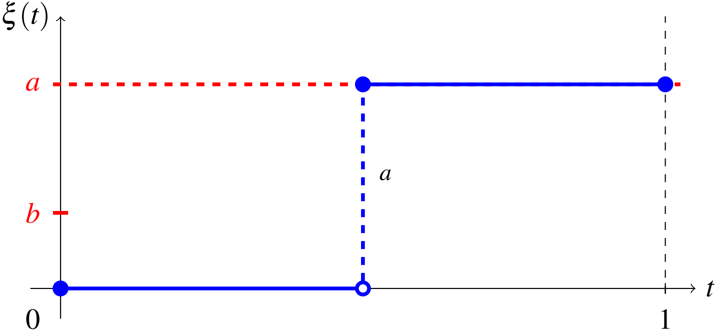

- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \neq b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

# $L_1$ topology is weak

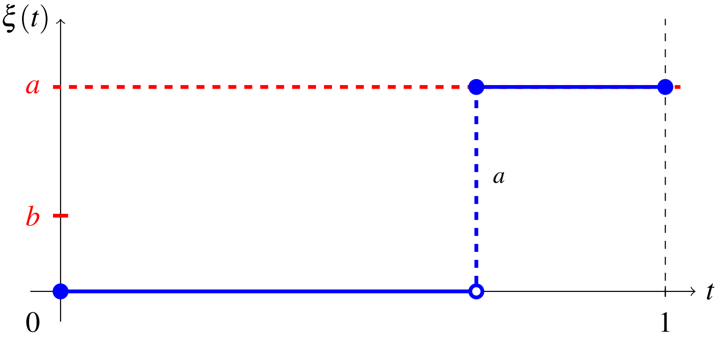



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi \in A^-} I(\xi)$

- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak

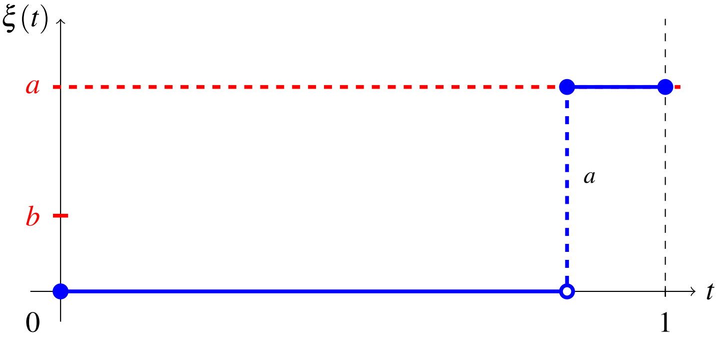



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \ \lesssim \ \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

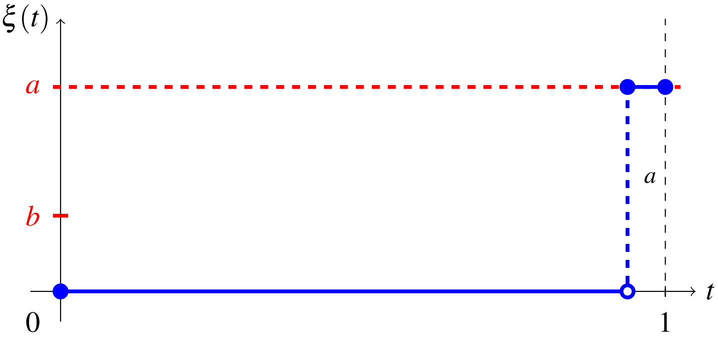
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak



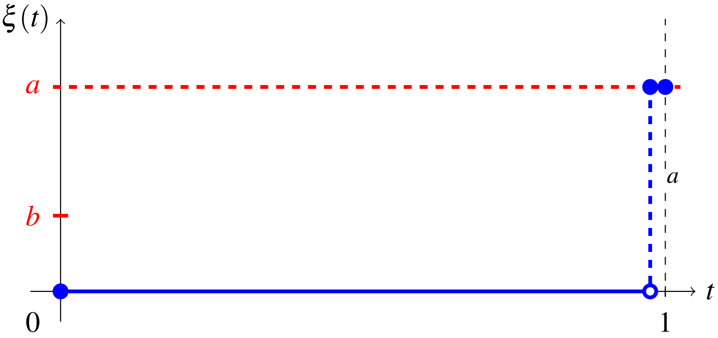

- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak

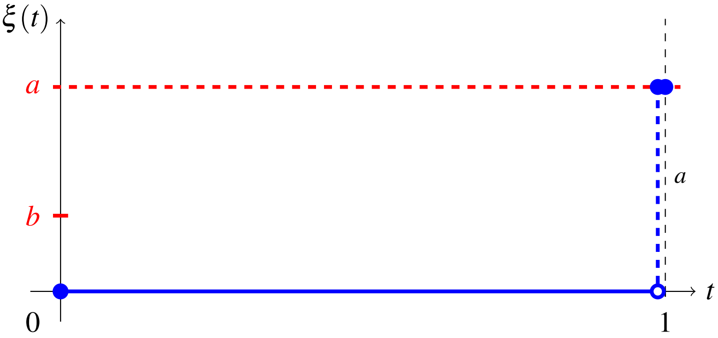



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

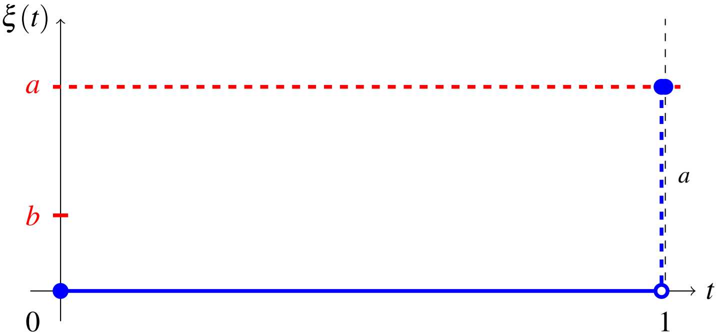



- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$

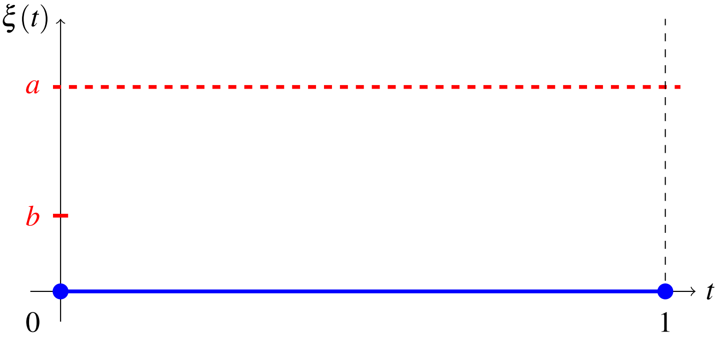
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

## $L_1$ topology is weak



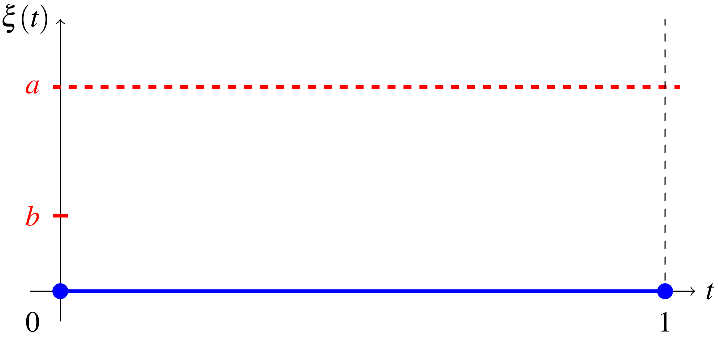

- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak



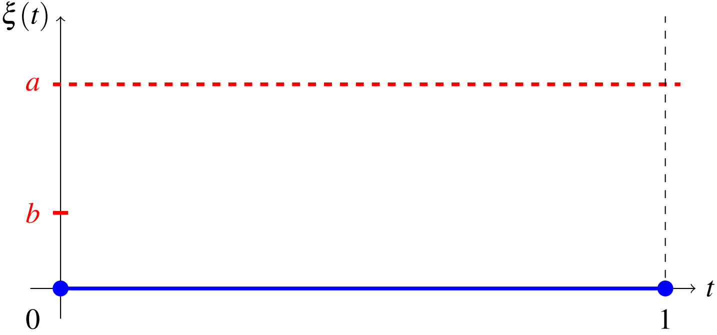

- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = 0 < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad 0 \in A^-$

