

# **Heavy-Tailed Large Deviations, Metastability, and Eliminating Sharp Minima from SGD**

Chang-Han Rhee

Northwestern University

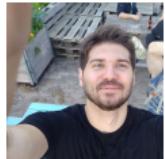
Actuarial Mathematics & Statistics, Heriot-Watt

July 12, 2023

Based on the joint works with

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

# Team



Mihail Bazhba  
U. of Amsterdam



Jose Blanchet  
Stanford



Bohan Chen  
Munich Re



Sewoong Oh  
U. of Washington



Xingyu Wang  
Northwestern



Zhe Su  
Northwestern



Bert Zwart  
CWI

# **Generalization Mystery of Deep Learning**

# Modern “Artificial Intelligence”

“Artificial Intelligence is like electricity.”

- Andrew Ng

# Empirical Success of Deep Neural Networks (DNNs)

“Deep Learning is eating the world.”

- Jorge Nocedal

## Empirical Success of Deep Neural Networks (DNNs)

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

# Empirical Success of Deep Neural Networks (DNNs)

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

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Training Set



Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

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

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Training Set



Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

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

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Training Set



Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

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

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Training Set



Test Set image source

# Empirical Success of Deep Neural Networks (DNNs)

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

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Training Set



Test Set image source

## Empirical Success of Deep Neural Networks (DNNs)

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

A Central Mystery of Deep Learning

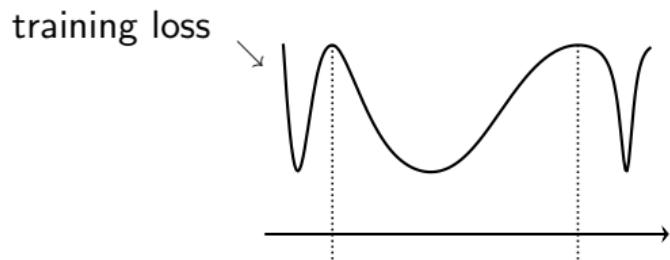
**Heavy Tails may have something to do with the Mystery**

## Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

## Heavy Tails may have something to do with the Mystery

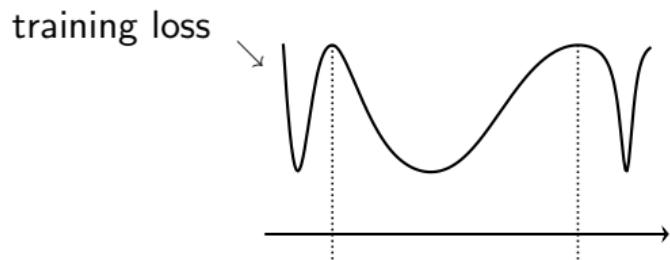
- Popular explanation: SGD somehow finds flat local minima.



## Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

↑ tends to generalize well

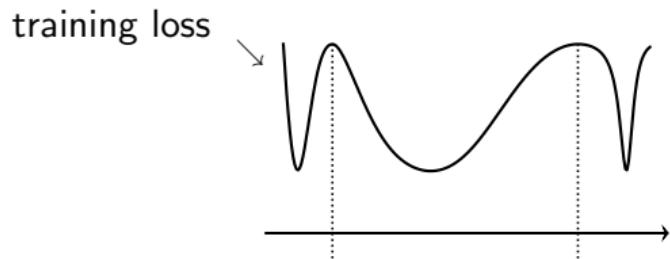


## Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

tends to generalize well

- But how?

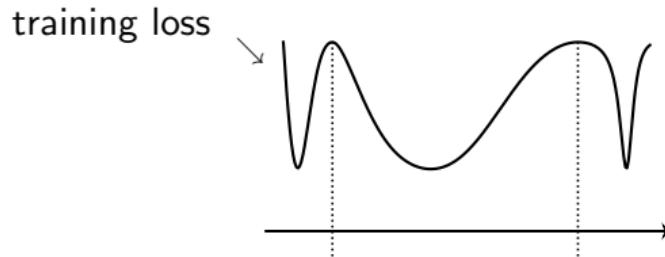


# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

tends to generalize well

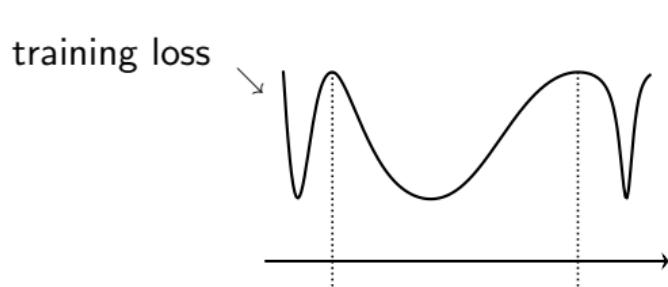
- But how?



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

# Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.  
    ↑ tends to generalize well
- But how? Previous theoretical attempts failed to explain how.



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

## Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.  
     tends to generalize well
- But how? Previous theoretical attempts failed to explain how.
- Evidence/arguments for heavy-tails in training DNNs with SGDs!  
eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

## Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.  
     tends to generalize well
- But how? Previous theoretical attempts failed to explain how.
- Evidence/arguments for heavy-tails in training DNNs with SGDs!  
eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

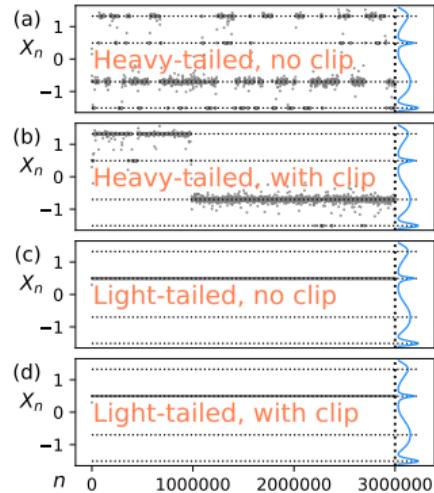
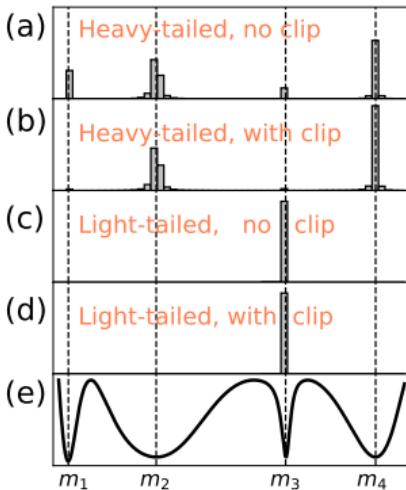
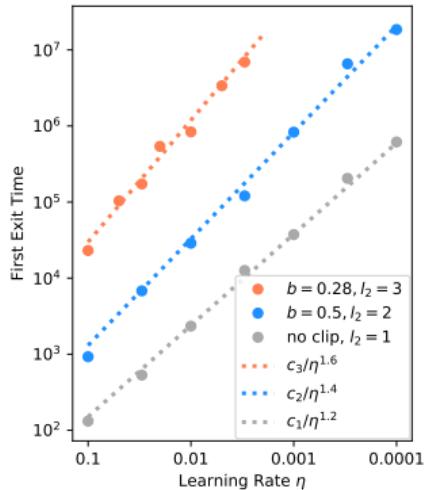
**Heavy-tailed SGD escapes local minima and prefers flat local minima.**

## Heavy Tails may have something to do with the Mystery

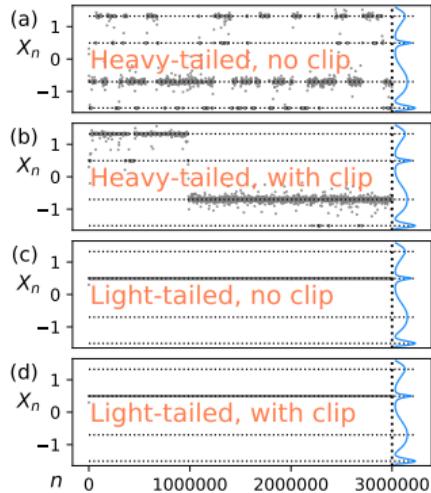
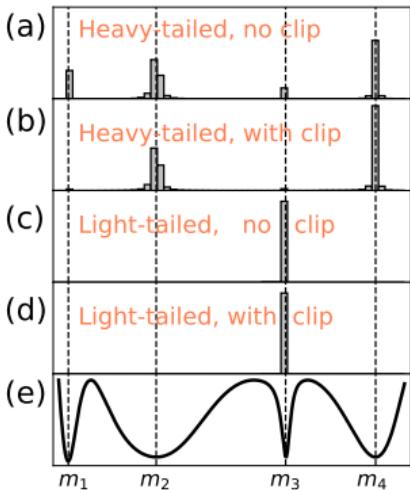
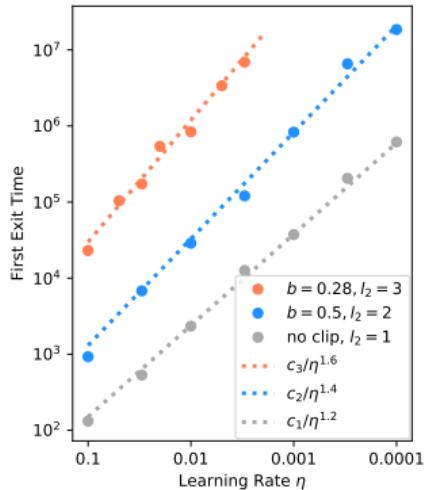
- Popular explanation: SGD somehow finds flat local minima.  
     tends to generalize well
- But how? Previous theoretical attempts failed to explain how.
- Evidence/arguments for heavy-tails in training DNNs with SGDs!  
eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

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

# Entirely Different Global Dynamics Depending on Tail Behaviors

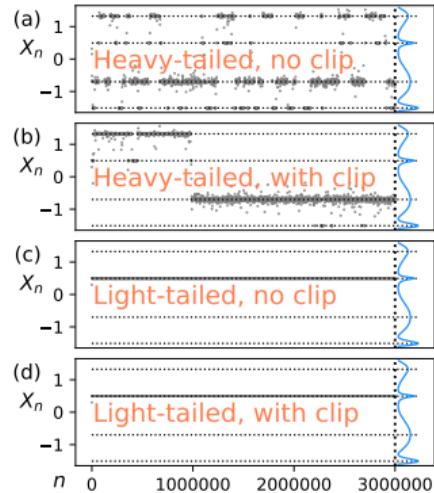
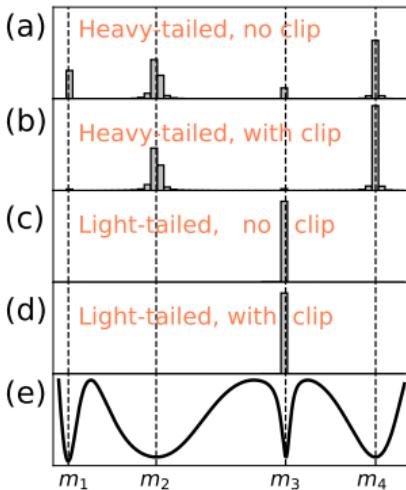
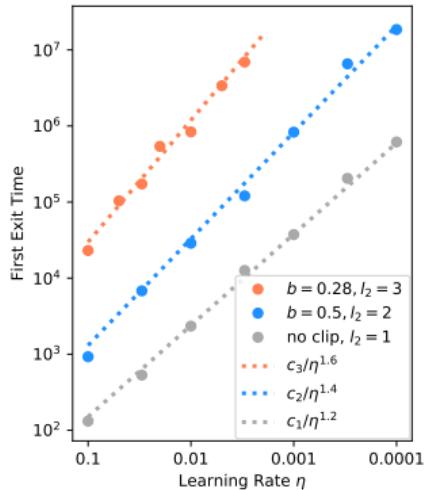


# Entirely Different Global Dynamics Depending on Tail Behaviors



Explanation?

# Entirely Different Global Dynamics Depending on Tail Behaviors



Explanation? Catastrophe Principle.

## **Heavy Tails and Catastrophe Principle**

# Rare Events depend on “Tail Behaviors”

## Light-Tailed Distributions

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



## Heavy-Tailed Distributions

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



Instagram

# Rare Events depend on “Tail Behaviors”

## Light-Tailed Distributions

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



## Heavy-Tailed Distributions

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



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

# Rare Events depend on “Tail Behaviors”

## Light-Tailed Distributions

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

## Heavy-Tailed Distributions

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

Systemwide rare events

arise because

EVERYTHING goes wrong.

(Conspiracy Principle)



Instagram

Structural difference in the way systemwide rare events arise.

# Rare Events depend on “Tail Behaviors”

## Light-Tailed Distributions

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

## Heavy-Tailed Distributions

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

Systemwide rare events

arise because

**EVERYTHING** goes wrong.

(Conspiracy Principle)

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

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

Poisson Arrival

Initial Capital

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

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

## Insurance Example: Capital Reserve

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

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

## Typical Scenario

## Typical Scenario

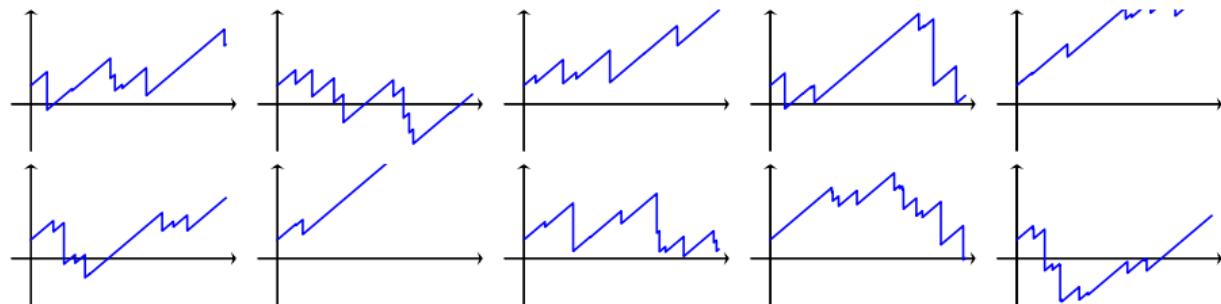
Sample paths of  $\bar{Y}_n$ :

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

## Typical Scenario

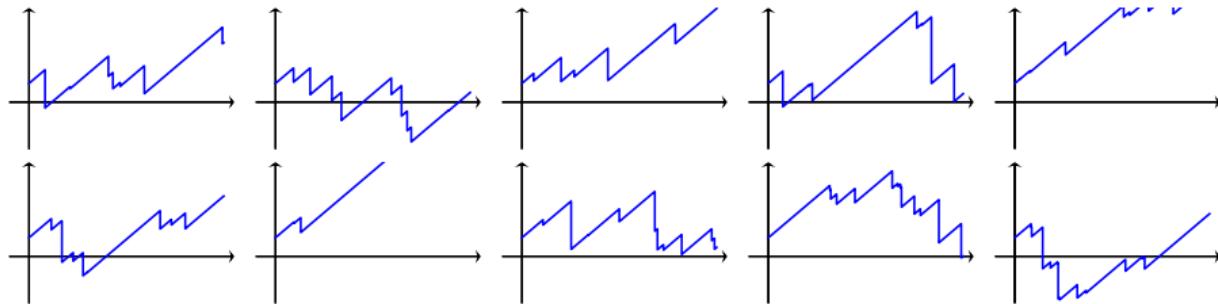
Sample paths of  $\bar{Y}_n$ :

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



## Typical Scenario

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



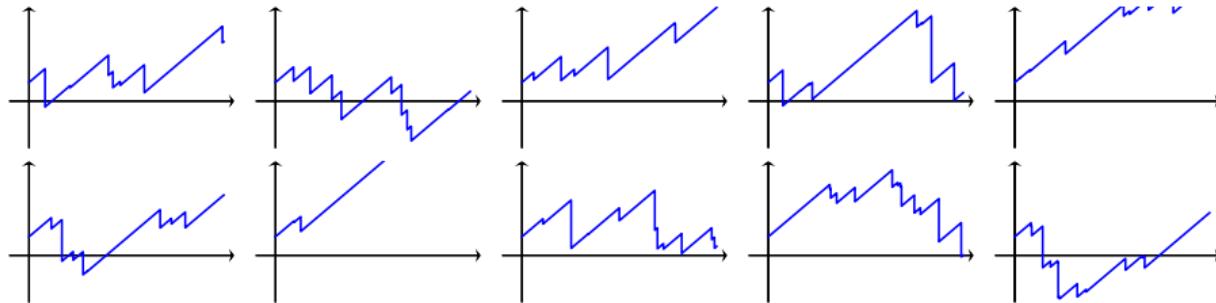
$n=10$  & claim sizes are **light-tailed**

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

$n=10$  & claim sizes are **heavy-tailed**

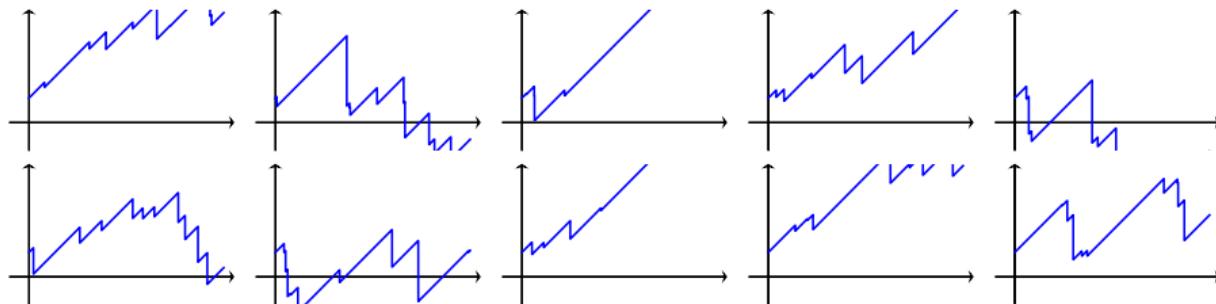
## Typical Scenario

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



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

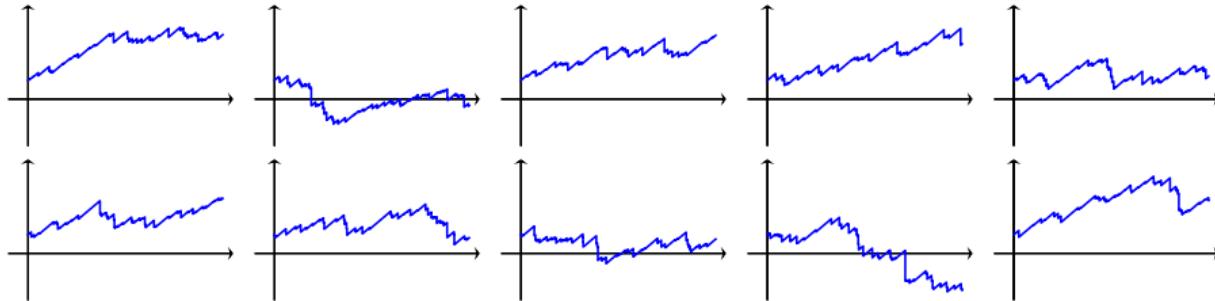
Sample paths of  $\bar{Y}_n$ :