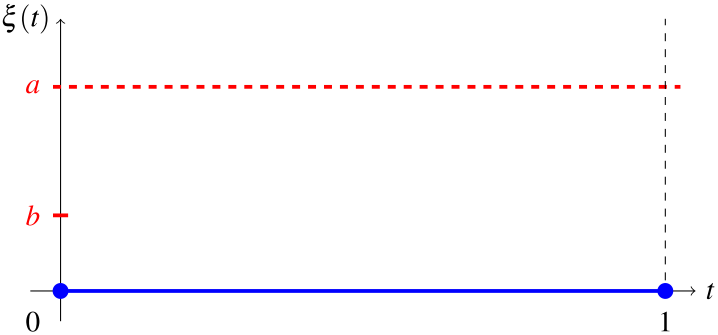
# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) = 0 < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad 0 \in A^-$

**No information!**

# $L_1$ topology is weak




- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^{-})\right)$, $\quad I(A^{-}) = \inf_{\xi \in A^{-}} I(\xi)$
- $I(A^{-}) = 0 < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad 0 \in A^{-}$

**Want a stronger topology, ideally $J_1$ topology!**

# LDP w.r.t. $J_1$ Topology is Impossible

**Counterexample (Bazhba, Blanchet, R., Zwart 2020)**

*There exists a closed set $A \subseteq \mathbb{D}$ such that*

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}} \log \mathbf{P}\left(\bar{S}_n \in A\right) \not\leq -\inf_{\xi\in A} I(\xi)$$

- $L(n) = 1$ and $\alpha = \frac{1}{2}$ so that $\mathbf{P}(X_1 > x) = e^{-\sqrt{x}}$
- Paths in $A$ have $m$ increases of size $O\left(\frac{1}{m^2}\right)$, for some $m$

# “Extended” Large Deviation Principle

**Theorem (Bazhba, Blanchet, R., Zwart 2020)**

$\bar{X}_n$ *satisfies an “extended LDP” w.r.t. $J_1$ topology, i.e.,*

$$\limsup_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(\bar{X}_n(t) \in A\right) \le -\lim_{\varepsilon\to 0} \inf_{\xi\in A^\varepsilon} I(\xi)$$

$$\liminf_{n\to\infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(\bar{X}_n(t) \in A\right) \ge -\inf_{\xi\in A^\circ} I(\xi)$$

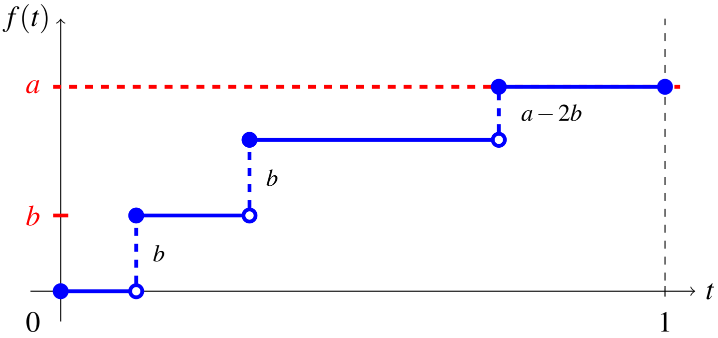
*where*

$$I(\xi) = \begin{cases} \sum_{\{t:\xi(t)\neq\xi(t^-)\}} \left(\xi(t)-\xi(t^-)\right)^\alpha, & \text{if } \xi \text{ is a nondecreasing pure jump function} \\ \infty, & o.w. \end{cases}$$

**Corollary**

*If $\phi$ is Lipschitz, $\phi\left(\bar{X}_n\right)$ satisfies a LDP, if the resulting rate function is good.*

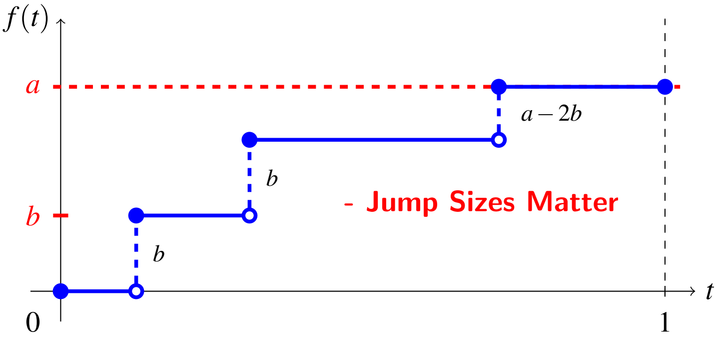
# Back to our reinsurance example




- $A = \{f \in \mathbb{D} : f$ crosses level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \;\sim\; \exp\left(-n^{\alpha} I^{*}\right)$

  where $I^{*} = b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$.

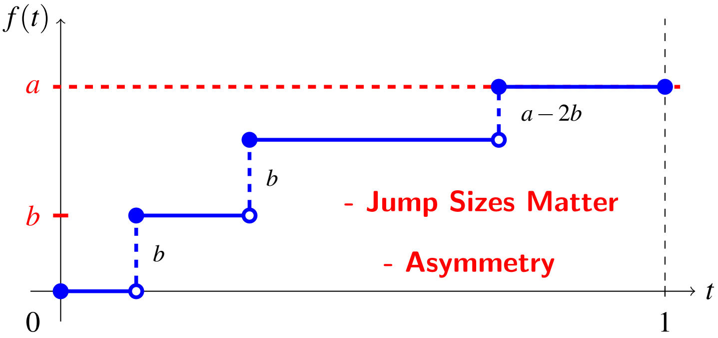
# Back to our reinsurance example




- $A = \{f \in \mathbb{D} : f$ crosses level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \sim \exp\left(-n^{\alpha} I^*\right)$

  where $I^* = b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$.

# Back to our reinsurance example

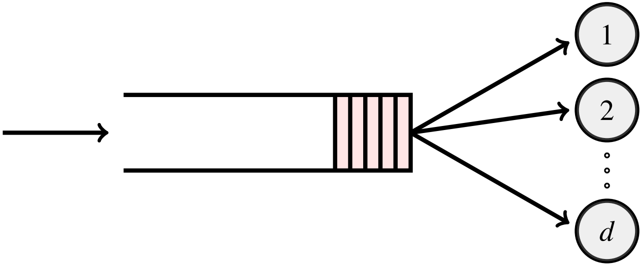



- $A = \{f \in \mathbb{D} : f$ crosses level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \;\sim\; \exp\left(-n^{\alpha} I^{*}\right)$

  where $I^{*} = b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$.

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

In case service time distribution is Weibull, i.e., $\mathbf{P}(S > x) \sim \exp(-x^{\alpha})$,

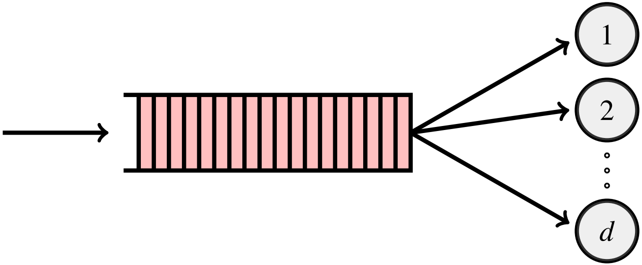



**Tail Asymptotics? Most likely scenario?**

**Previously, NOT EVEN a reasonable conjecture was available!**

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

In case service time distribution is Weibull, i.e., $\mathbf{P}(S > x) \sim \exp(-x^{\alpha})$,




**Tail Asymptotics? Most likely scenario?**

**Previously, NOT EVEN a reasonable conjecture was available!**

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^{\alpha})$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^{\alpha}\right)$$

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^{\alpha})$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^{\alpha}\right)$$

where

$$\begin{aligned} c^* = \quad & \min \sum_{i=1}^{d} x_i^{\alpha} \quad \text{s.t.} \\ & \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma], \\ & x_1, \ldots, x_d \geq 0 \,. \end{aligned}$$

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^\alpha)$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^\alpha\right)$$

where

$$\begin{aligned} c^* = \quad & \min \sum_{i=1}^{d} x_i^\alpha \quad \text{s.t.} \\ & \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma], \\ & x_1, \ldots, x_d \geq 0\,. \end{aligned}$$

- Special case of $L^\alpha$-norm minimization problem.
- We have explicit solution.

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^\alpha)$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^\alpha\right)$$

where

**- Nontrivial Tradeoff between Size and Number of Jumps!**

$$\begin{aligned} c^* = \quad & \min \sum_{i=1}^{d} x_i^\alpha \quad \text{s.t.} \\ & \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma], \\ & x_1, \ldots, x_d \geq 0 \, . \end{aligned}$$

- Special case of $L^\alpha$-norm minimization problem.
- We have explicit solution.

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^\alpha)$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^\alpha\right)$$

where

**- Nontrivial Tradeoff between Size and Number of Jumps!**

**- Asymmetry in Jump Sizes!**

$$\begin{aligned} c^* = \quad & \min \sum_{i=1}^{d} x_i^\alpha \quad \text{s.t.} \\ & \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma], \\ & x_1, \ldots, x_d \geq 0 \, . \end{aligned}$$

- Special case of $L^\alpha$-norm minimization problem.
- We have explicit solution.

# Queue Length Asymptotics for the Weibull GI/GI/d Queue

If service time distribution is Weibull, i.e., $\mathbf{P}(S_1 > x) \sim \exp(-x^\alpha)$, then

$$\mathbf{P}(Q(\gamma n) > n) \sim \exp\left(-c^* n^\alpha\right)$$

where

**- Nontrivial Tradeoff between Size and Number of Jumps!**

**- Asymmetry in Jump Sizes!**

$$\begin{aligned} c^* = \quad & \min \sum_{i=1}^{d} x_i^\alpha \quad \text{s.t.} \\ & \lambda s - \sum_{i=1}^{d} (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma], \\ & x_1, \ldots, x_d \geq 0\,. \end{aligned}$$

- Special case of $L^\alpha$-norm minimization problem.
- We have explicit solution.

**Solution to Open Question Posed by Whitt (2000) and Foss (2009)**

# More Specifically,

If $\gamma > \frac{1}{\lambda - \lfloor \lambda \rfloor}$ : job/jump sizes (of the most likely scenario) are **symmetric**

- If $\lfloor \lambda \rfloor \leq \frac{\lambda - \alpha d}{1 - \alpha}$: smallest possible number of big jobs to block enough servers
  $\Rightarrow$ **same as the power law case**

- If $\lfloor \lambda \rfloor > \frac{\lambda - \alpha d}{1 - \alpha}$: larger number of moderately big jobs might be more likely
  $\Rightarrow$ **qualitatively different from the power law case**

If $\gamma < \frac{1}{\lambda - \lfloor \lambda \rfloor}$ : job/jump sizes may be **asymmetric** (upto 3 different sizes)

# Other Topic III: Rare Event Simulation

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Head})$

- Flip the coin 100 times
- Count the number of head
- Divide by 100 and report the number

Should be reasonably close to $1/2$

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge})$

- Flip the coin 100 times
- Count the number of Edge
- Divide by 100 and report the number

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge})$

- Flip the coin 100 times
- Count the number of Edge : most likely to be 0
- Divide by 100 and report the number

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge})$

- Flip the coin 100 times
- Count the number of Edge : most likely to be 0
- Divide by 100 and report the number : and hence, likely to be 0

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge})$

- Flip the coin 100 times
- Count the number of Edge : most likely to be 0
- Divide by 100 and report the number : and hence, likely to be 0

**Is 0 a useful answer?**

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge})$

- Flip the coin 100 times
- Count the number of Edge : most likely to be 0
- Divide by 100 and report the number : and hence, likely to be 0

**Is 0 a useful answer? No.**

**e.g., Nuclear Meltdown, Large-Scale Blackout, Large Financial Loss**

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge}) \overset{\textit{Suppose}}{\approx} 10^{-6}$

- Flip the coin 100 times
- Count the number of Edge
- Divide by 100 and report the number

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge}) \overset{\textit{Suppose}}{\approx} 10^{-6}$

- Flip the coin ~~100~~ a few million times
- Count the number of Edge
- Divide by the total number of flips and report the number

# Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

e.g. Coin flip: want to estimate $\mathbf{P}(\text{Edge}) \overset{\textit{Suppose}}{\approx} 10^{-6}$

- Flip the coin ~~100~~ a few million times
- Count the number of Edge
- Divide by the total number of flips and report the number

**Much harder than $\mathbf{P}$(Head)**

# Importance Sampling

- Construct an alternative universe

# Importance Sampling

- Construct an alternative universe

- Perform experiments there

# Importance Sampling

- Construct an alternative universe
- Perform experiments there
- Recover the true probability using the relationship between the two parallel universes

# Importance Sampling

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}$

- Perform experiments there

- Recover the true probability using the relationship between the two parallel universes

# Importance Sampling

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}$

- Perform experiments there
  - i.e., Sample $\mathbb{I}^{(1)}_{\text{Edge}}, \ldots, \mathbb{I}^{(m)}_{\text{Edge}}$ from $\mathbf{Q}$

- Recover the true probability using the relationship between the two parallel universes

# Importance Sampling

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}$
- Perform experiments there
  - i.e., Sample $\mathbb{I}_{\text{Edge}}^{(1)}, \ldots, \mathbb{I}_{\text{Edge}}^{(m)}$ from $\mathbf{Q}$
    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes

# Importance Sampling

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}$
- Perform experiments there
  - i.e., Sample $\mathbb{I}^{(1)}_{\text{Edge}}, \ldots, \mathbb{I}^{(m)}_{\text{Edge}}$ from $\mathbf{Q}$

    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report $\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}^{(i)}_{\text{Edge}} \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(i)}$ as an estimate of $\mathbf{P}(\text{Edge})$

# Importance Sampling for $\mathbf{P}(\bar{S}_n \in A)$

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}$

- Perform experiments there
  - i.e., Sample $\mathbb{I}_{\text{Edge}}^{(1)}, \ldots, \mathbb{I}_{\text{Edge}}^{(m)}$ from $\mathbf{Q}$

    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(m)}$

- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report $\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\text{Edge}}^{(i)} \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(i)}$ as an estimate of $\mathbf{P}(\text{Edge})$

# Importance Sampling for $\mathbf{P}(\bar{S}_n \in A)$

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}_n$

- Perform experiments there
  - i.e., Sample $\mathbb{I}^{(1)}_{\{\bar{S}_n \in A\}}, \ldots, \mathbb{I}^{(m)}_{\{\bar{S}_n \in A\}}$ from $\mathbf{Q}_n$

    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(m)}$

- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report $\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}^{(i)}_{\{\bar{S}_n \in A\}} \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(i)}$ as an estimate of $\mathbf{P}(\bar{S}_n \in A)$

# Importance Sampling for $\mathbf{P}(\bar{S}_n \in A)$

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}_n$

- Perform experiments there
  - i.e., Sample $\mathbb{I}^{(1)}_{\{\bar{S}_n \in A\}}, \ldots, \mathbb{I}^{(m)}_{\{\bar{S}_n \in A\}}$ from $\mathbf{Q}_n$

    as well as the likelihood ratio $\left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(1)}, \ldots, \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(m)}$

- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report $\frac{1}{m}\sum_{i=1}^{m} \underbrace{\mathbb{I}^{(i)}_{\{\bar{S}_n \in A\}} \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(i)}}_{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n} \text{ : IS estimator}}$ as an estimate of $\mathbf{P}(\bar{S}_n \in A)$

# Importance Sampling for $\mathbf{P}(\bar{S}_n \in A)$

- Construct an alternative universe
  - i.e., Consider an "importance distribution" $\mathbf{Q}_n$

- Perform experiments there
  - i.e., Sample $\mathbb{I}^{(1)}_{\{\bar{S}_n \in A\}}, \ldots, \mathbb{I}^{(m)}_{\{\bar{S}_n \in A\}}$ from $\mathbf{Q}_n$

    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(m)}$

- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report $\frac{1}{m}\sum_{i=1}^{m} \underbrace{\mathbb{I}^{(i)}_{\{\bar{S}_n \in A\}} \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^{(i)}}_{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n} \text{ : IS estimator}}$ as an estimate of $\mathbf{P}(\bar{S}_n \in A)$

**Finding a good alternative universe $\mathbf{Q}_n$ is crucial.**

# What is a good alternate universe $\mathbf{Q}_n$ for $\mathbf{P}(\bar{S}_n \in A)$?

General principle for making $\overbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ an efficient estimator:

# What is a good alternate universe $\mathbf{Q}_n$ for $\mathbf{P}(\bar{S}_n \in A)$?

General principle for making $\overbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ an efficient estimator:

- **Choose $\mathbf{Q}_n(\cdot)$ as close to $\mathbf{P}(\cdot|\bar{S}_n \in A)$ as possible.**

# What is a good alternate universe $\mathbf{Q}_n$ for $\mathbf{P}(\bar{S}_n \in A)$?

General principle for making $\overbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ an efficient estimator:

- **Choose $\mathbf{Q}_n(\cdot)$ as close to $\mathbf{P}(\cdot | \bar{S}_n \in A)$ as possible.**

- **Make sure that $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}$ does not blow up.**

# Goal: Strongly Efficient IS Estimator

$\underbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ is a **strongly efficient** estimator for $\mathbf{P}(\bar{S}_n \in A)$, if

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \quad \sim \quad \mathbf{P}(\bar{S}_n \in A)^2$$

# Goal: Strongly Efficient IS Estimator

$\underbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ is a **strongly efficient** estimator for $\mathbf{P}(\bar{S}_n \in A)$, if

$$\overbrace{\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2}^{\text{Error}^2} \quad \sim \quad \mathbf{P}(\bar{S}_n \in A)^2$$

# Goal: Strongly Efficient IS Estimator

$\underbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}_{\text{IS estimator}}$ is a **strongly efficient** estimator for $\mathbf{P}(\bar{S}_n \in A)$, if

$$\underbrace{\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2}_{\text{Error}^2} \quad \sim \quad \underbrace{\mathbf{P}(\bar{S}_n \in A)^2}_{\text{Target Quantity}^2}$$

# Goal: Strongly Efficient IS Estimator

$\underbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}}^{\text{IS estimator}}$ is a **strongly efficient** estimator for $\mathbf{P}(\bar{S}_n \in A)$, if

$$\overbrace{\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2}^{\text{Error}^2} \quad \sim \quad \overbrace{\mathbf{P}(\bar{S}_n \in A)^2}^{\text{Target Quantity}^2}$$

$\Rightarrow$ Number of simulation runs required remains bounded.