$n=10$  & claim sizes are **heavy-tailed**

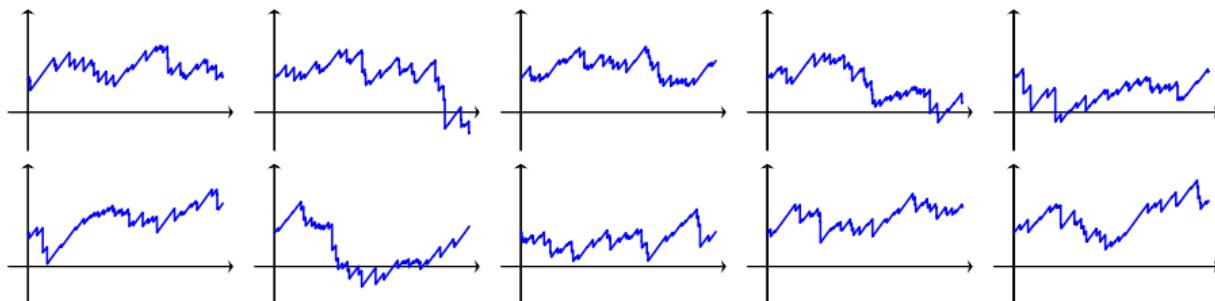
## Typical Scenario

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



$n=50$  & claim sizes are **light-tailed**

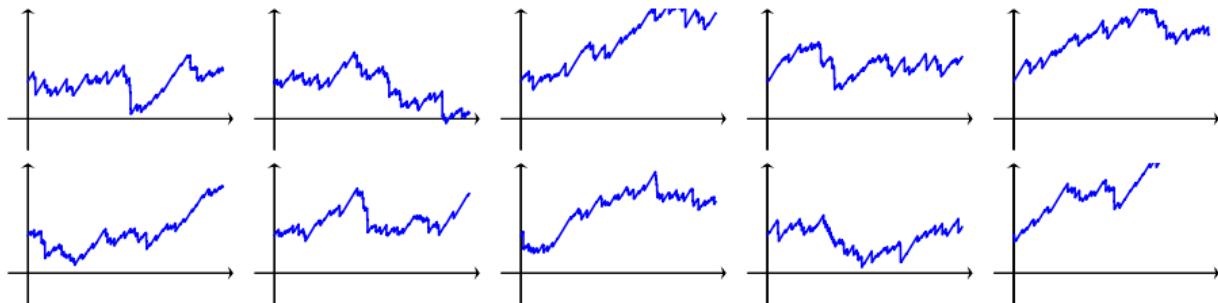
Sample paths of  $\bar{Y}_n$ :



$n=50$  & claim sizes are **heavy-tailed**

## Typical Scenario

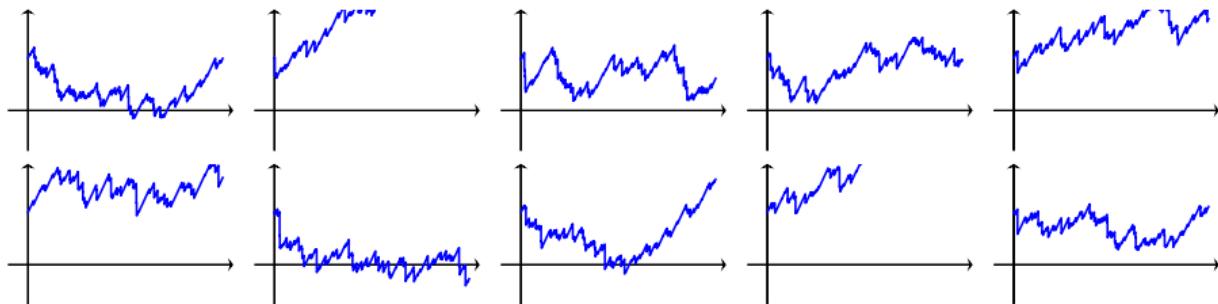
Sample paths of  $\bar{Y}_n$ :



$n=100$  & claim sizes are **light-tailed**

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

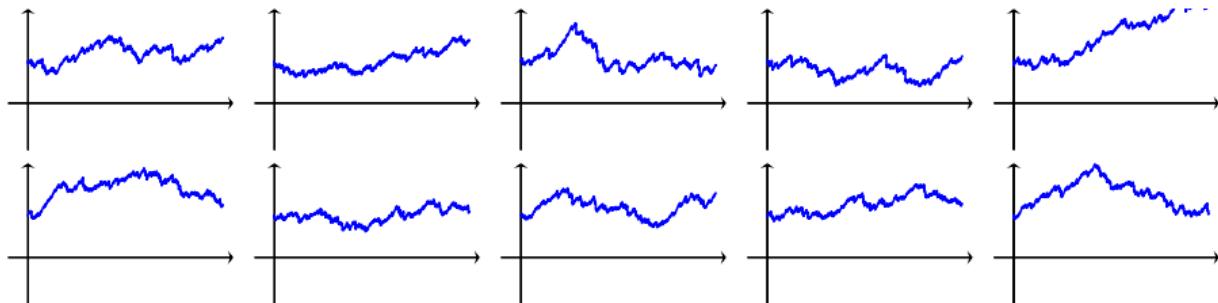
$n=100$  & claim sizes are **heavy-tailed**



## Typical Scenario

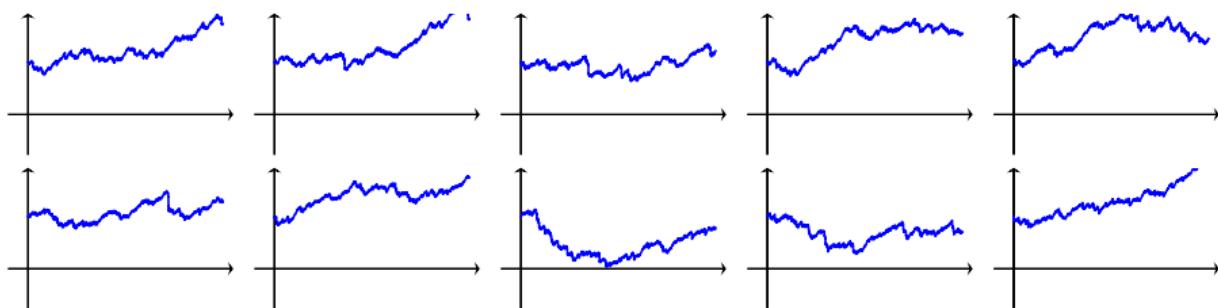
Sample paths of  $\bar{Y}_n$ :

$n=500$  & claim sizes are **light-tailed**



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

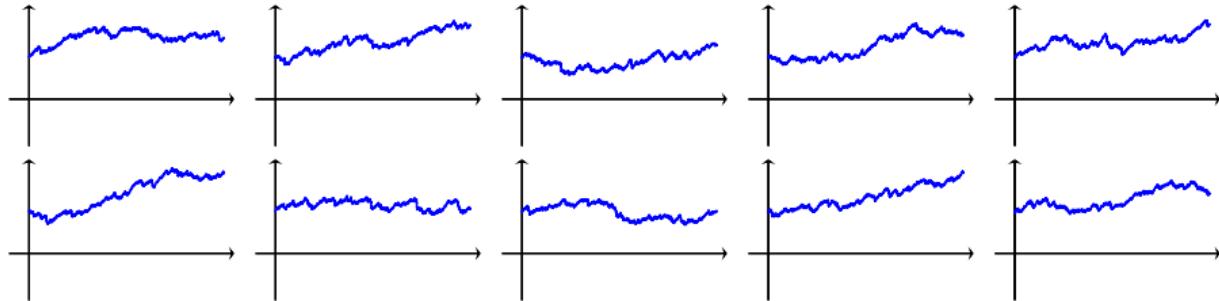
$n=500$  & claim sizes are **heavy-tailed**



## Typical Scenario

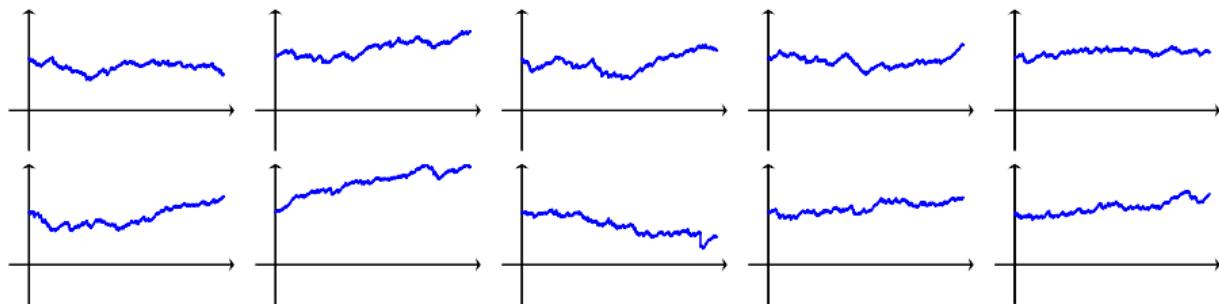
Sample paths of  $\bar{Y}_n$ :

$n=1000$  & claim sizes are **light-tailed**



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

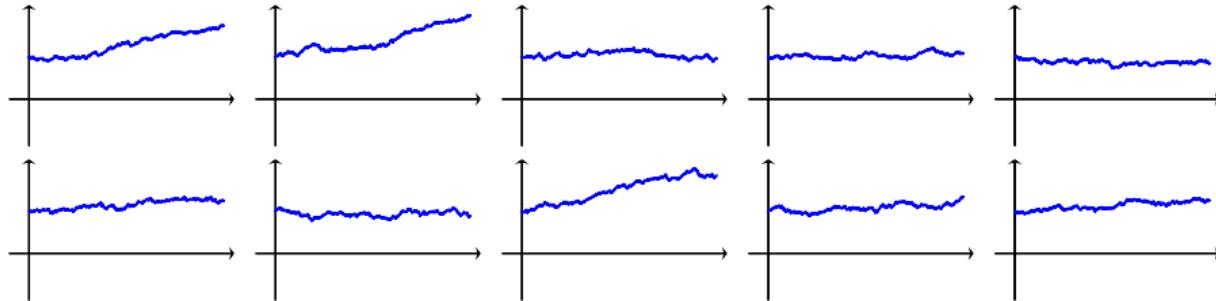
$n=1000$  & claim sizes are **heavy-tailed**



## Typical Scenario

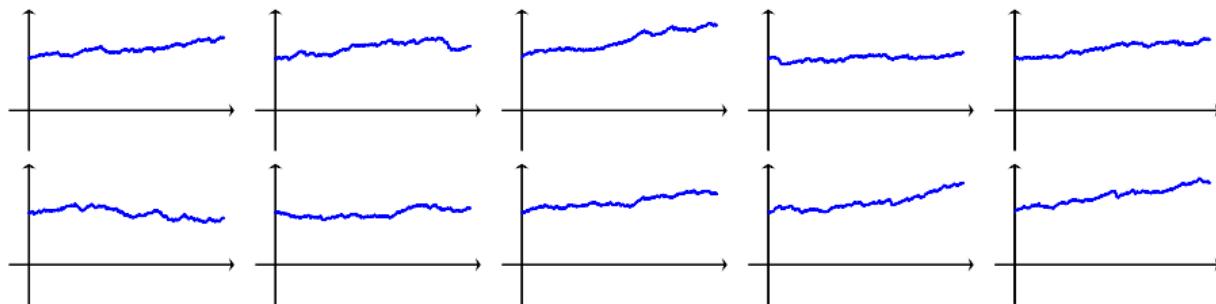
Sample paths of  $\bar{Y}_n$ :

$n=2500$  & claim sizes are **light-tailed**



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

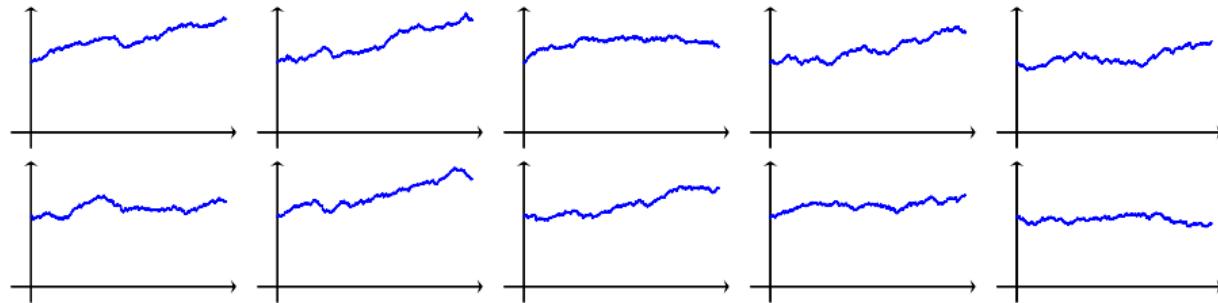
$n=2500$  & claim sizes are **heavy-tailed**



## Typical Scenario

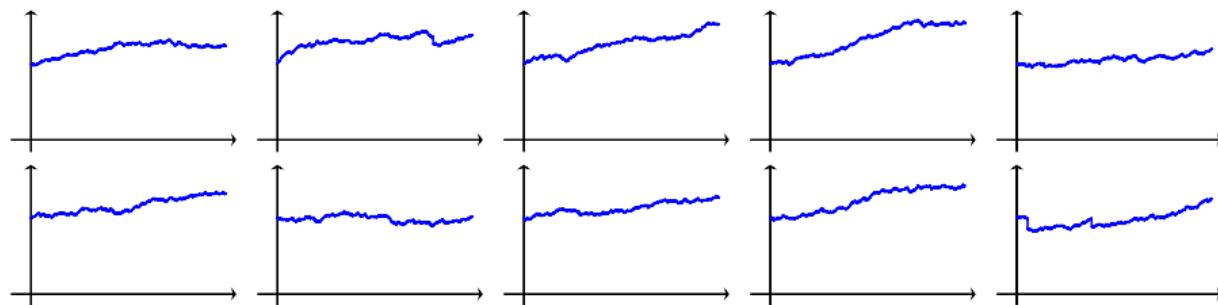
Sample paths of  $\bar{Y}_n$ :

$n=5000$  & claim sizes are **light-tailed**



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

$n=5000$  & claim sizes are **heavy-tailed**

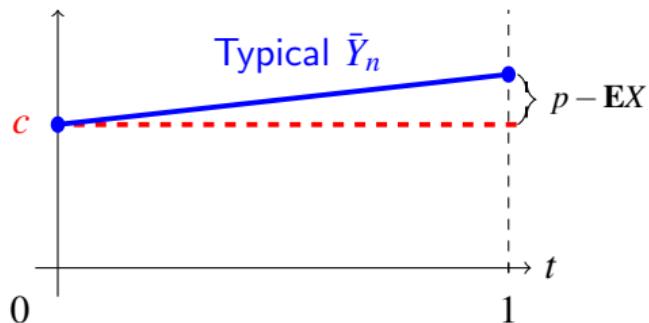


## Typical Scenario

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

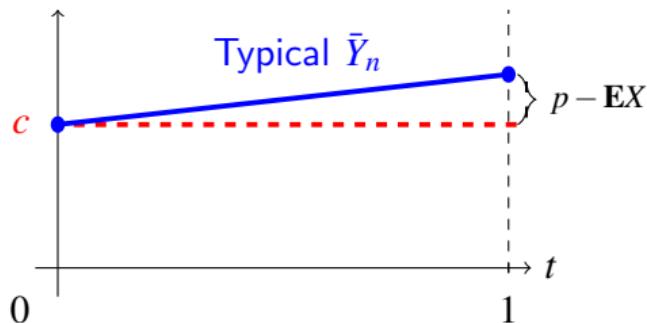
## Typical Scenario

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



## Typical Scenario

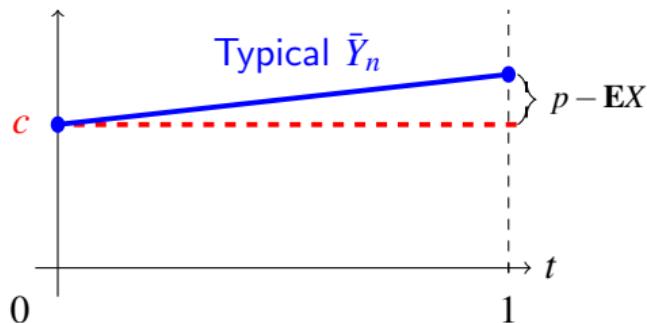
That is,  $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$  for large  $n$ .



**Typically, your business will flourish**

## Typical Scenario

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

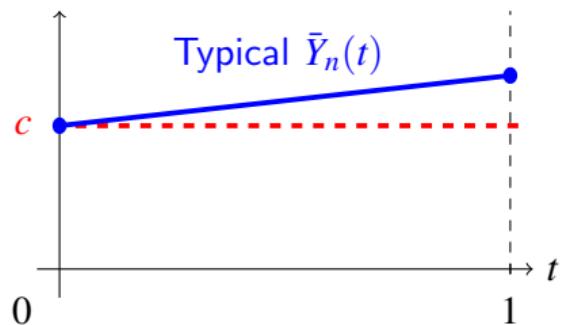


Typically, your business will flourish

regardless of the tail distributions.