# Goal: Strongly Efficient IS Estimator

$\underbrace{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}}_{\text{IS estimator}}$ is a **strongly efficient** estimator for $\mathbf{P}(\bar{S}_n \in A)$, if

$$\underbrace{\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2}_{\text{Error}^2} \sim \underbrace{\mathbf{P}(\bar{S}_n \in A)^2}_{\text{Target Quantity}^2}$$

$\Rightarrow$ Number of simulation runs required remains bounded.

**Considered Notoriously Hard for Heavy-Tailed Processes.**

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\,\bar{X}_n \in B^{\gamma})$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- Fix $w \in (0,1)$ and define

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n} \text{ not too big}$$

$$\uparrow$$

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- Fix $w \in (0,1)$ and define

$\frac{d\mathbf{P}}{d\mathbf{Q}_n}$ not too big

$\mathbf{Q}_n(\cdot)$ close to $\mathbf{P}(\cdot|\bar{X}_n \in A)$

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\,\bar{X}_n \in B^\gamma)$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $> \gamma\}$

- Fix $w \in (0,1)$ and define

$\frac{d\mathbf{P}}{d\mathbf{Q}_n}$ not too big

$\mathbf{Q}_n(\cdot)$ close to $\mathbf{P}(\cdot|\bar{X}_n \in A)$

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\,\bar{X}_n \in B^\gamma)$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $> \gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$\frac{d\mathbf{P}}{d\mathbf{Q}_n}$ not too big

$\mathbf{Q}_n(\cdot)$ close to $\mathbf{P}(\cdot|\bar{X}_n \in A)$

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\,\bar{X}_n \in B^\gamma)$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\} \Rightarrow \mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$\frac{d\mathbf{P}}{d\mathbf{Q}_n}$ not too big

$\mathbf{Q}_n(\cdot)$ close to $\mathbf{P}(\cdot|\bar{X}_n \in A)$

$$\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot\,|\,\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $> \gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}$ not too big

$\mathbf{Q}_n(\cdot)$ close to $\mathbf{P}(\cdot|\bar{X}_n \in A)$

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\,\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A)\cdot\mathbf{P}(\bar{X}_n \in B^\gamma)$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq$ {paths w/ at least $\mathcal{J}(A)$ jumps of size $>\gamma$} $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}^{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}} + \overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^\gamma)}^{n^{-\alpha\mathcal{J}(B^\gamma)}}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $>\gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}^{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}} + \overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^\gamma)}^{n^{-\alpha\mathcal{J}(B^\gamma)}}$$

## First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^{\gamma} \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^{\gamma})$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^{\gamma})$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^{\gamma}) \geq \mathcal{J}(A) + \mathcal{J}(B^{\gamma})$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^{\gamma})}^{n^{-\alpha\mathcal{J}(A \setminus B^{\gamma})}} + \underbrace{\overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^{\gamma})}^{n^{-\alpha\mathcal{J}(B^{\gamma})}}}_{\sim \mathbf{P}(\bar{X}_n \in A)^2}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $>\gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot\,|\,\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}^{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}} + \underbrace{\overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^\gamma)}^{n^{-\alpha\mathcal{J}(B^\gamma)}}}_{\sim \mathbf{P}(\bar{X}_n \in A)^2}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\} \;\Rightarrow\; \mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}^{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}} + \underbrace{\overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^\gamma)}^{n^{-\alpha\mathcal{J}(B^\gamma)}}}_{\sim\, \mathbf{P}(\bar{X}_n \in A)^2}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $> \gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$
- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\underbrace{\overset{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}}{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}}_{O(\mathbf{P}(\bar{X}_n \in A)^2)} + \underbrace{\overset{n^{-\alpha\mathcal{J}(A)}}{\mathbf{P}(\bar{X}_n \in A)} \cdot \overset{n^{-\alpha\mathcal{J}(B^\gamma)}}{\mathbf{P}(\bar{X}_n \in B^\gamma)}}_{\sim\, \mathbf{P}(\bar{X}_n \in A)^2}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\} \;\Rightarrow\; \mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w} \underbrace{\overbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}^{n^{-\alpha\mathcal{J}(A\setminus B^\gamma)}}}_{O(\mathbf{P}(\bar{X}_n \in A)^2)} + \underbrace{\overbrace{\mathbf{P}(\bar{X}_n \in A)}^{n^{-\alpha\mathcal{J}(A)}} \cdot \overbrace{\mathbf{P}(\bar{X}_n \in B^\gamma)}^{n^{-\alpha\mathcal{J}(B^\gamma)}}}_{\sim \mathbf{P}(\bar{X}_n \in A)^2}$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $>\gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\underbrace{\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma)}_{O(\mathbf{P}(\bar{X}_n \in A)^2)} + \underbrace{\mathbf{P}(\bar{X}_n \in A)\cdot\mathbf{P}(\bar{X}_n \in B^\gamma)}_{\sim\,\mathbf{P}(\bar{X}_n \in A)^2}$$

$$\sim \big(\mathbf{P}(\bar{X}_n \in A)\big)^2$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $>\gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

$$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A)\cdot\mathbf{P}(\bar{X}_n \in B^\gamma)$$

$$\sim \big(\mathbf{P}(\bar{X}_n \in A)\big)^2$$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\} \;\Rightarrow\; \mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\,\cdot\,) \triangleq w\mathbf{P}(\,\cdot\,) + (1-w)\mathbf{P}(\,\cdot\,|\bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

Error$^2$ $\rightarrow$ $$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A) \cdot \mathbf{P}(\bar{X}_n \in B^\gamma)$$

$$\sim \big(\mathbf{P}(\bar{X}_n \in A)\big)^2$$ $\leftarrow$ Target Quantity$^2$

# First Universal Rare-Event Simulation Scheme for Heavy-Tails

- $B^\gamma \triangleq \{$paths w/ at least $\mathcal{J}(A)$ jumps of size $> \gamma\}$ $\Rightarrow$ $\mathcal{J}(A) = \mathcal{J}(B^\gamma)$

- Fix $w \in (0,1)$ and define

$$\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot \,|\, \bar{X}_n \in B^\gamma)$$

- If we choose $\gamma$ so that $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

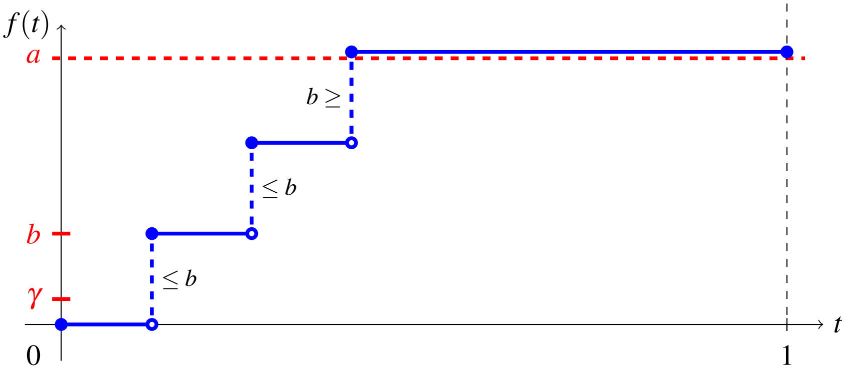
Error$^2$ $\rightarrow$ $$\mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A) \cdot \mathbf{P}(\bar{X}_n \in B^\gamma)$$

$$\sim \big(\mathbf{P}(\bar{X}_n \in A)\big)^2$$ $\leftarrow$ Target Quantity$^2$

$Z_n \triangleq \mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}$ **is strongly efficient for** $\mathbf{P}(\bar{X}_n \in A)$**!**

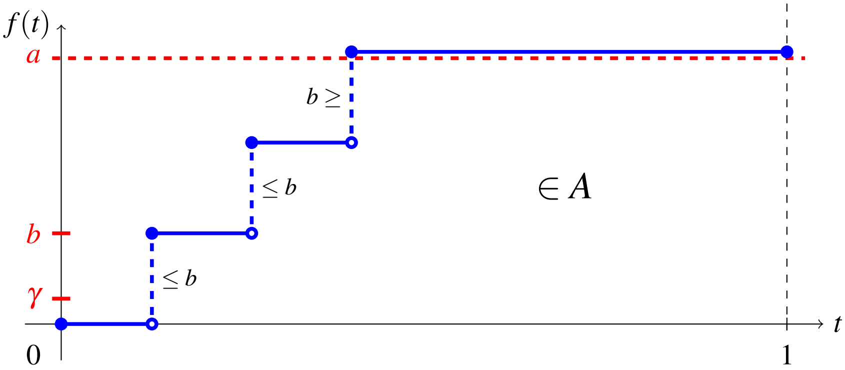
Chen, Blanchet, R., and Zwart (2019)

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



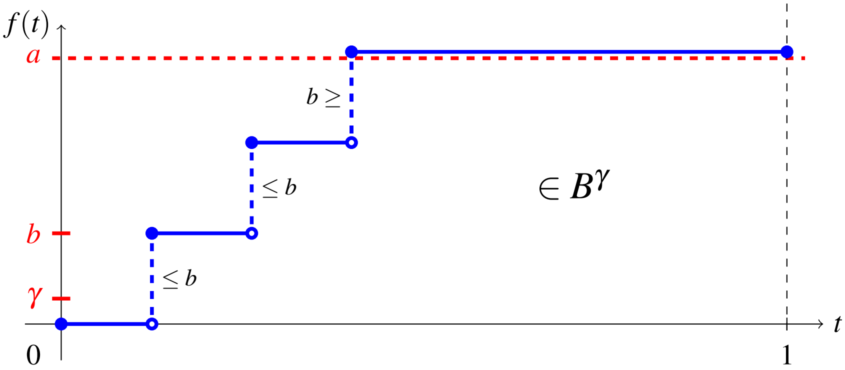

- $A$ $= \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma$ $= \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



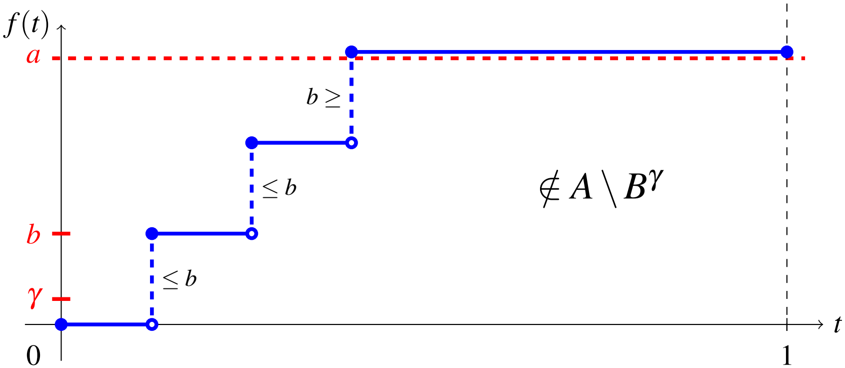

- $A \quad = \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



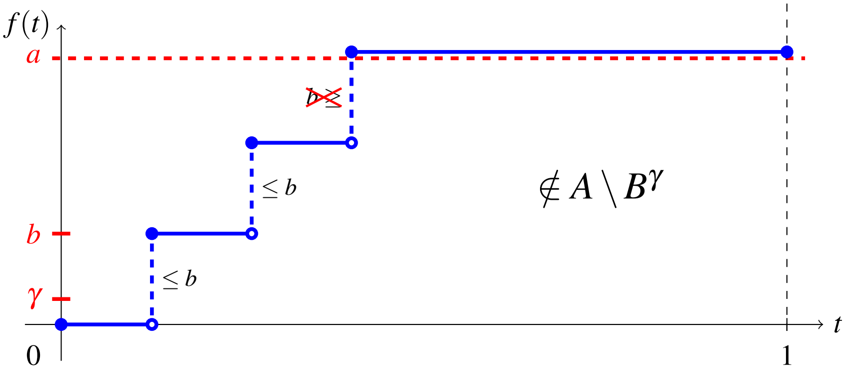

- $A \quad = \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



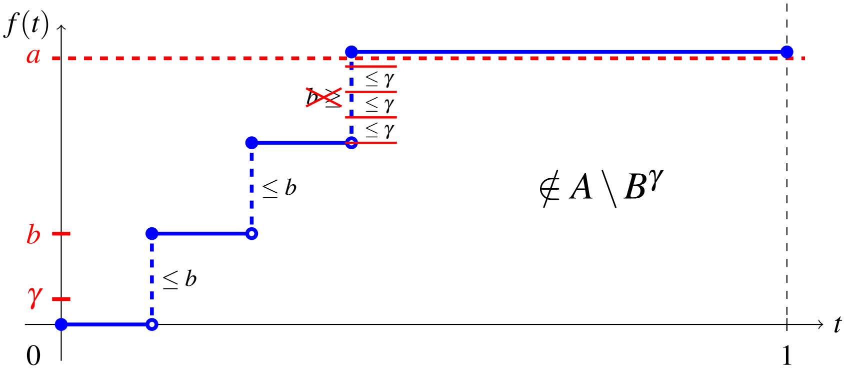

- $A$ $= \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma$ $= \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



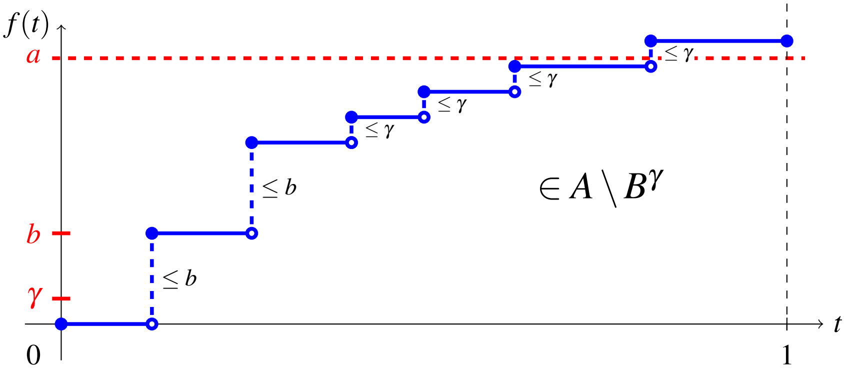

- $A \quad = \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



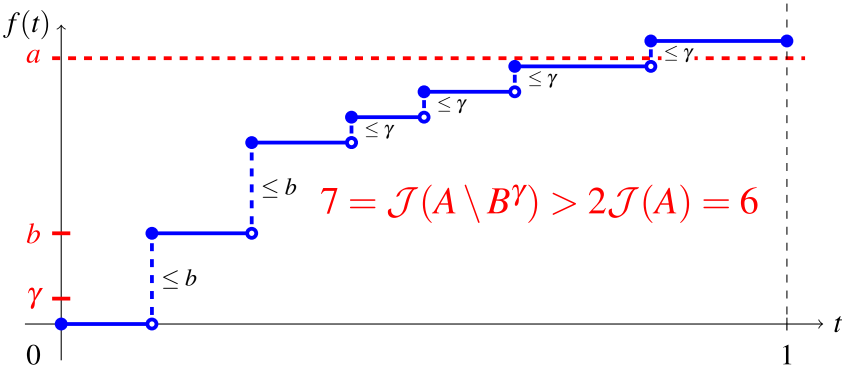

- $A$ $= \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma$ $= \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$ & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example