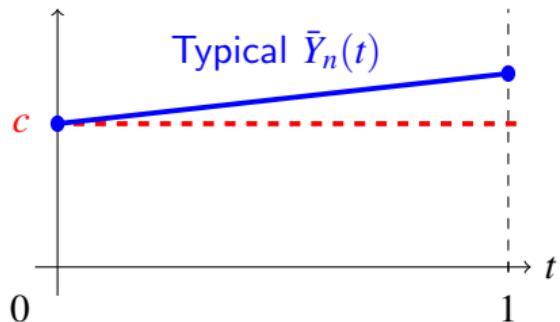
**What about atypical cases?**

## A Rare Event: Bankruptcy



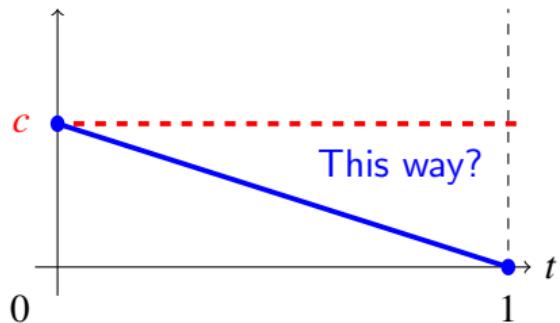
## A Rare Event: Bankruptcy

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



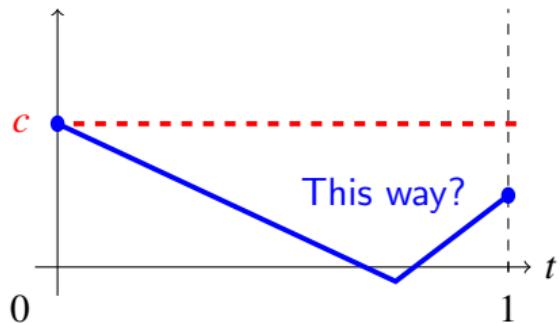
## A Rare Event: Bankruptcy

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



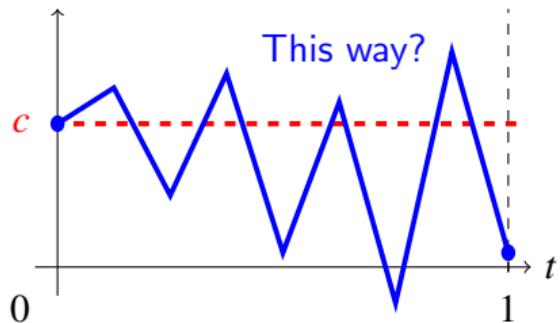
## A Rare Event: Bankruptcy

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



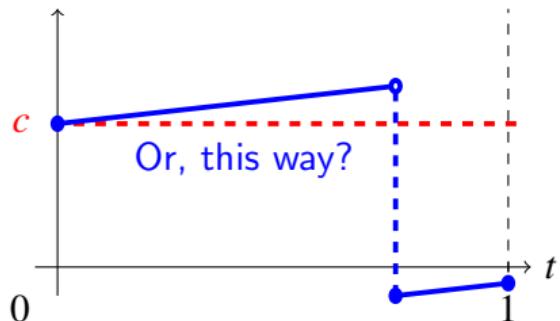
## A Rare Event: Bankruptcy

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



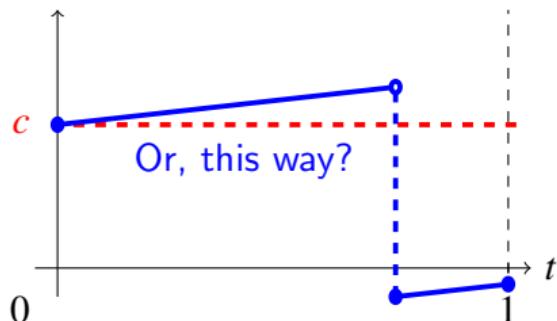
## A Rare Event: Bankruptcy

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



## A Rare Event: Bankruptcy

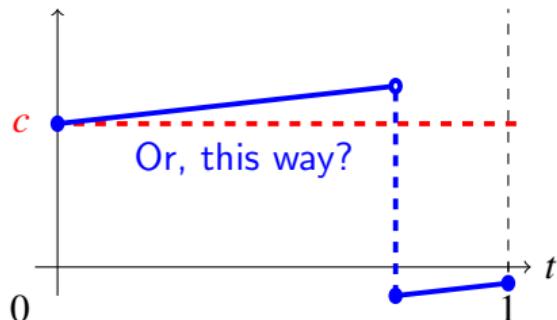
Consider  $B \triangleq \{ \bar{Y}_n \text{ falls below } 0 \text{ on } [0, 1] \}$ . (i.e., Bankruptcy)



Are we going to see clear patterns?

## A Rare Event: Bankruptcy

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

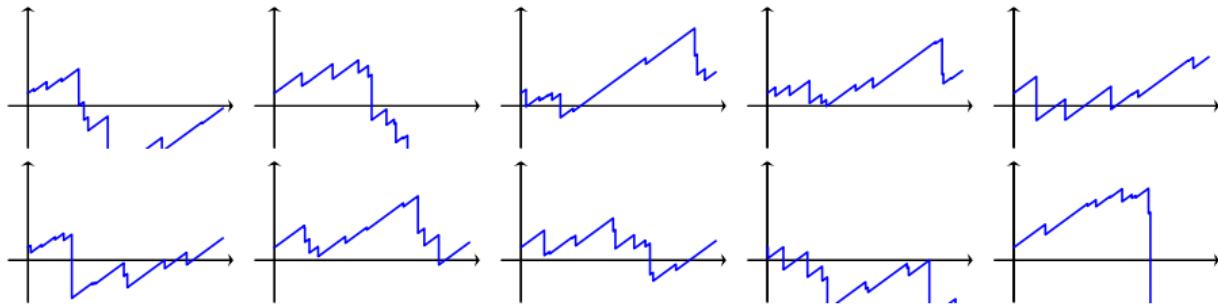


Are we going to see clear patterns?

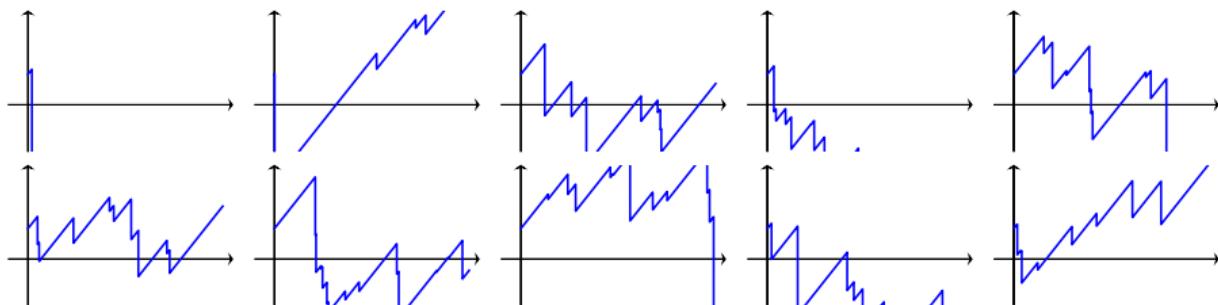
Do they depend on the tail distributions?

# A Rare Event: Bankruptcy

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

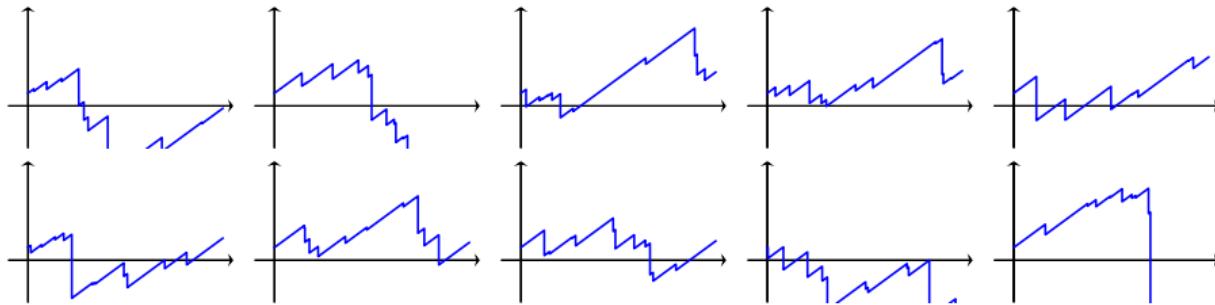


Sample paths of  $\bar{Y}_{10}$  conditional on  $B$  for **heavy-tailed** claims:

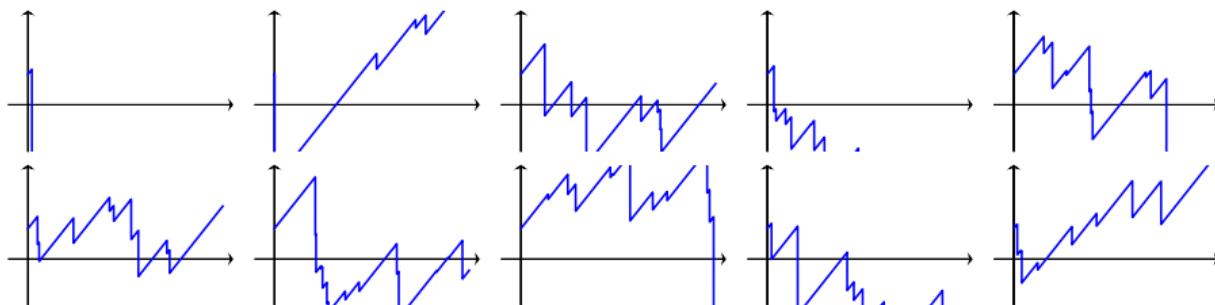


# A Rare Event: Bankruptcy

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

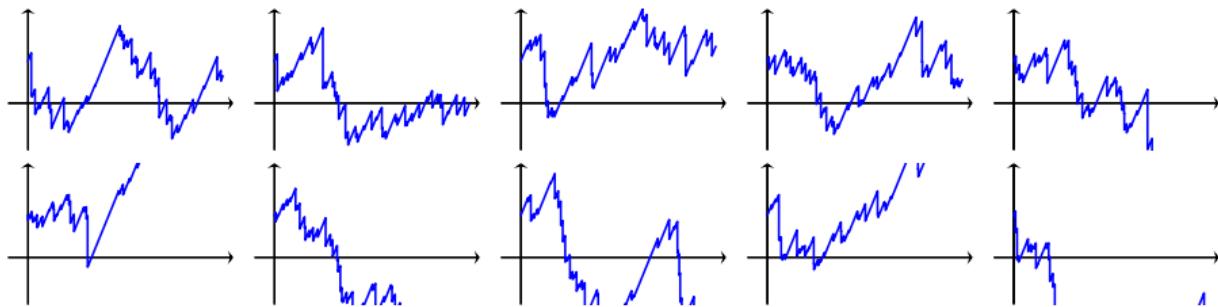


Sample paths of  $\bar{Y}_{10}$  conditional on  $B$  for **heavy-tailed** claims:

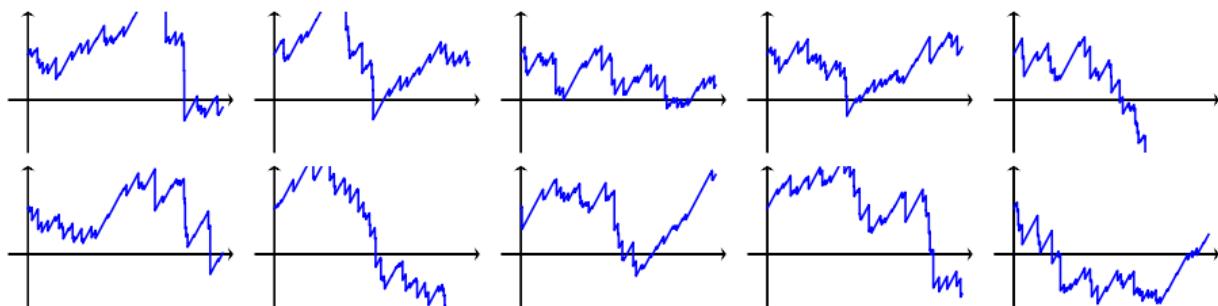


## A Rare Event: Bankruptcy

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

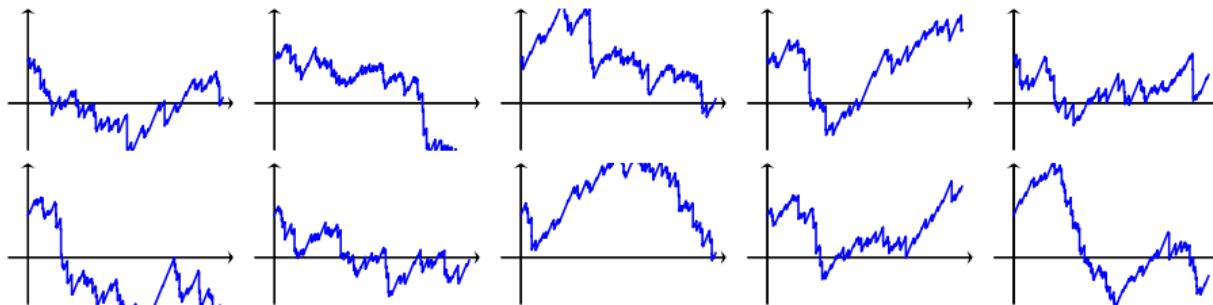


Sample paths of  $\bar{Y}_{50}$  conditional on  $B$  for **heavy-tailed** claims:

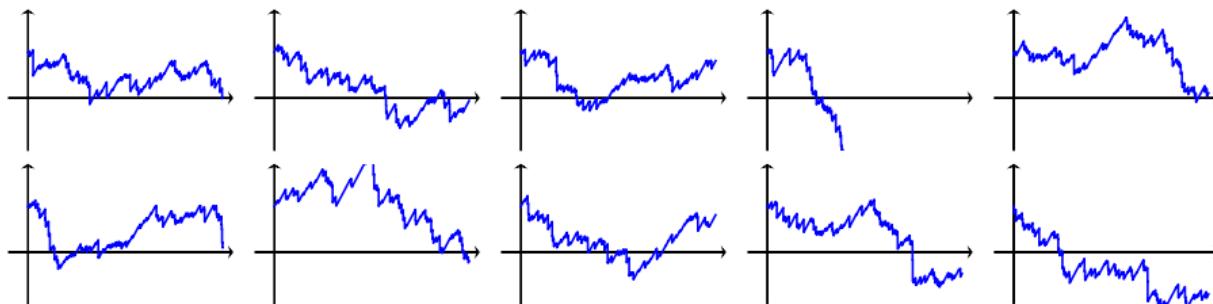


## A Rare Event: Bankruptcy

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

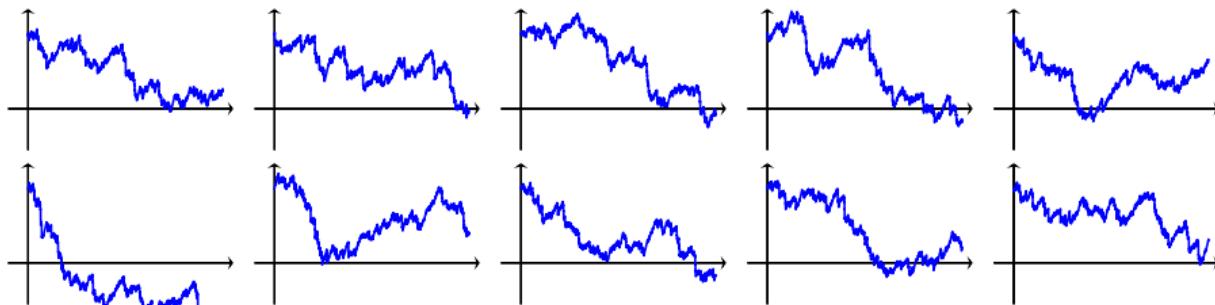


Sample paths of  $\bar{Y}_{100}$  conditional on  $B$  for **heavy-tailed** claims:

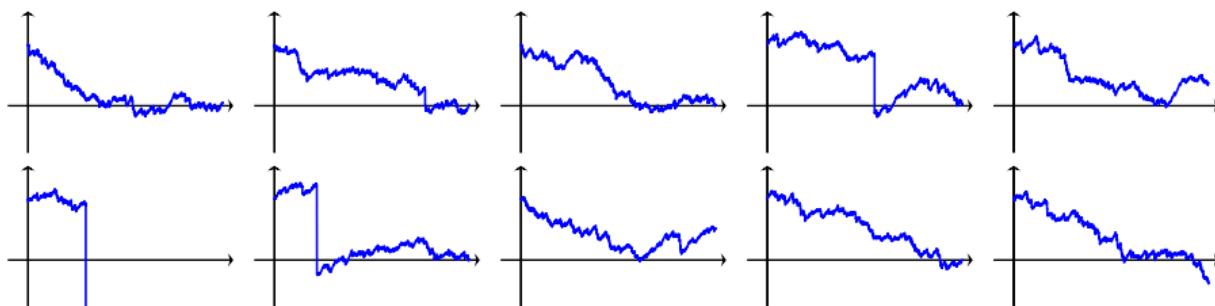


# A Rare Event: Bankruptcy

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

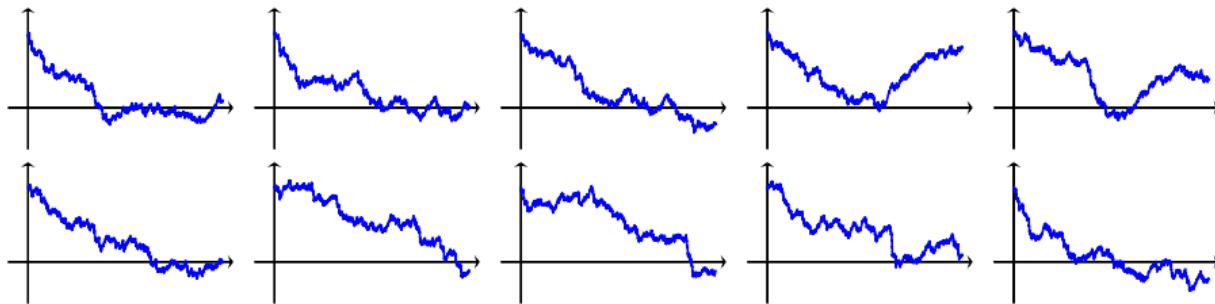


Sample paths of  $\bar{Y}_{500}$  conditional on  $B$  for **heavy-tailed** claims:

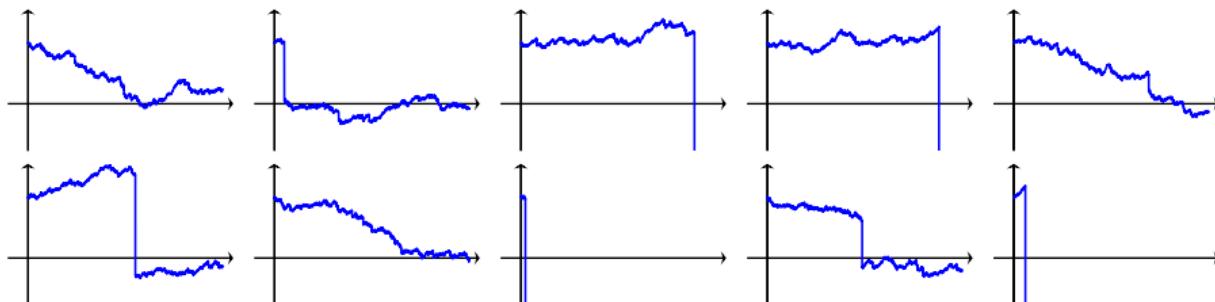


# A Rare Event: Bankruptcy

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

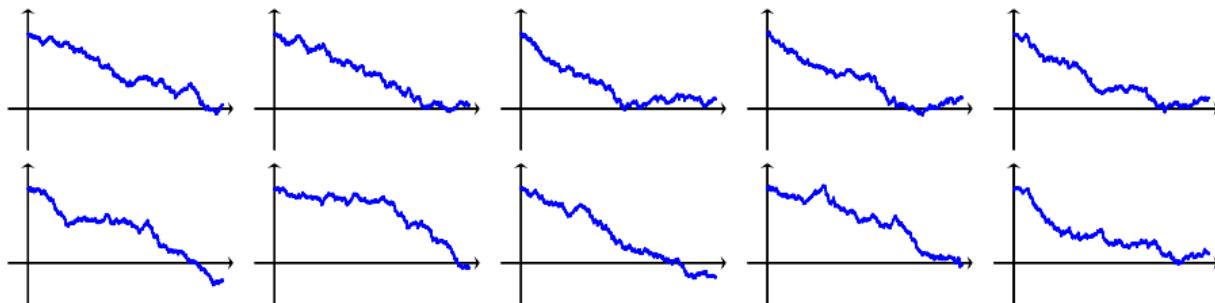


Sample paths of  $\bar{Y}_{1000}$  conditional on  $B$  for **heavy-tailed** claims:



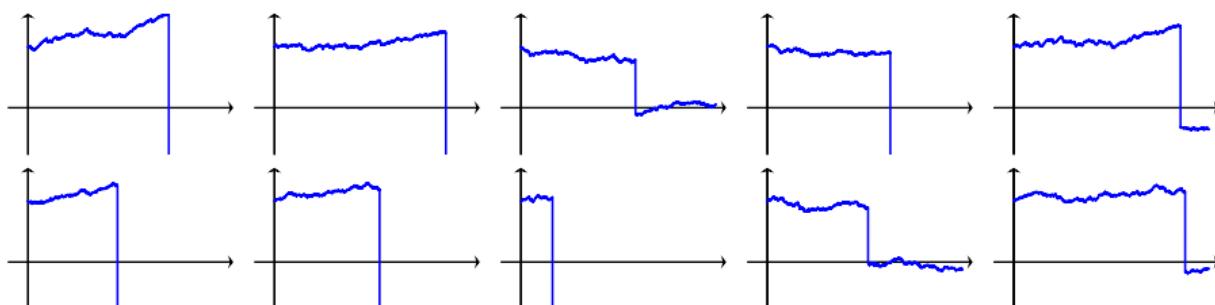
# A Rare Event: Bankruptcy

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



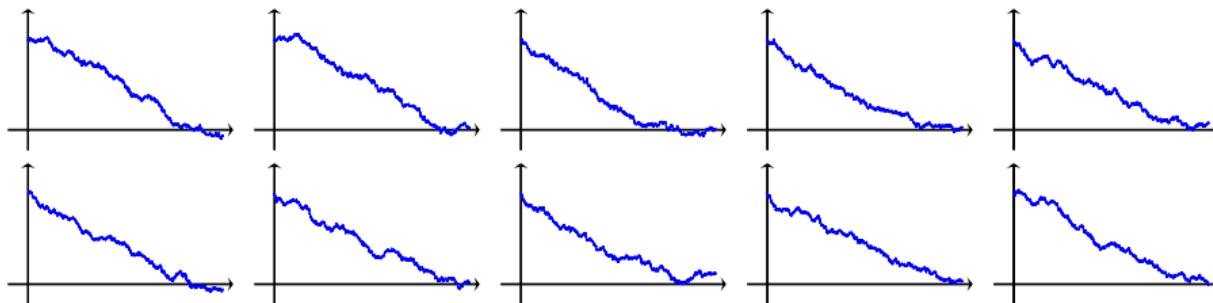
Bankruptcy

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



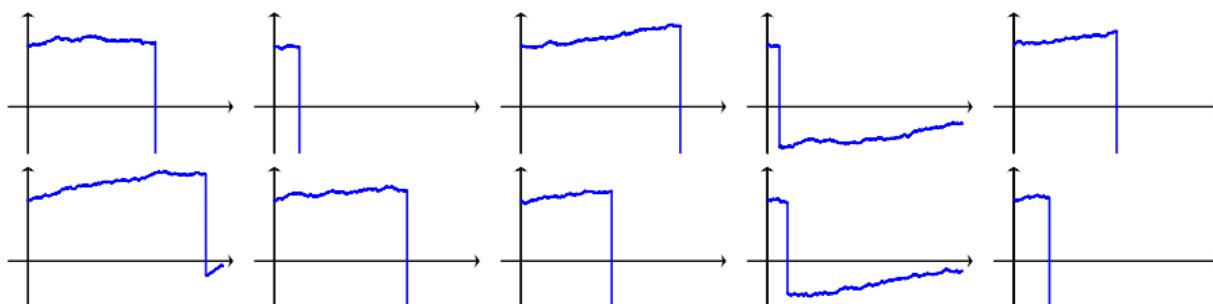
# A Rare Event: Bankruptcy

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

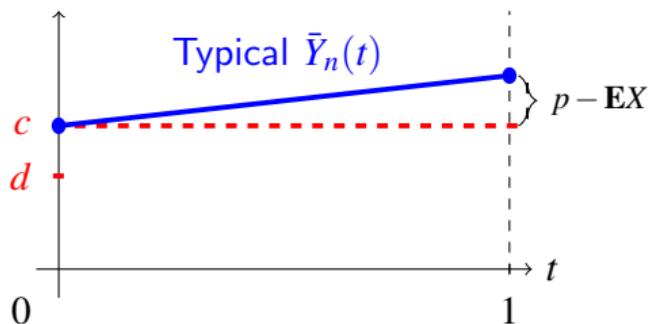


Bankruptcy

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

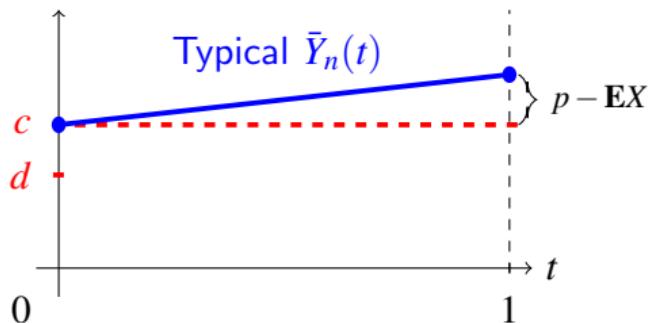


## Bankruptcy Despite Reinsurance



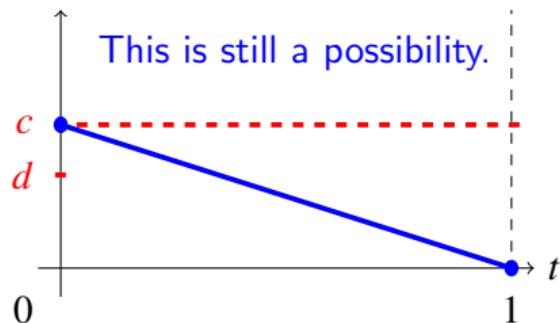
## Bankruptcy Despite of Reinsurance

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



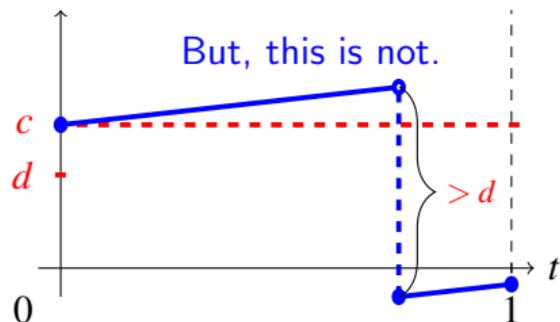
## Bankruptcy Despite of Reinsurance

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



# Bankruptcy Despite of Reinsurance

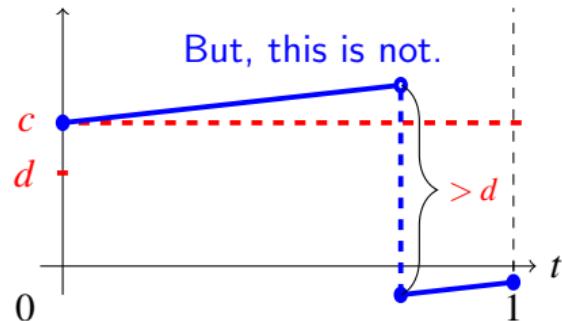
Consider  $R \triangleq \{ \bar{Y}_n \text{ falls below } 0 \text{ on } [0,1], \text{ jump sizes } \leq d \}$



But, this is not.

## Bankruptcy Despite of Reinsurance

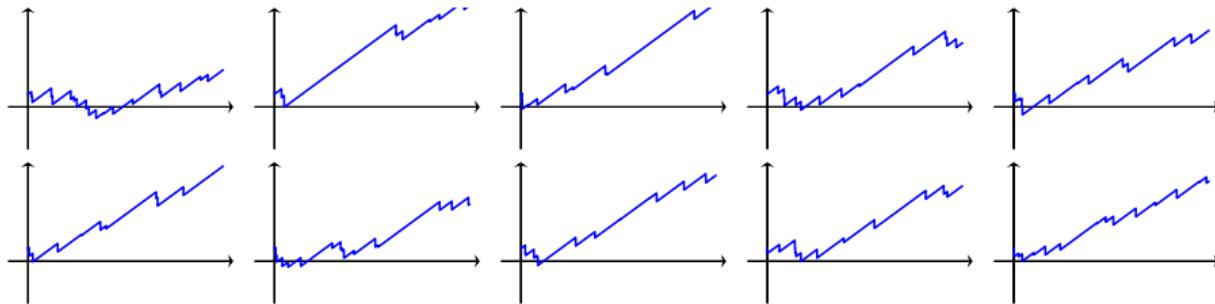
Consider  $R \triangleq \{ \bar{Y}_n \text{ falls below } 0 \text{ on } [0,1], \text{ jump sizes } \leq d \}$



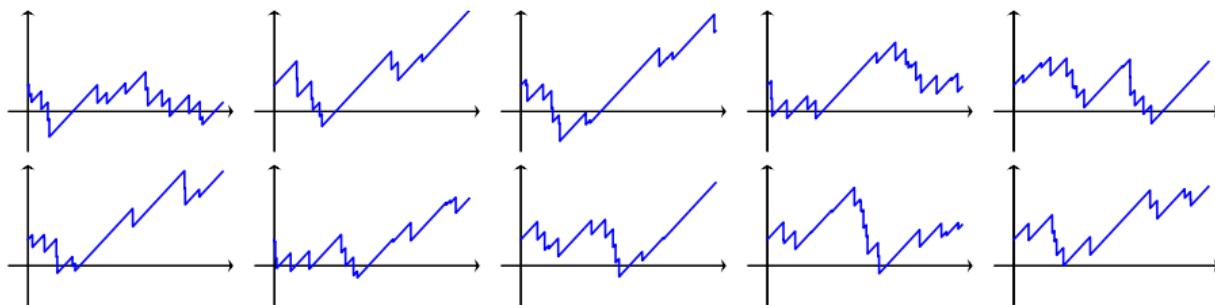
How does the pattern change in this case?

# Bankruptcy Despite Reinsurance

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

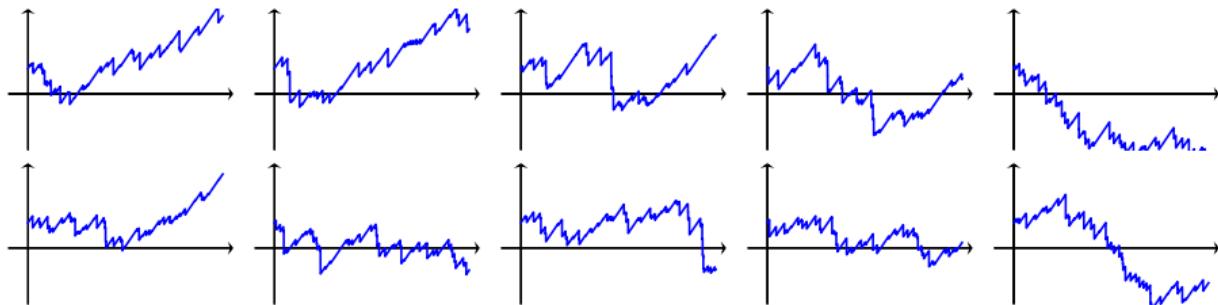


Sample paths of  $\bar{Y}_{10}$  conditional on  $R$  for **heavy-tailed** claims:

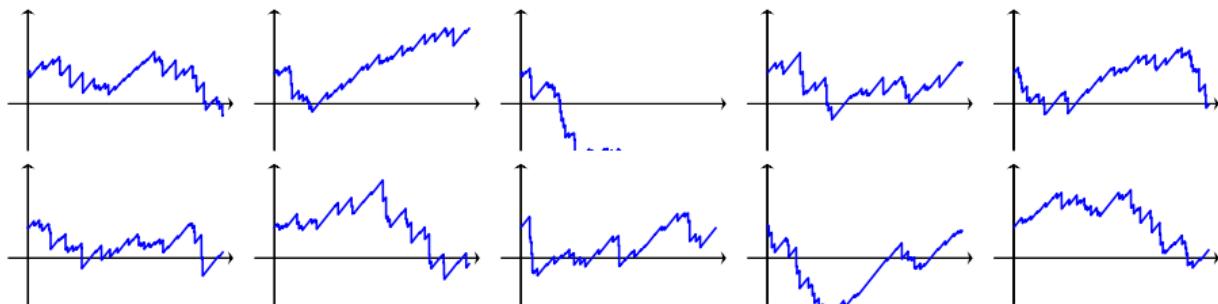


# Bankruptcy Despite Reinsurance

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

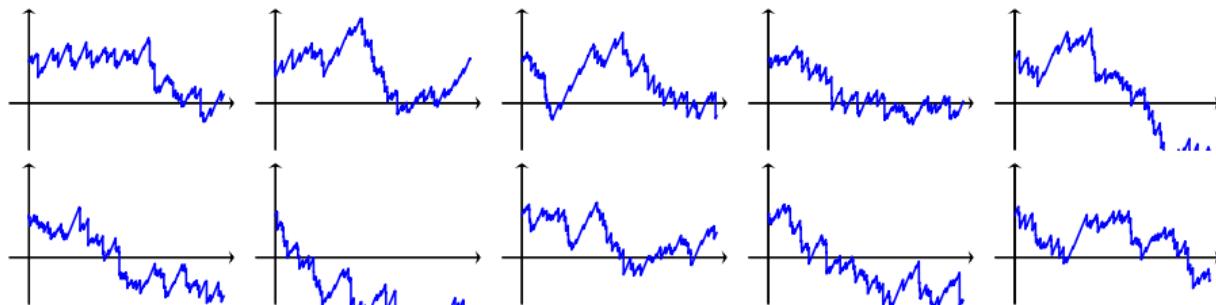


Sample paths of  $\bar{Y}_{50}$  conditional on  $R$  for **heavy-tailed** claims:

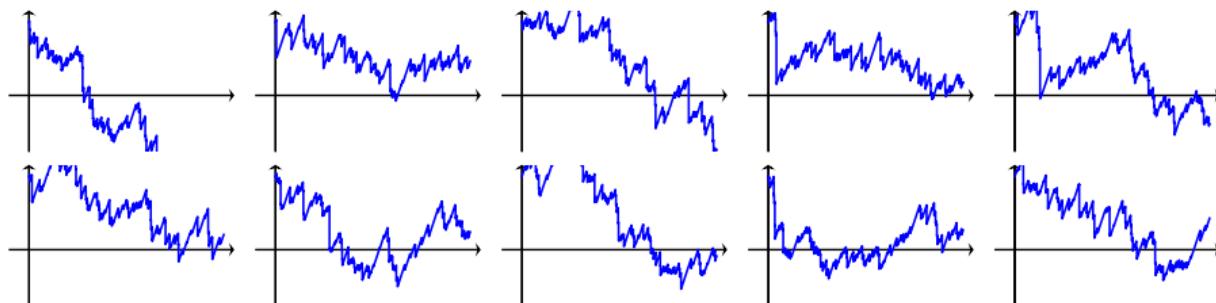


# Bankruptcy Despite Reinsurance

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

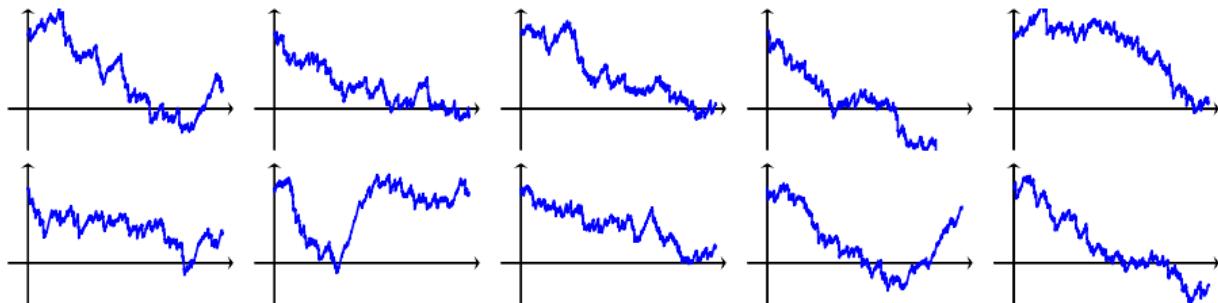


Sample paths of  $\bar{Y}_{100}$  conditional on  $R$  for **heavy-tailed** claims:

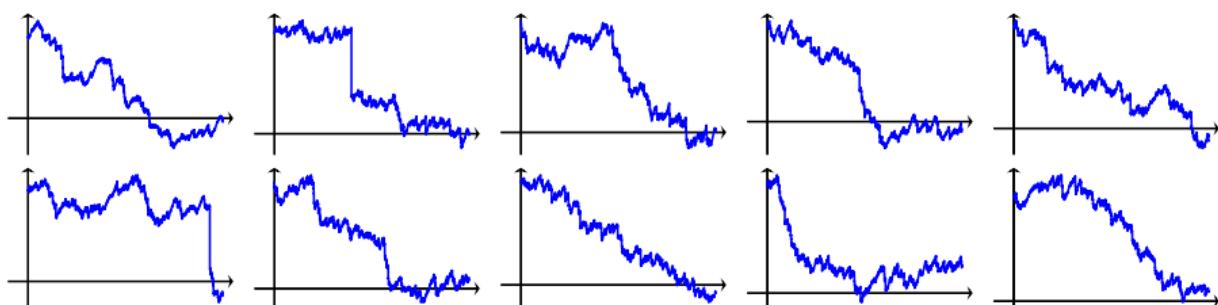


# Bankruptcy Despite Reinsurance

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

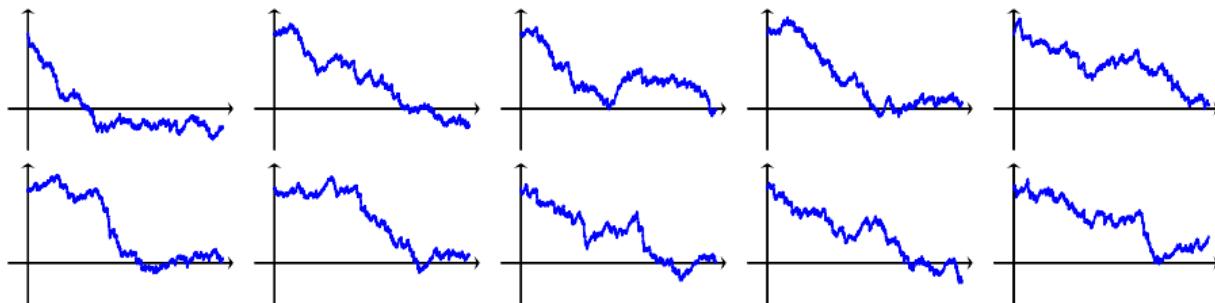


Sample paths of  $\bar{Y}_{500}$  conditional on  $R$  for **heavy-tailed** claims:

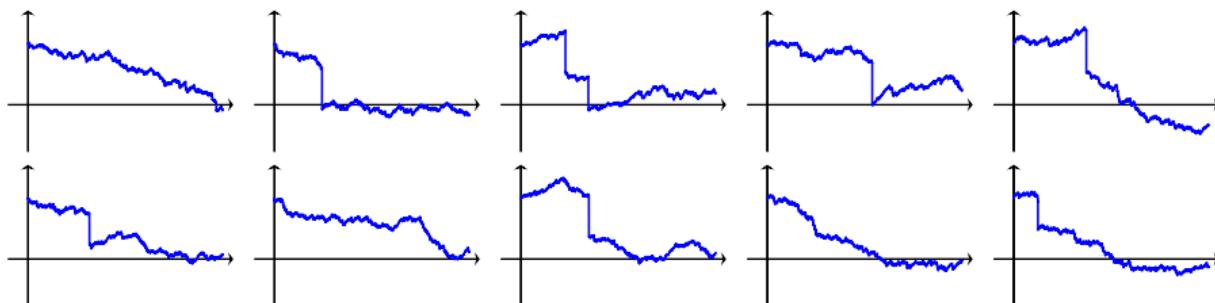


# Bankruptcy Despite Reinsurance

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

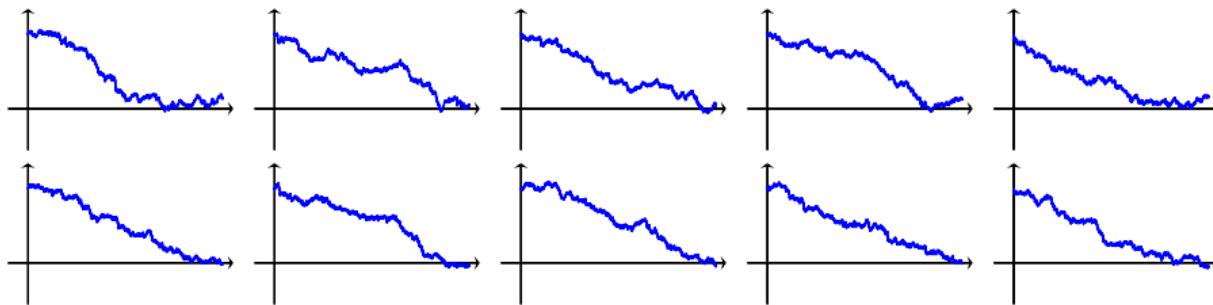


Sample paths of  $\bar{Y}_{1000}$  conditional on  $R$  for **heavy-tailed** claims:

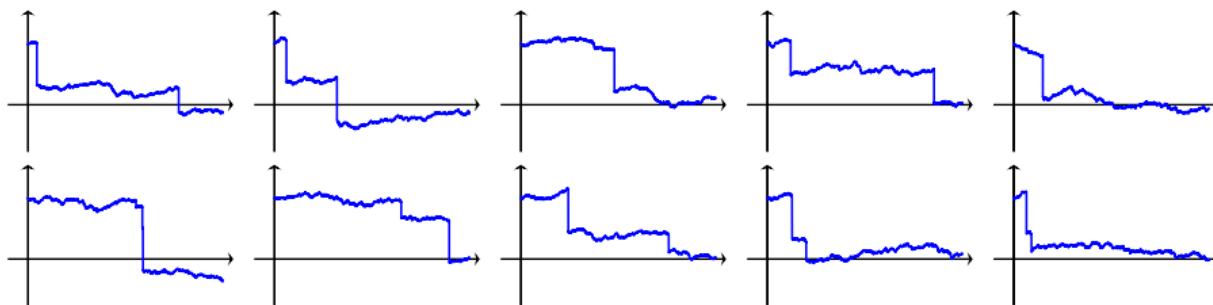


# Bankruptcy Despite Reinsurance

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

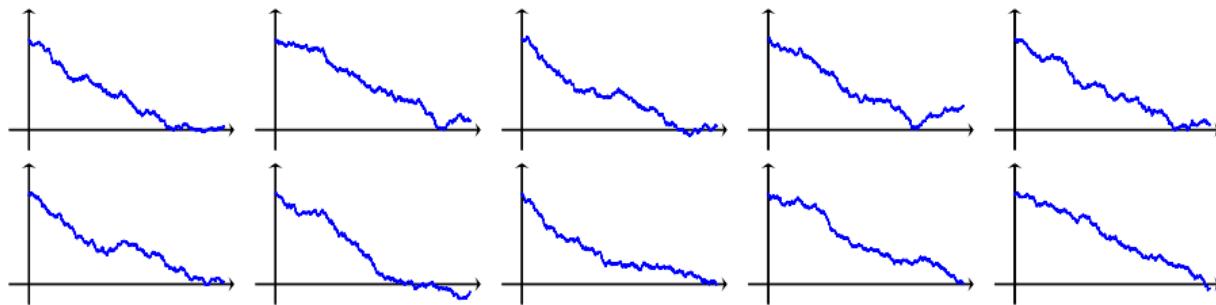


Sample paths of  $\bar{Y}_{2500}$  conditional on  $R$  for **heavy-tailed** claims:

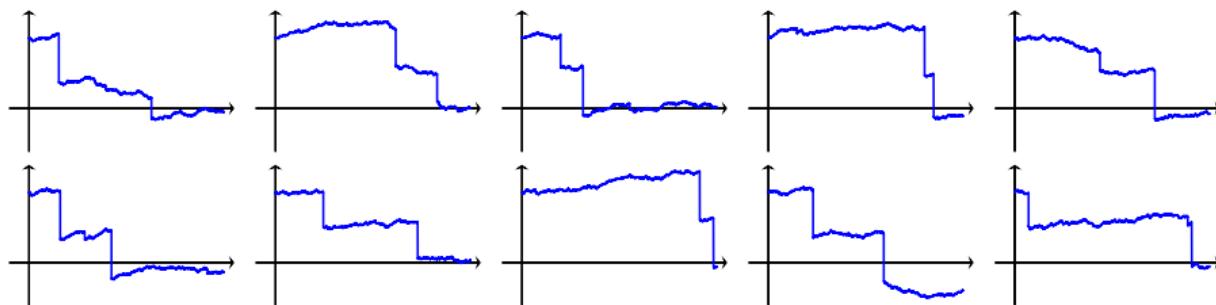


# Bankruptcy Despite of Reinsurance

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

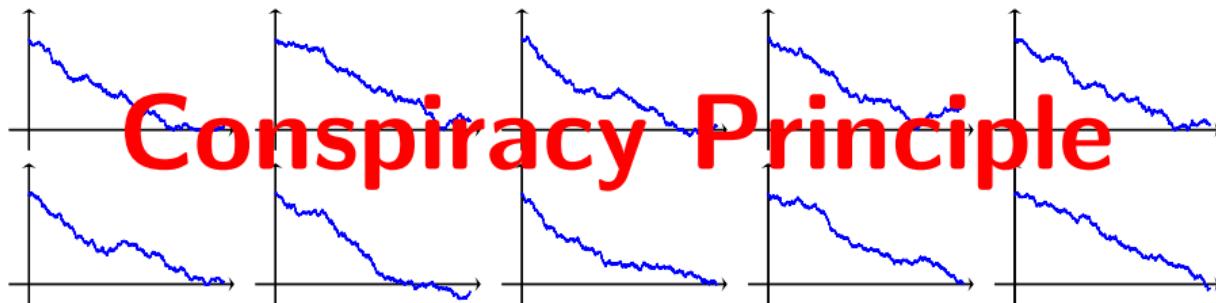


Sample paths of  $\bar{Y}_{5000}$  conditional on  $R$  for **heavy-tailed** claims:



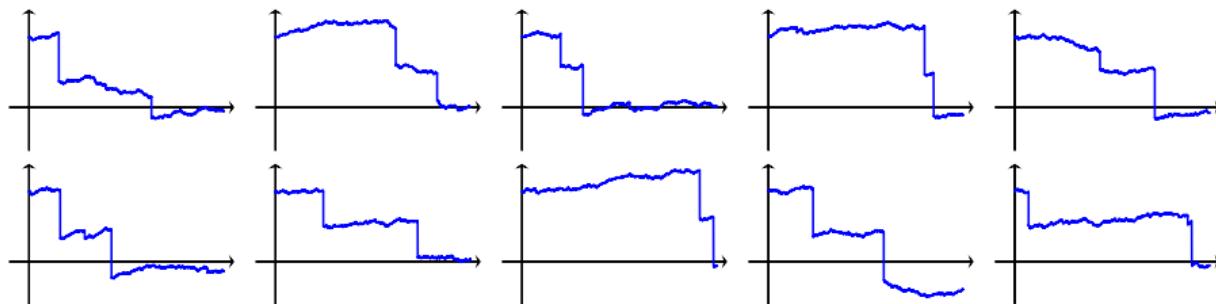
# Bankruptcy Despite of Reinsurance

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



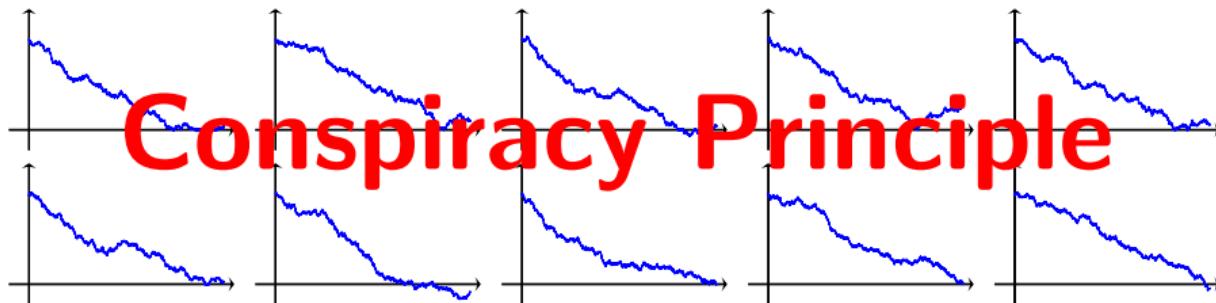
**Conspiracy Principle**

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



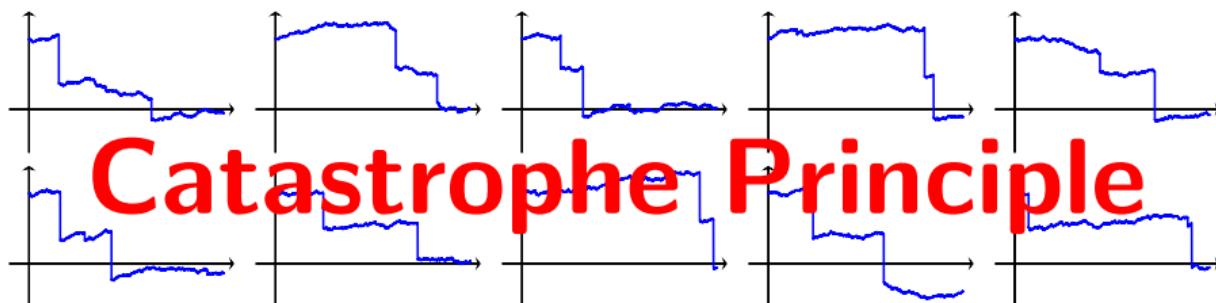
# Bankruptcy Despite of Reinsurance

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



**Conspiracy Principle**

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



**Catastrophe Principle**

**Heavy-Tailed Large Deviations:**

**Rigorous Characterization of Catastrophe Principle**

## Heavy-Tailed Large Deviations

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

# Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

For “general”  $A \subseteq \mathbb{D}$

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

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

# Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

For “general”  $A \subseteq \mathbb{D}$

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

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

# Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

For “general”  $A \subseteq \mathbb{D}$

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

LD power index ↗

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

## Implication: The Catastrophe Principle

Under certain regularity conditions on  $A$ ,

$$\mathcal{L}(\bar{S}_n | \bar{S}_n \in A) \rightarrow \mathcal{L}(\bar{S}_{|A})$$

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

## Implication: The Catastrophe Principle

Under certain regularity conditions on  $A$ ,

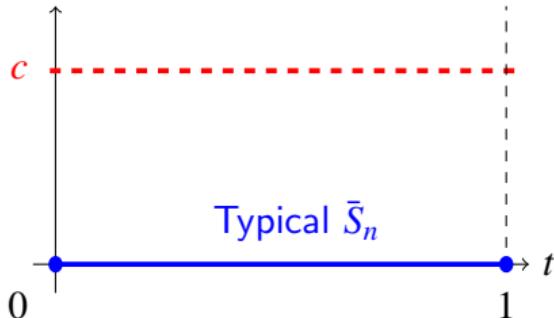
$$\mathcal{L}(\bar{S}_n | \bar{S}_n \in A) \rightarrow \mathcal{L}(\bar{S}_{|A})$$

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

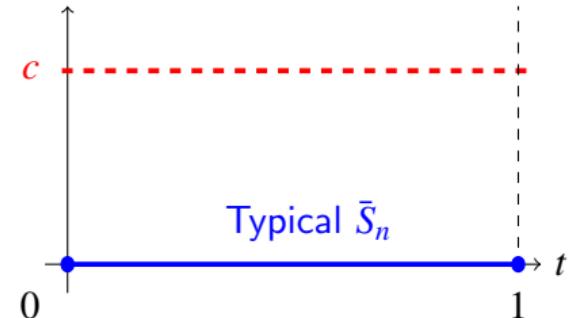
## Rigorous Characterization of Catastrophe Principle

# Conspiracy vs Catastrophe

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



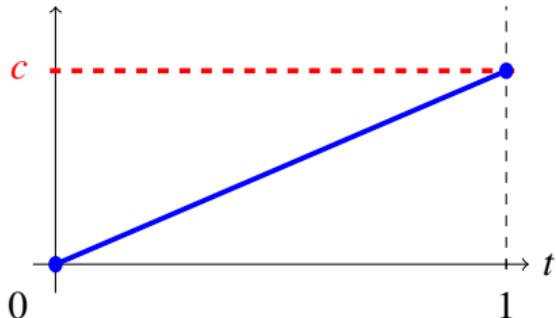
Light-Tailed Claim Size



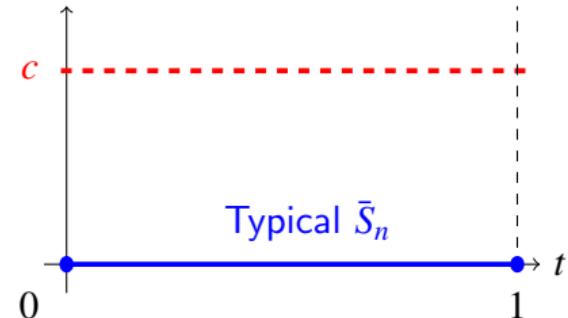
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