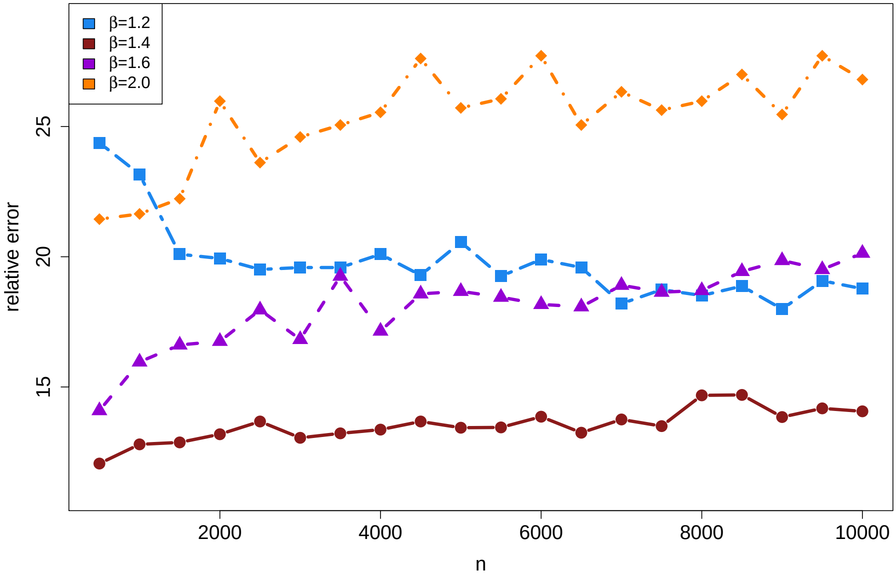

- $A$ $= \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma$ $= \{$paths with at least $3$ jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most $2$ jumps of size $> \gamma\}$

# How to ensure $\mathcal{J}(A \setminus B^\gamma) \geq 2\mathcal{J}(A)$: Reinsurance Example

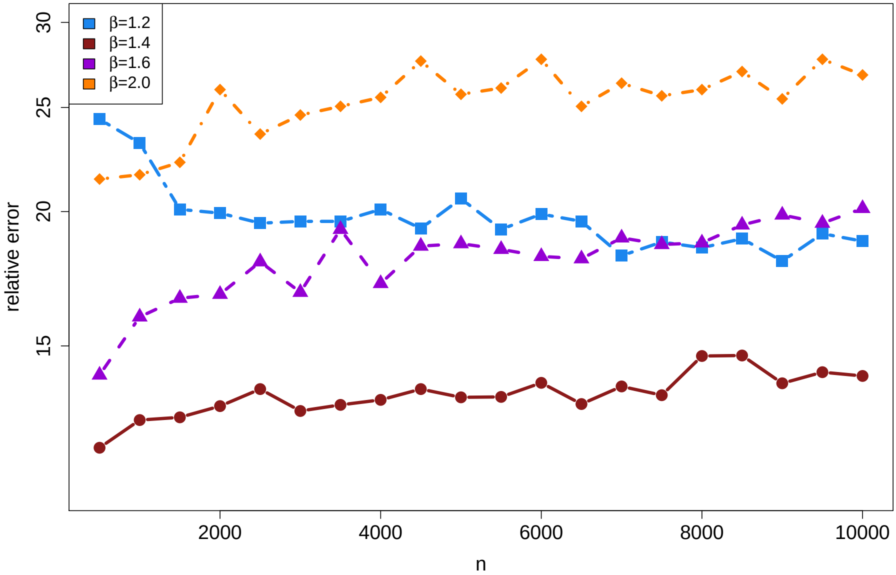



- $A \quad = \{$paths that cross level $a$ on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least $3$ jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level $a$ on $[0,1]$
  & all jump sizes $\leq b$ & at most $2$ jumps of size $> \gamma\}$

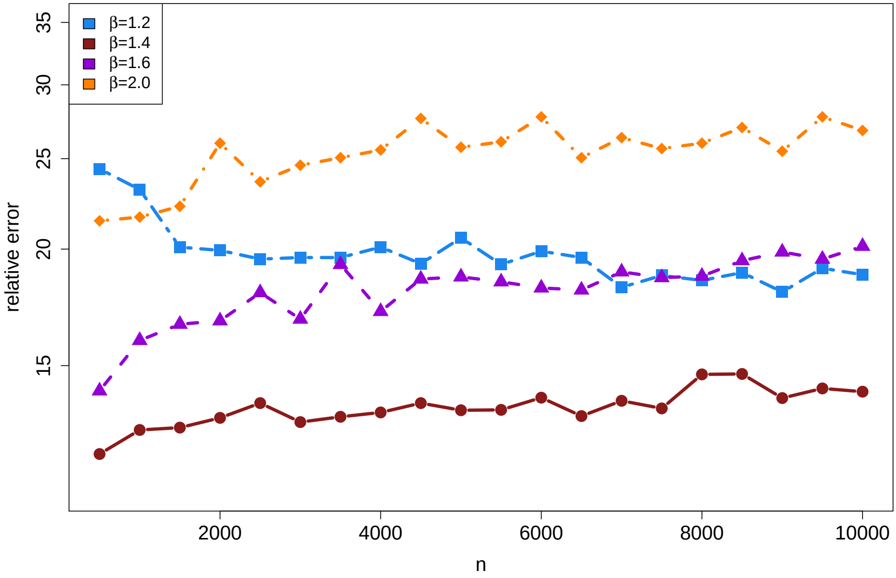
# Numerical Experiments for Reinsurance Example



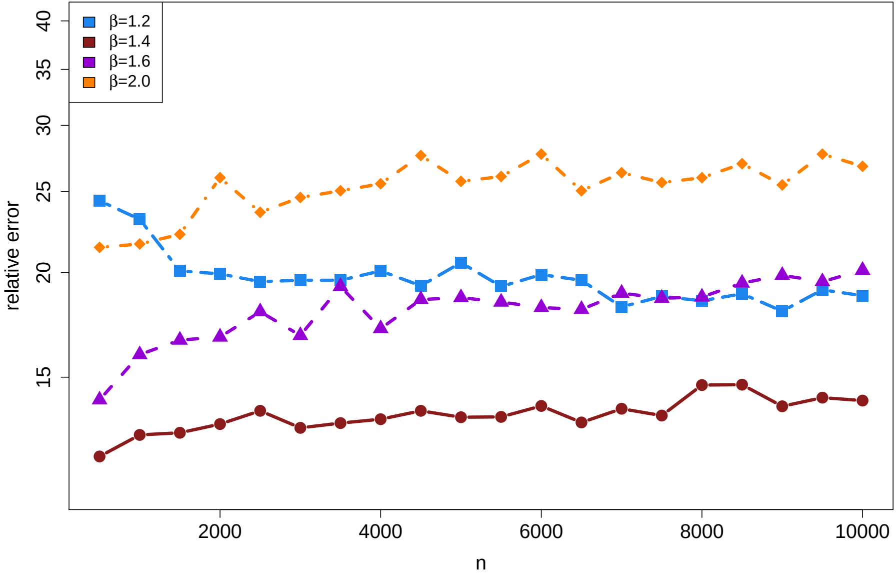
# Numerical Results for Reinsurance Example



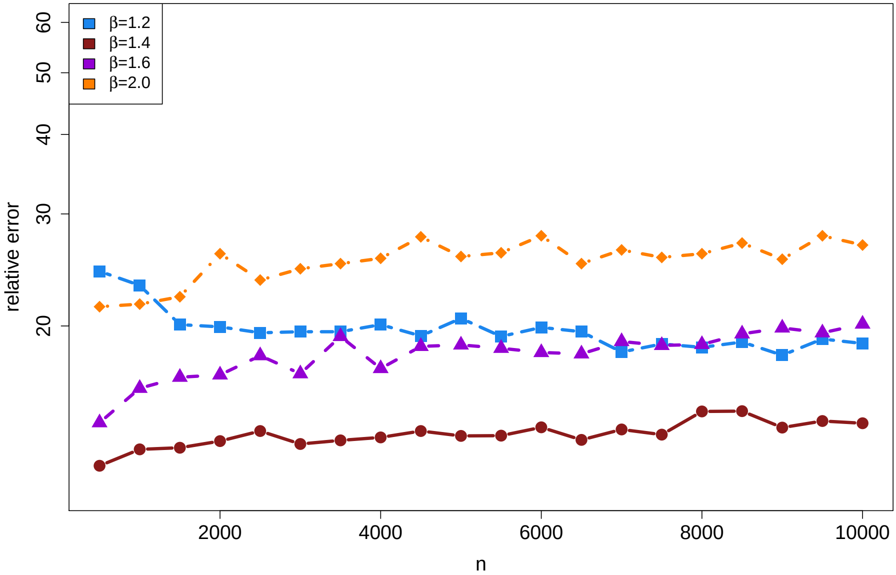
# Numerical Results for Reinsurance Example



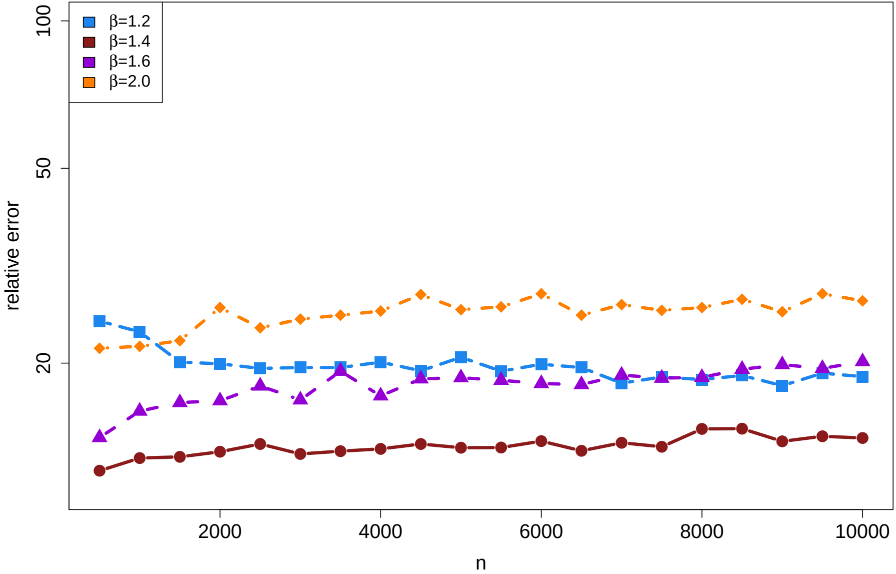
# Numerical Results for Reinsurance Example



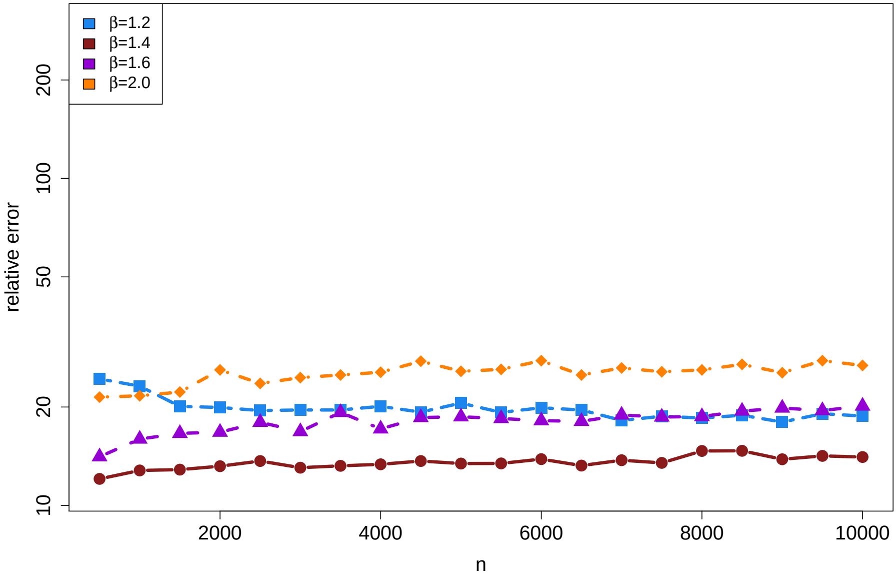
# Numerical Results for Reinsurance Example



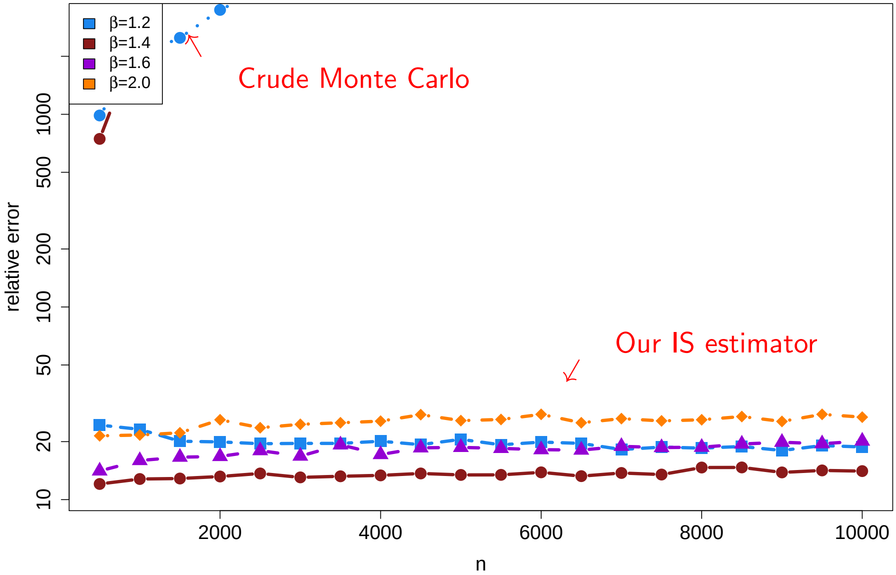
# Numerical Results for Reinsurance Example



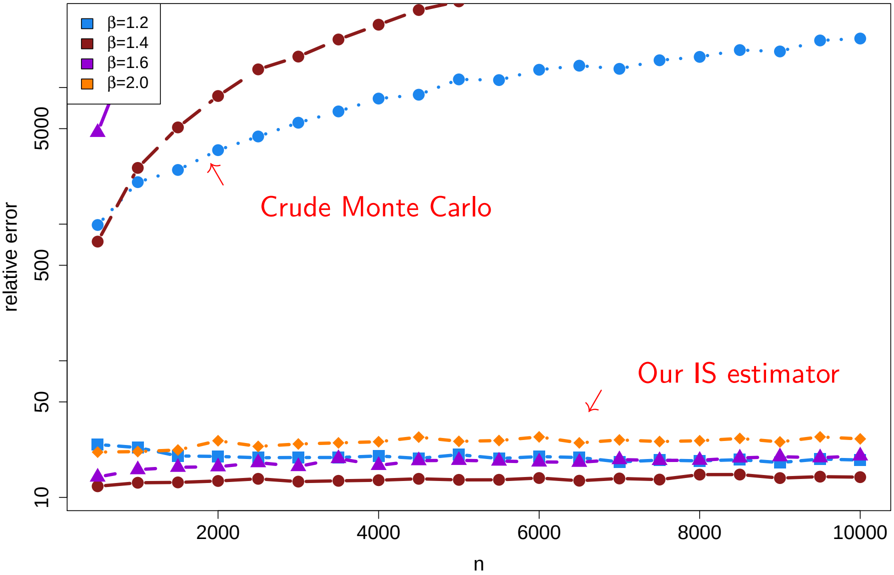
# Numerical Results for Reinsurance Example



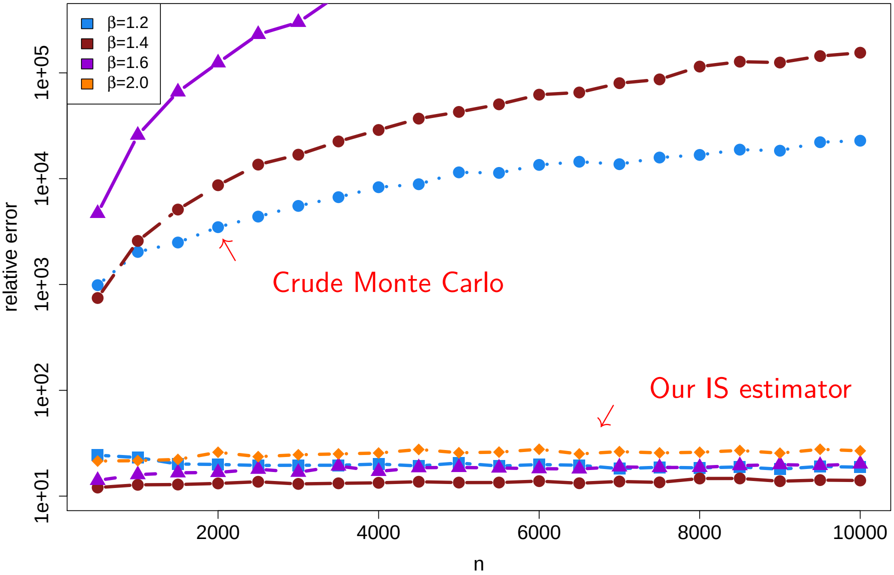
# Numerical Results for Reinsurance Example