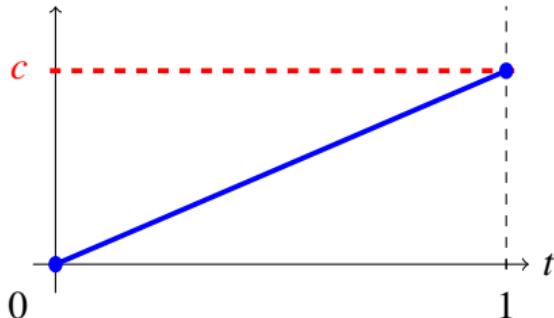
Light-Tailed Claim Size



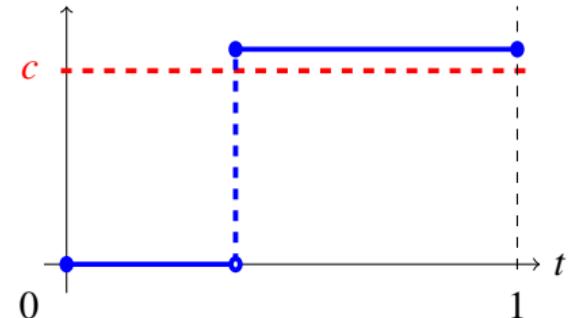
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



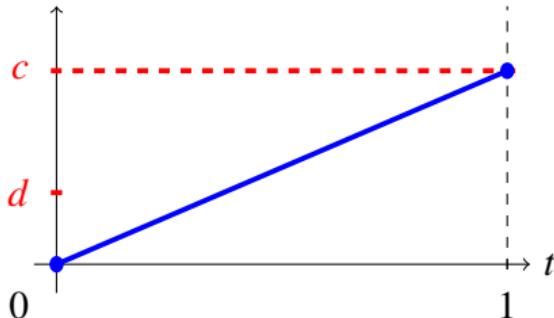
Light-Tailed Claim Size



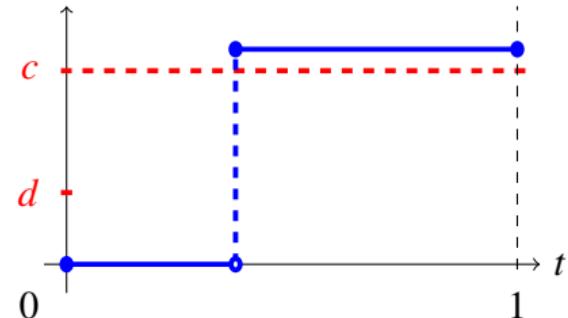
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



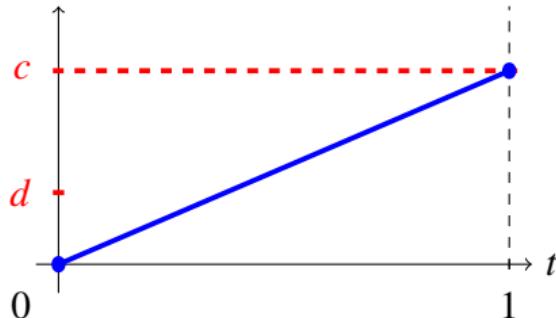
Light-Tailed Claim Size



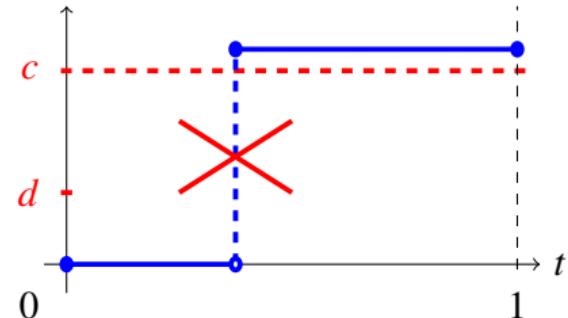
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



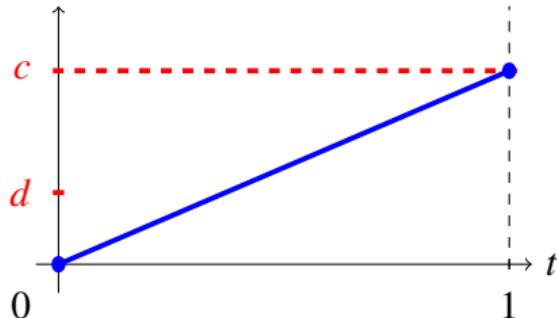
Light-Tailed Claim Size



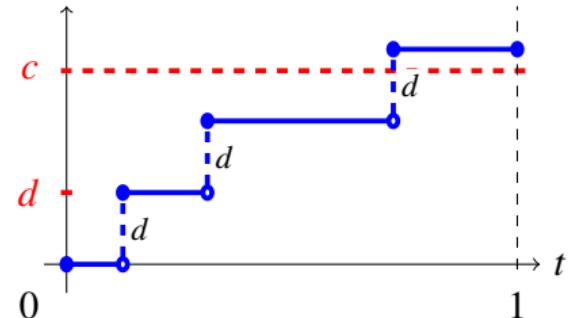
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



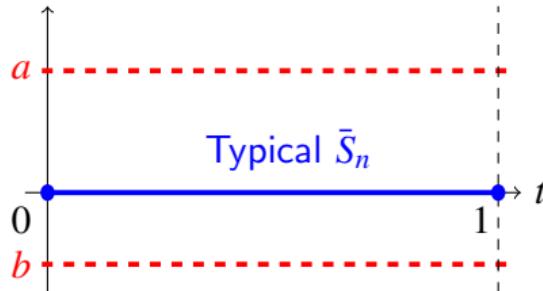
Light-Tailed Claim Size



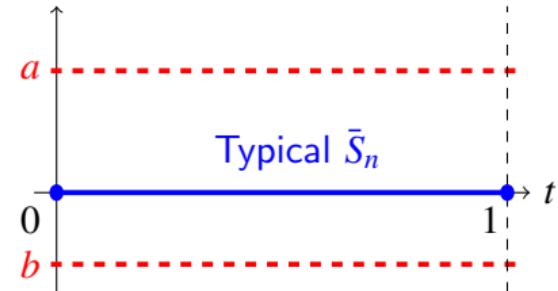
Heavy-Tailed Claim Size

# Conspiracy vs Catastrophe

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



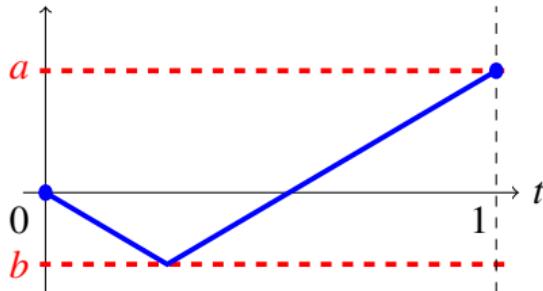
Light-Tailed Increments



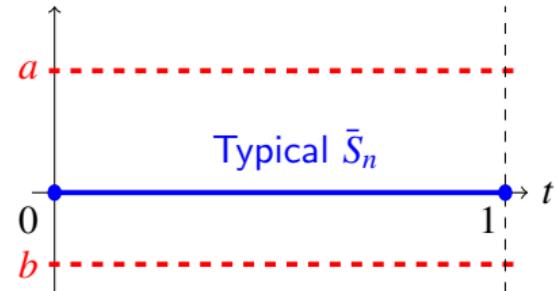
Heavy-Tailed Increments

# Conspiracy vs Catastrophe

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



Light-Tailed Increments



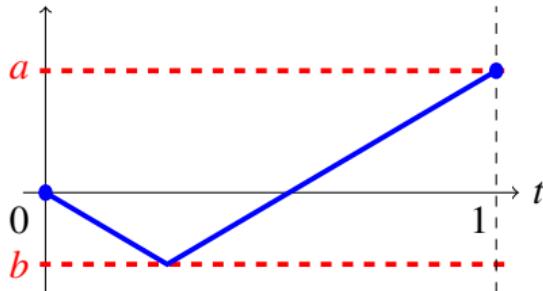
Heavy-Tailed Increments

# Conspiracy

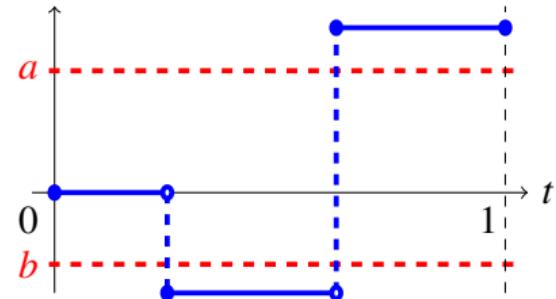
vs

# Catastrophe

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



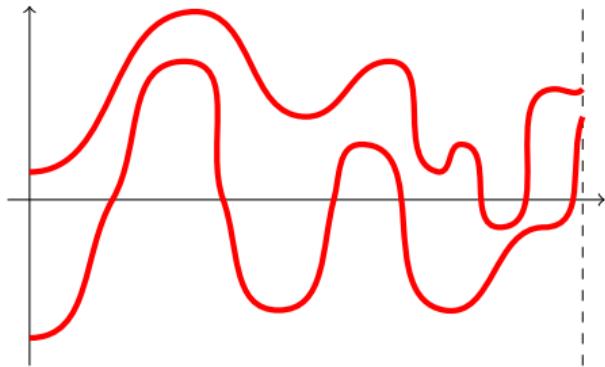
Light-Tailed Increments



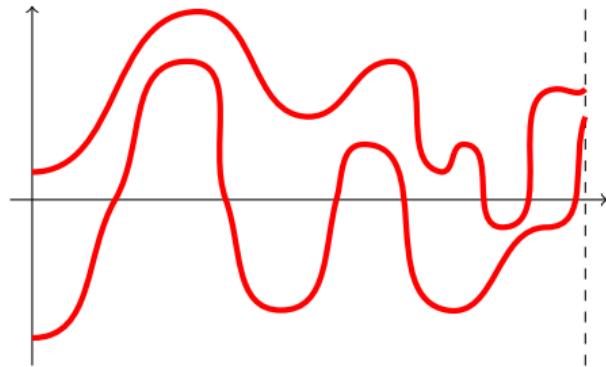
Heavy-Tailed Increments

# Conspiracy vs Catastrophe

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



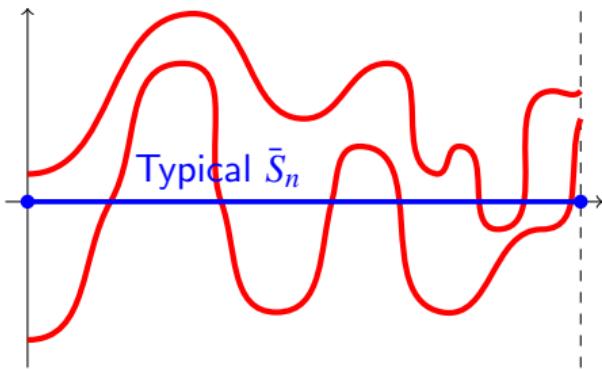
Light-Tailed Increments



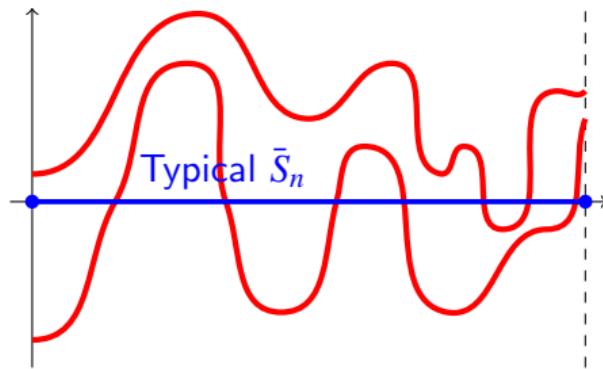
Heavy-Tailed Increments

# Conspiracy vs Catastrophe

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



Light-Tailed Increments



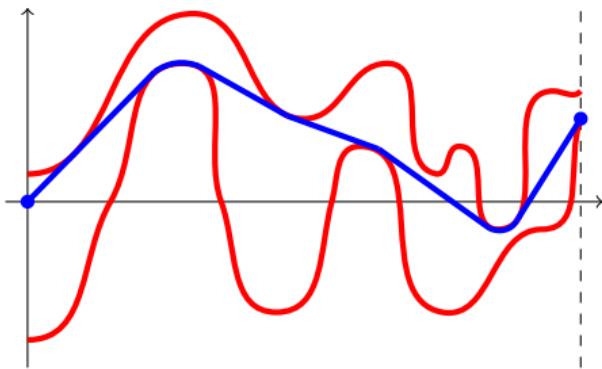
Heavy-Tailed Increments

# Conspiracy

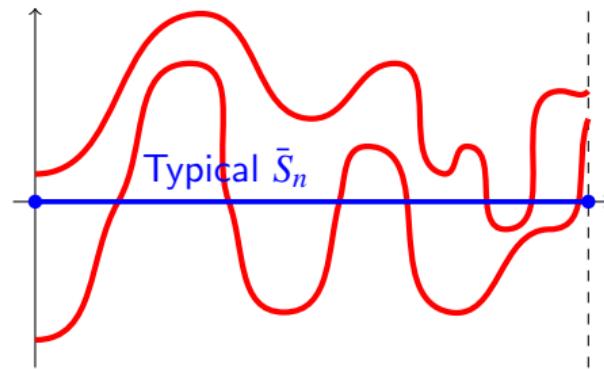
vs

# Catastrophe

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



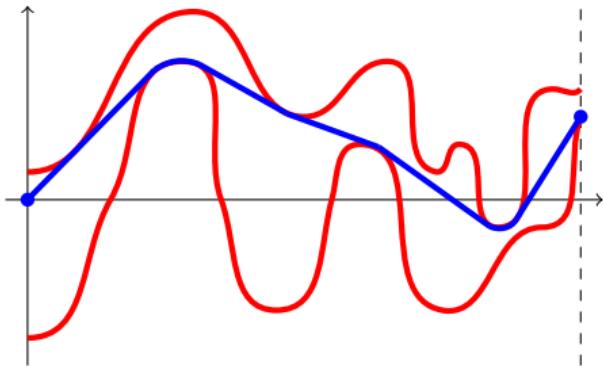
Light-Tailed Increments



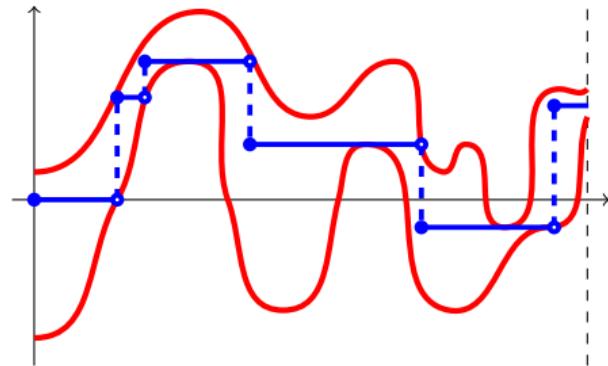
Heavy-Tailed Increments

# Conspiracy vs Catastrophe

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



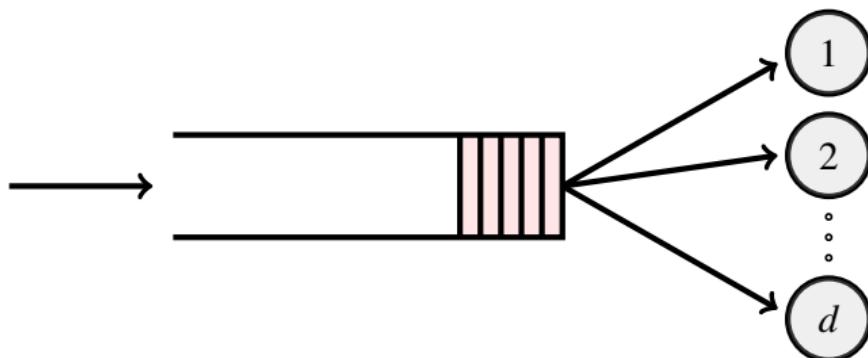
Light-Tailed Increments



Heavy-Tailed Increments

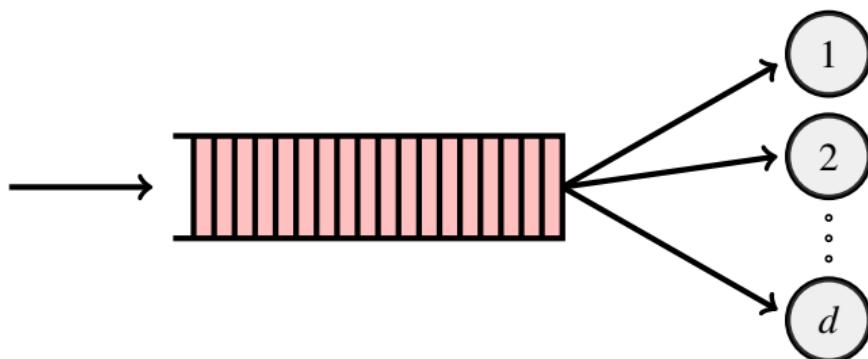
# Conspiracy vs Catastrophe

Congestion of Multiple Server Queue:



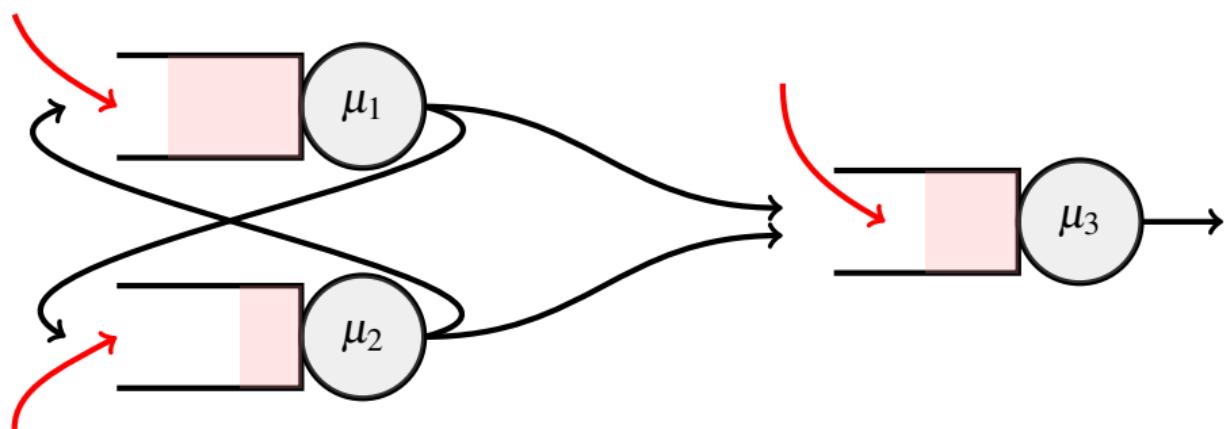
# Conspiracy vs Catastrophe

Congestion of Multiple Server Queue:



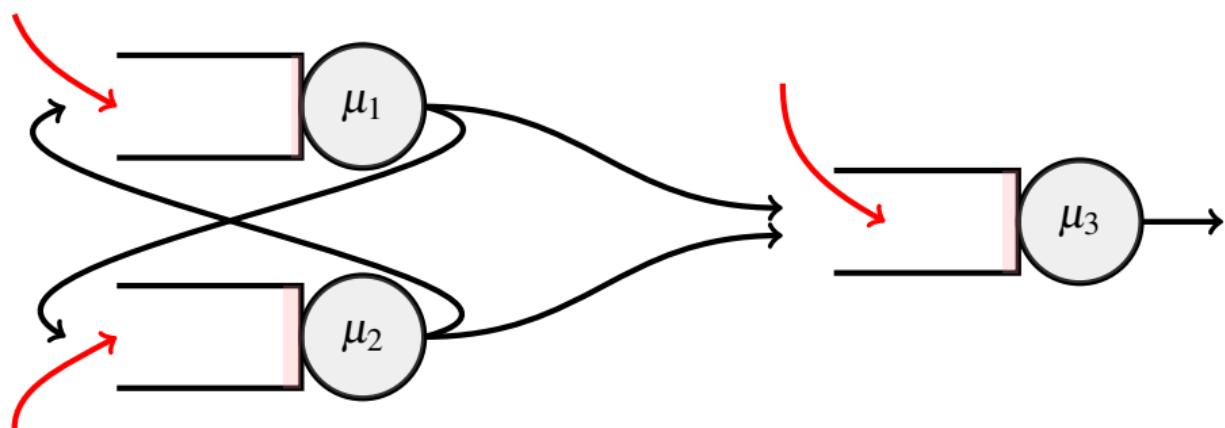
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



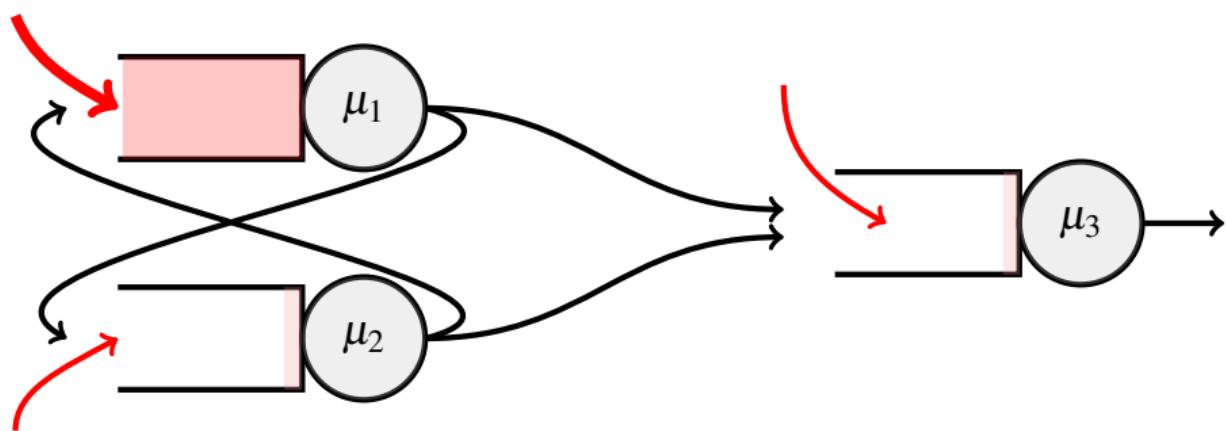
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



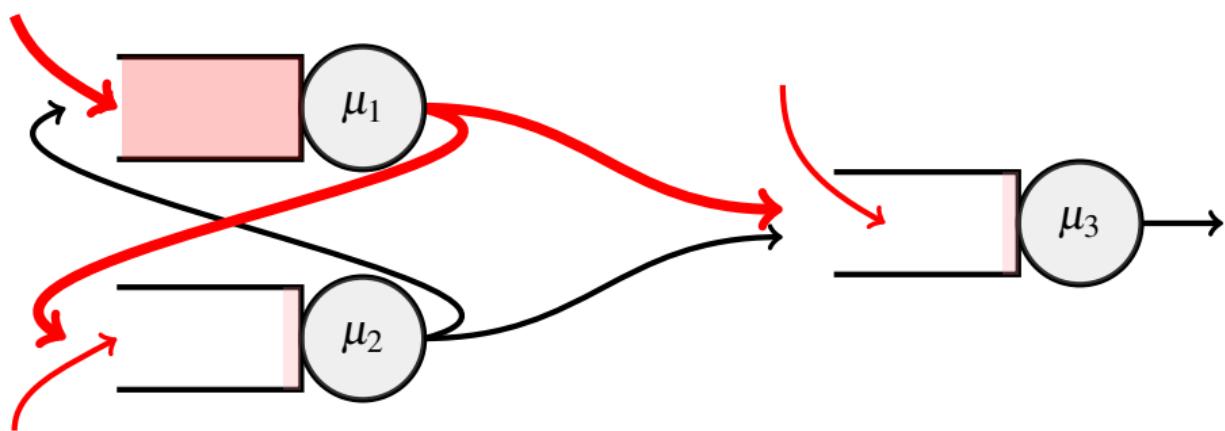
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



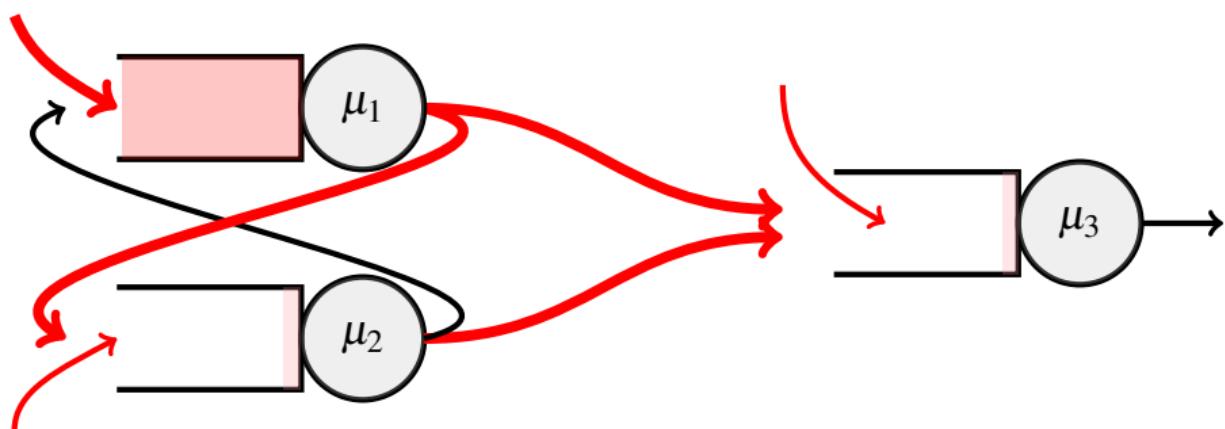
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



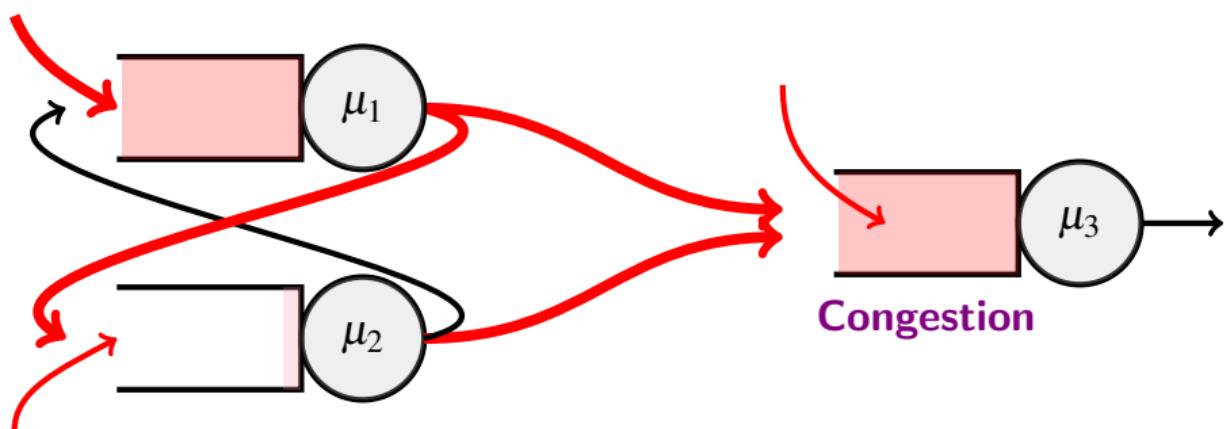
# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



# Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



The point is

Many heavy-tailed rare events can be written as

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

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

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

# Catastrophe Principle Extends to General Heavy-Tailed Systems

Heavy-Tailed Large Deviations for

- Continuous-Time Processes

R., Blanchet, Zwart (2019), Bazhba, Blanchet, R., Zwart (2020), Su, Wang, R. (2023+)

- Processes with Spatial and Temporal Correlations

Chen, R., Zwart (2023+), Bazhba, Blanchet, R., Zwart (2023+), Su, R. (2023+), Wang, R. (2023+)

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

# 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  $\mathbf{P}(\text{Edge}) \stackrel{\text{Suppose}}{\approx} 10^{-6}$

- 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  $\mathbf{P}(\text{Edge}) \stackrel{\text{Suppose}}{\approx} 10^{-6}$

- 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  $\mathbf{P}(\text{Edge}) \stackrel{\text{Suppose}}{\approx} 10^{-6}$

- 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  $\mathbf{P}(\text{Head})$**

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

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

- Construct an alternative universe
  - i.e., Consider an "importance distribution"  $\mathbf{Q}$
- Perform experiments there
  - i.e., Sample  $\mathbb{I}_{\text{Edge}}^{(1)}, \dots, \mathbb{I}_{\text{Edge}}^{(m)}$  from  $\mathbf{Q}$   
as well as the likelihood ratio  $\left(\frac{d\mathbf{P}}{d\mathbf{Q}}\right)^{(1)}, \dots, \left(\frac{d\mathbf{P}}{d\mathbf{Q}}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report  $\frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\text{Edge}}^{(i)} \left(\frac{d\mathbf{P}}{d\mathbf{Q}}\right)^{(i)}$  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  $\mathbb{I}_{\{\bar{S}_n \in A\}}^{(1)}, \dots, \mathbb{I}_{\{\bar{S}_n \in A\}}^{(m)}$  from  $\mathbf{Q}_n$   
as well as the likelihood ratio  $\left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(1)}, \dots, \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report  $\frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\{\bar{S}_n \in A\}}^{(i)} \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(i)}$  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  $\mathbb{I}_{\{\bar{S}_n \in A\}}^{(1)}, \dots, \mathbb{I}_{\{\bar{S}_n \in A\}}^{(m)}$  from  $\mathbf{Q}_n$   
as well as the likelihood ratio  $\left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(1)}, \dots, \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes
  - i.e., Report  $\underbrace{\frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\{\bar{S}_n \in A\}}^{(i)} \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(i)}}_{\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}} \text{ : IS estimator}$  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  $\mathbb{I}_{\{\bar{S}_n \in A\}}^{(1)}, \dots, \mathbb{I}_{\{\bar{S}_n \in A\}}^{(m)}$  from  $\mathbf{Q}_n$   
as well as the likelihood ratio  $\left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(1)}, \dots, \left(\frac{d\mathbf{P}}{d\mathbf{Q}_n}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes

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

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  $\mathbb{I}_{\{\bar{S}_n \in A\}} \underbrace{\frac{d\mathbf{P}}{d\mathbf{Q}_n}}$  an efficient estimator:

IS estimator

$$\frac{d\mathbf{P}}{d\mathbf{Q}_n}$$

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

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

IS estimator

$$\overbrace{\phantom{\int}}^{d\mathbf{P}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}$$

- 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  $\mathbb{I}_{\{\bar{S}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}$  an efficient estimator:

IS estimator

$$\overbrace{\phantom{\frac{d\mathbf{P}}{d\mathbf{Q}_n}}}^{d\mathbf{P}} \frac{d\mathbf{P}}{d\mathbf{Q}_n}$$

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

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

## Goal: Strongly Efficient IS Estimator

IS estimator

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

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

## Goal: Strongly Efficient IS Estimator

IS estimator

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

Error<sup>2</sup>

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

# Goal: Strongly Efficient IS Estimator

## IS estimator

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

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

# Goal: Strongly Efficient IS Estimator

## IS estimator

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

$$\text{Error}^2 \quad \quad \quad \text{Target Quantity}^2$$

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

⇒ Number of simulation runs required remains bounded.

# Goal: Strongly Efficient IS Estimator

## IS estimator

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

$$\text{Error}^2 \quad \quad \quad \text{Target Quantity}^2$$

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

⇒ 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

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

$\uparrow$

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

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

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

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

$\uparrow$   $\frac{d\mathbf{P}}{d\mathbf{Q}_n}$  not too big       $\rightarrow$   $\mathbf{Q}_n(\cdot)$  close to  $\mathbf{P}(\cdot | \bar{X}_n \in A)$

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

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\}$
- Fix  $w \in (0, 1)$  and define  $\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot | \bar{X}_n \in B^\gamma)$   
 $\frac{d\mathbf{P}}{d\mathbf{Q}_n}$  not too big  
 $\mathbf{Q}_n(\cdot)$  close to  $\mathbf{P}(\cdot | \bar{X}_n \in A)$

# 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  $\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot | \bar{X}_n \in B^\gamma)$   
 $\frac{d\mathbf{P}}{d\mathbf{Q}_n}$  not too big  
 $\mathbf{Q}_n(\cdot)$  close to  $\mathbf{P}(\cdot | \bar{X}_n \in A)$

# 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  $\uparrow$   $\mathbf{Q}_n(\cdot)$  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

- $B^\gamma \triangleq \{\text{paths w/ at least } \mathcal{J}(A) \text{ jumps of size } > \gamma\} \Rightarrow \mathcal{J}(A) = \mathcal{J}(B^\gamma)$
- Fix  $w \in (0, 1)$  and define  $\mathbf{Q}_n(\cdot) \triangleq w\mathbf{P}(\cdot) + (1-w)\mathbf{P}(\cdot | \bar{X}_n \in B^\gamma)$ 
  - $\frac{d\mathbf{P}}{d\mathbf{Q}_n}$  not too big
  - $\mathbf{Q}_n(\cdot)$  close to  $\mathbf{P}(\cdot | \bar{X}_n \in A)$
- If we choose  $\gamma$  so that  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

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

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

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

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

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

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

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

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

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

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

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

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

# 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  $\mathcal{J}(A \setminus B^\gamma) \geq \mathcal{J}(A) + \mathcal{J}(B^\gamma)$

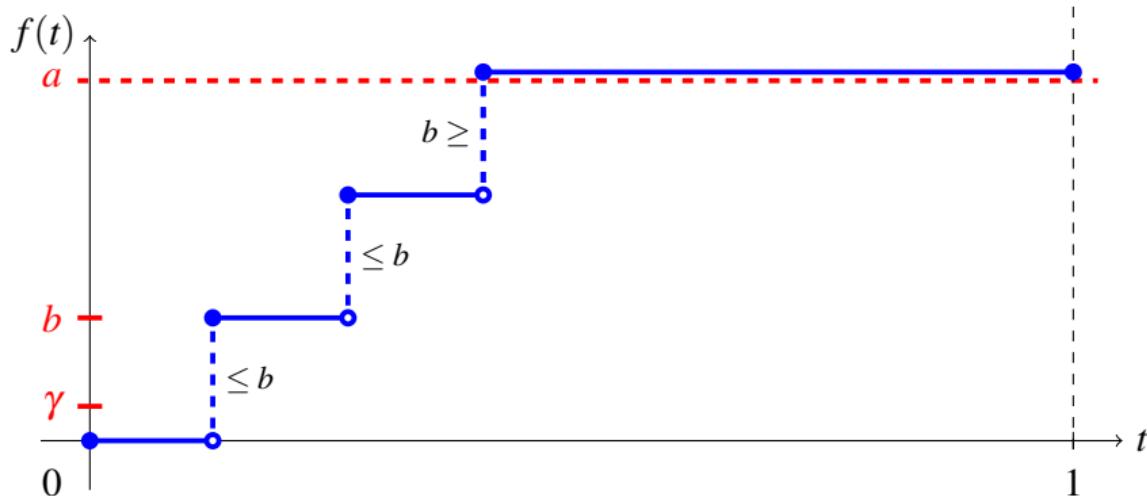
Error<sup>2</sup> →  $E^{\mathbf{Q}_n} \left( \mathbb{1}_{\{\bar{X}_n \in A\}} \frac{d\mathbf{P}}{d\mathbf{Q}_n} \right)^2 \leq \frac{1}{w} \mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A) \cdot \mathbf{P}(\bar{X}_n \in B^\gamma)$