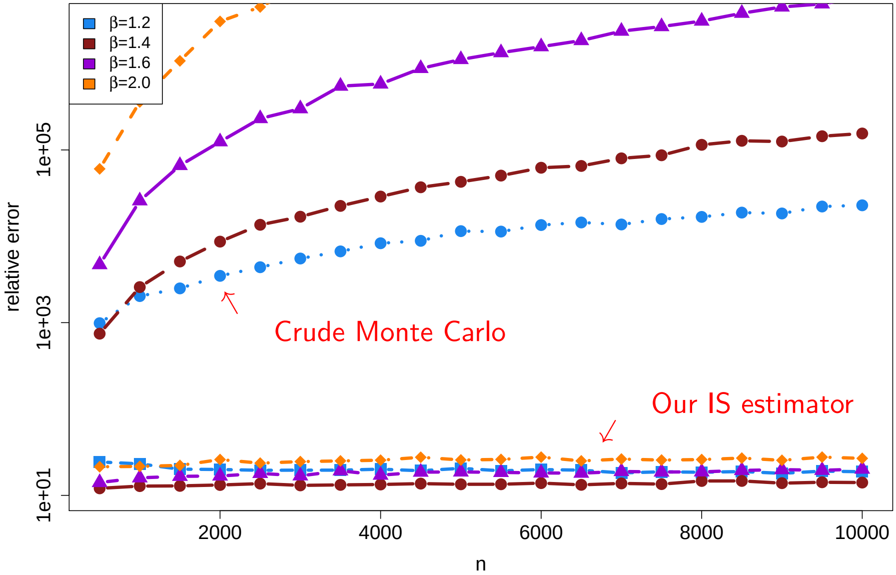
# Numerical Results for Reinsurance Example



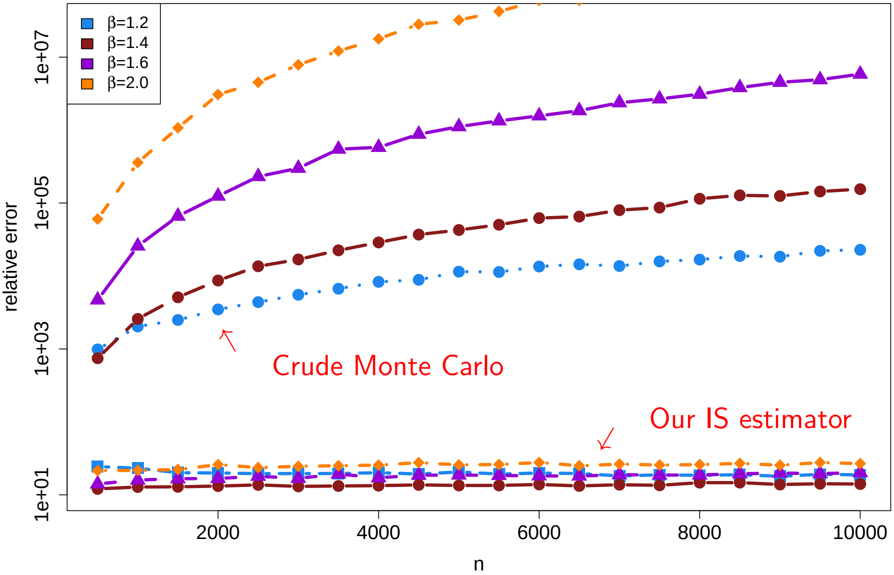
# Numerical Results for Reinsurance Example



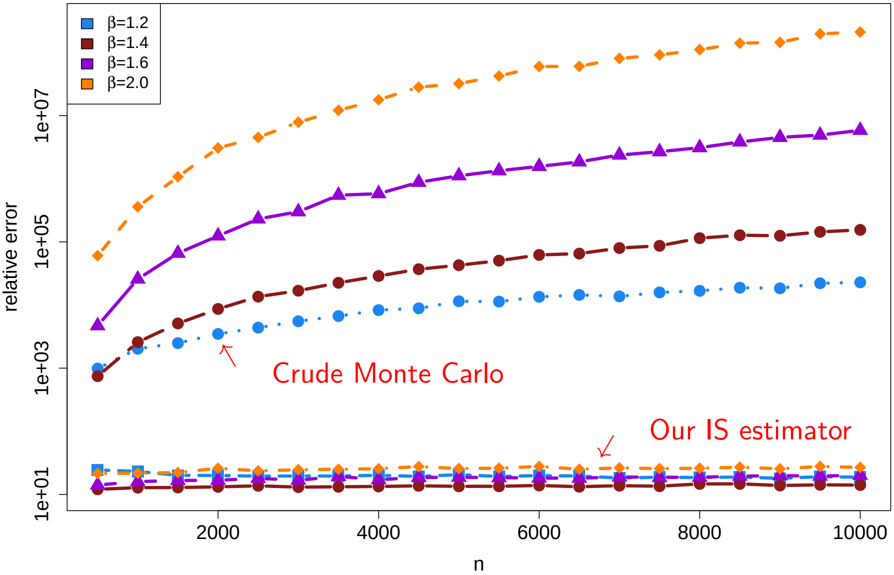
# Numerical Results for Reinsurance Example



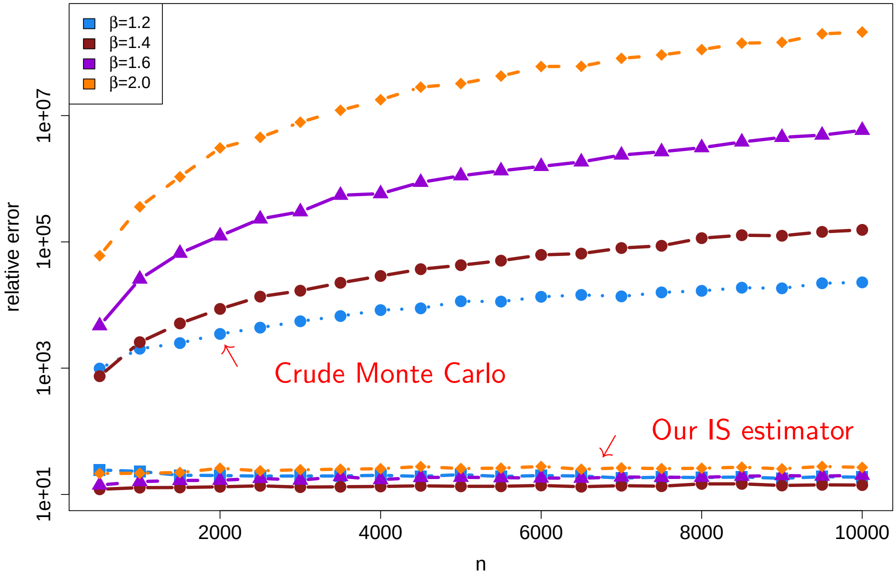
# Numerical Results for Reinsurance Example



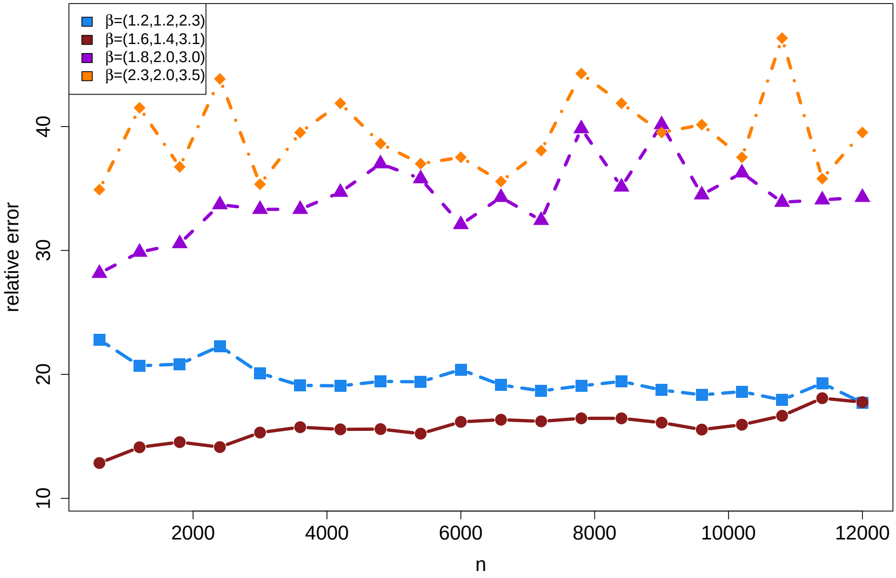
# Numerical Results for Reinsurance Example



# Numerical Results for Reinsurance Example



# Numerical Results for Fluid Network Example



# Summary

# Summary

- Catastrophe vs Conspiracy Principle

# Summary

- Catastrophe vs Conspiracy Principle

- Catastrophe via Heavy-Tailed Large Deviation

# Summary

- Catastrophe vs Conspiracy Principle
- Catastrophe via Heavy-Tailed Large Deviation
- Global Dynamics of Heavy-Tailed Dynamical Systems

# Summary

- Catastrophe vs Conspiracy Principle
- Catastrophe via Heavy-Tailed Large Deviation
- Global Dynamics of Heavy-Tailed Dynamical Systems
- Elimination of sharp local minima from SGD & Other Related Topics