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

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

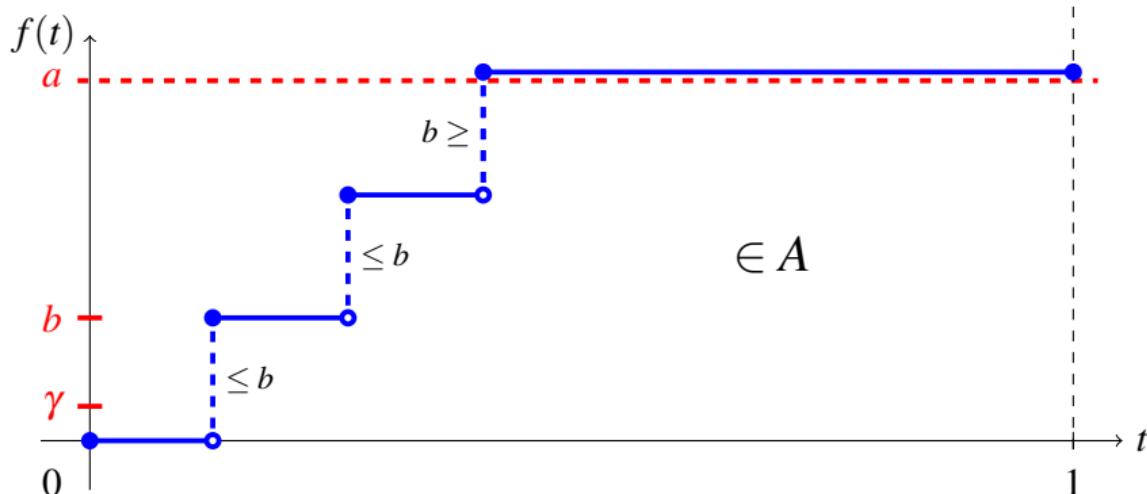
Chen, Blanchet, R., and Zwart (2019)

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



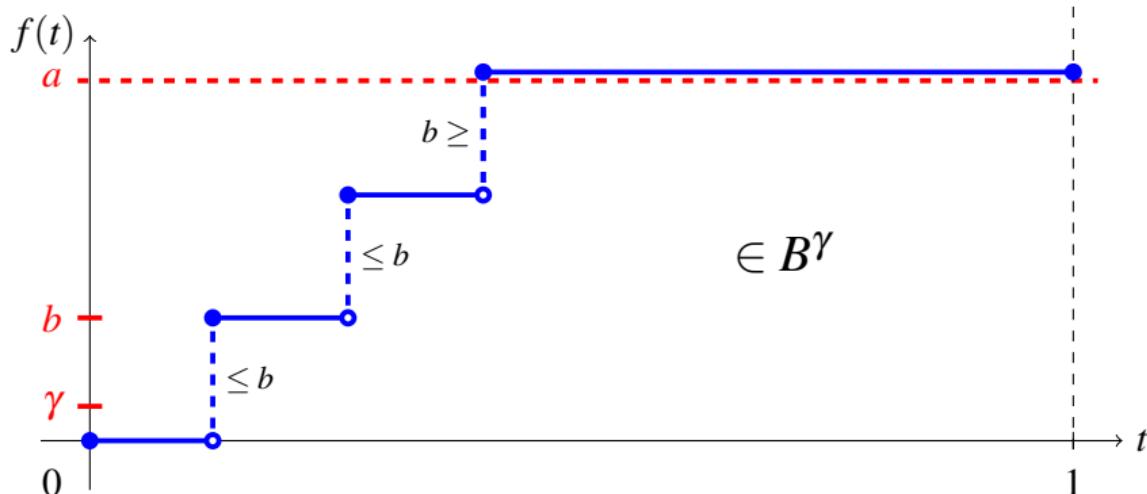
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



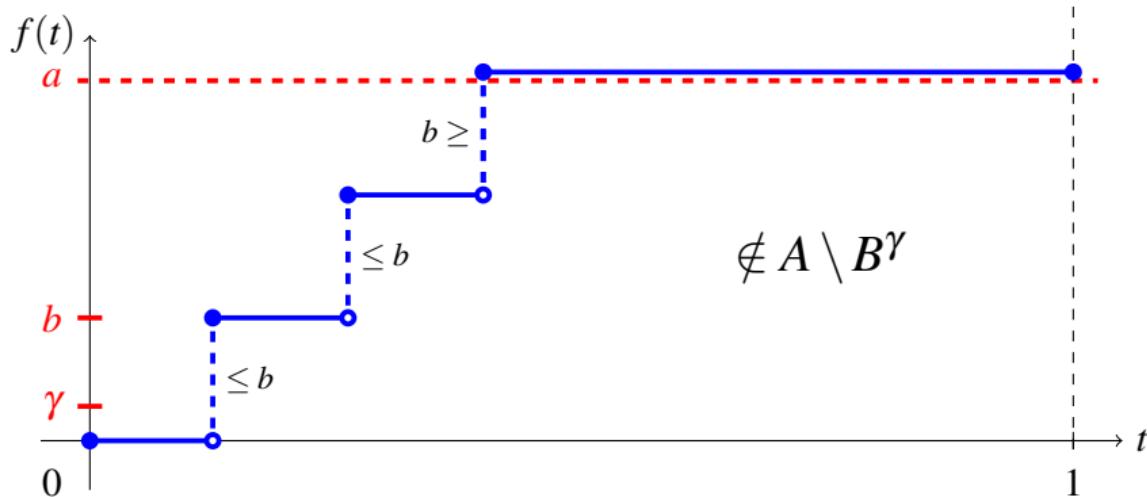
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



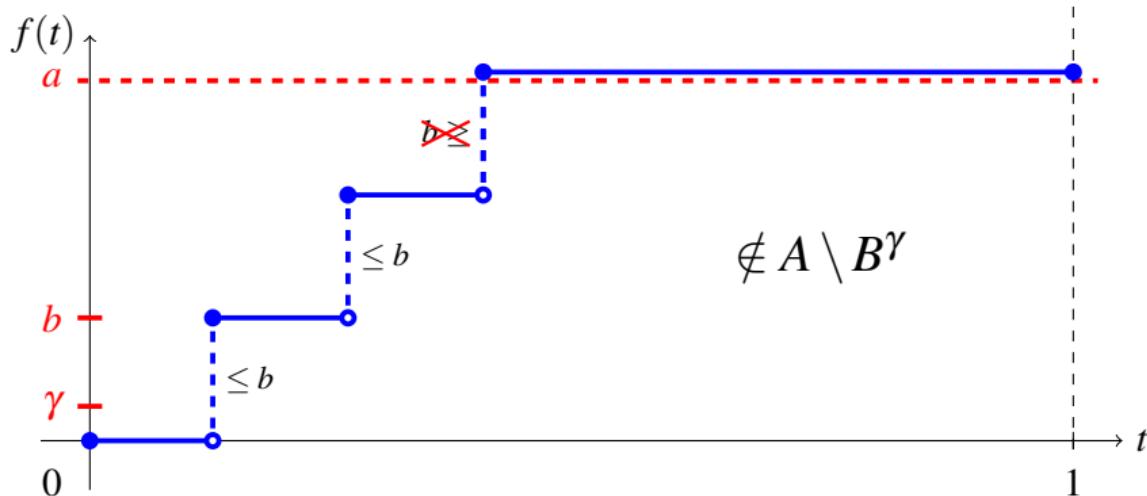
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



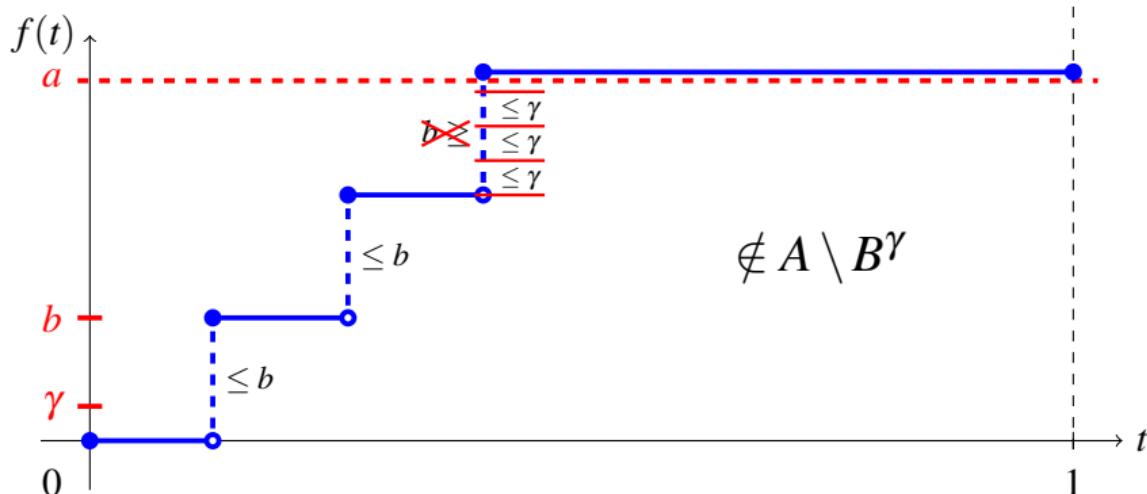
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



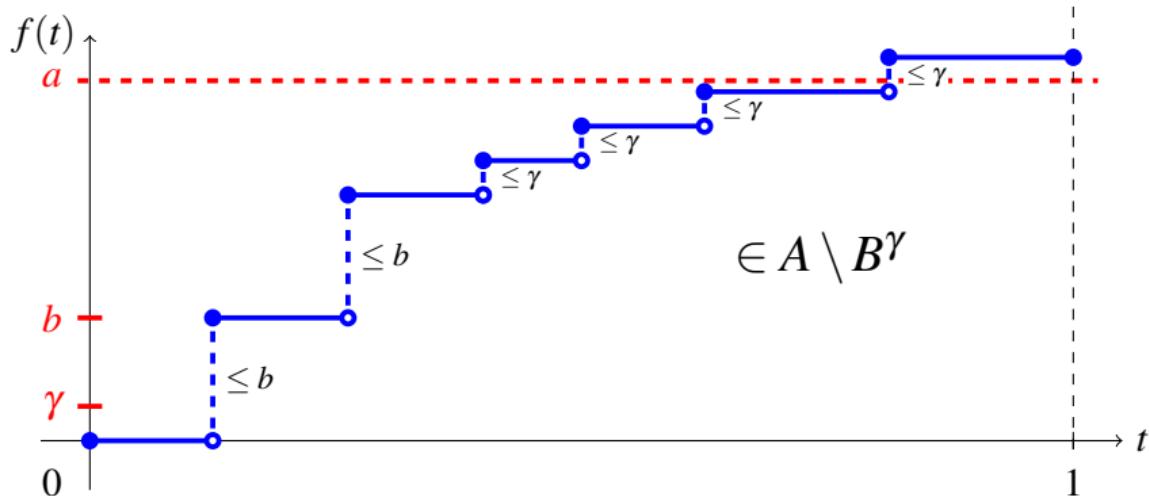
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



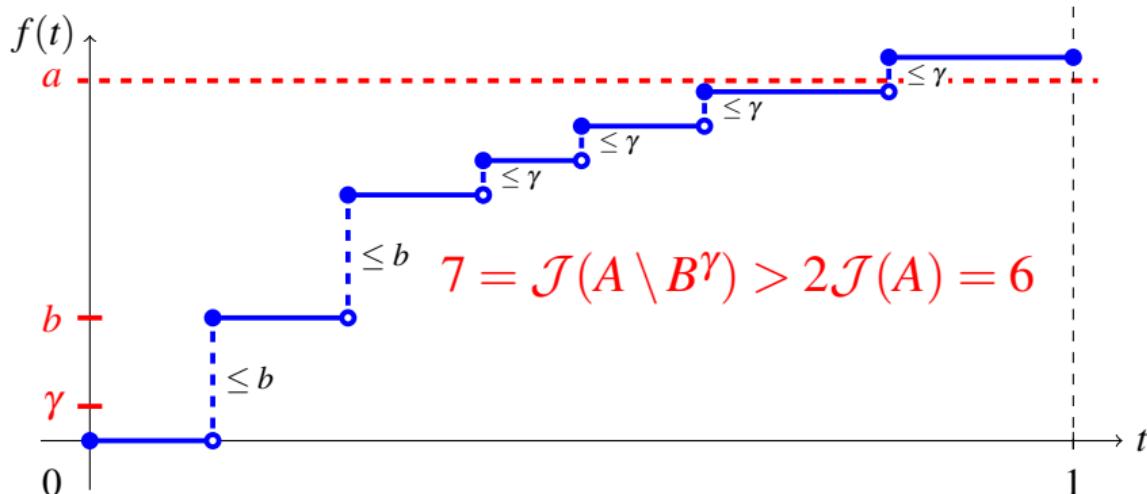
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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



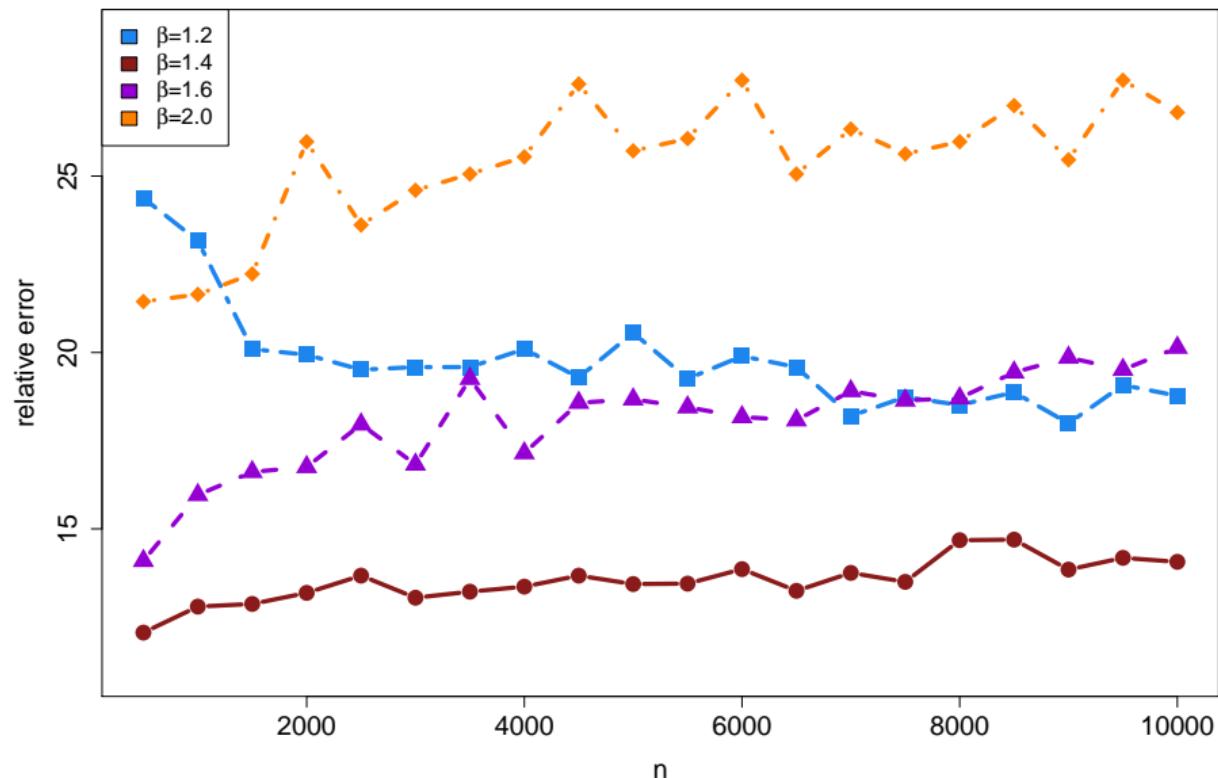
- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

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

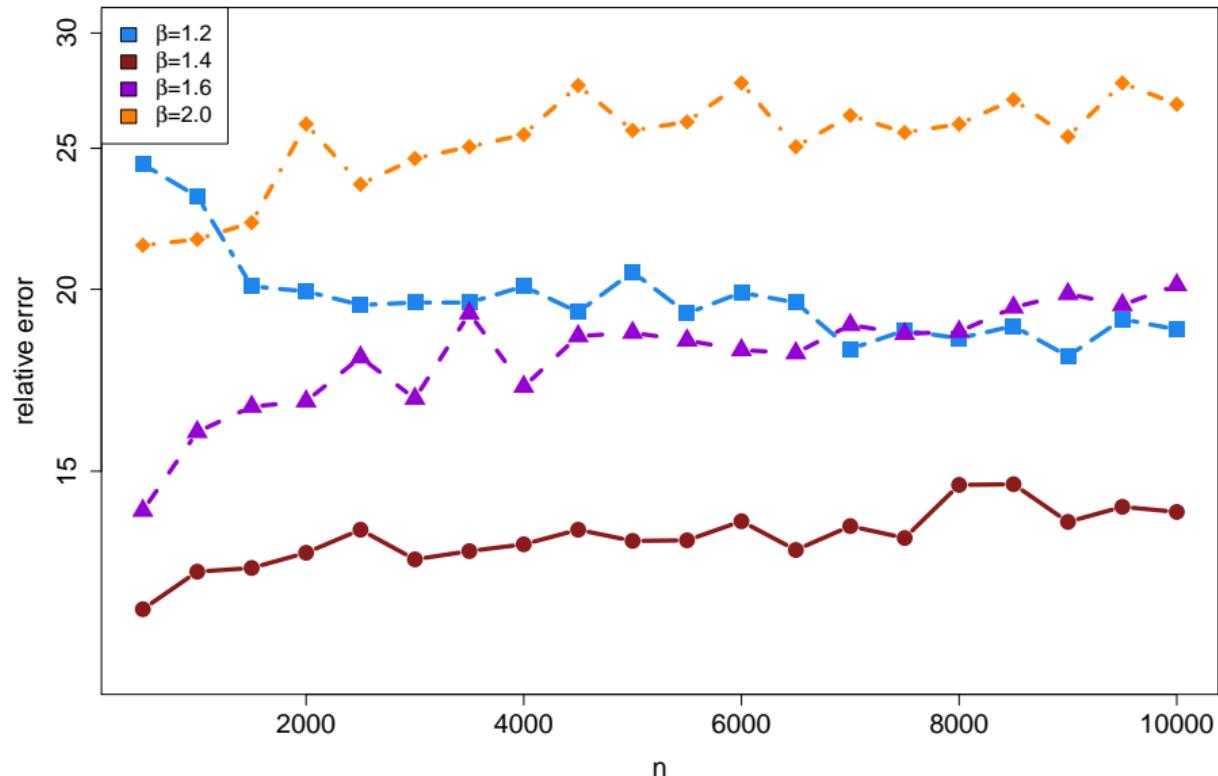


- $A = \{\text{paths that cross level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $B^\gamma = \{\text{paths with at least 3 jumps of size } > \gamma\}$
- $A \setminus B^\gamma = \{\text{paths that cross level } a \text{ on } [0, 1]$   
  & all jump sizes  $\leq b$  & at most 2 jumps of size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

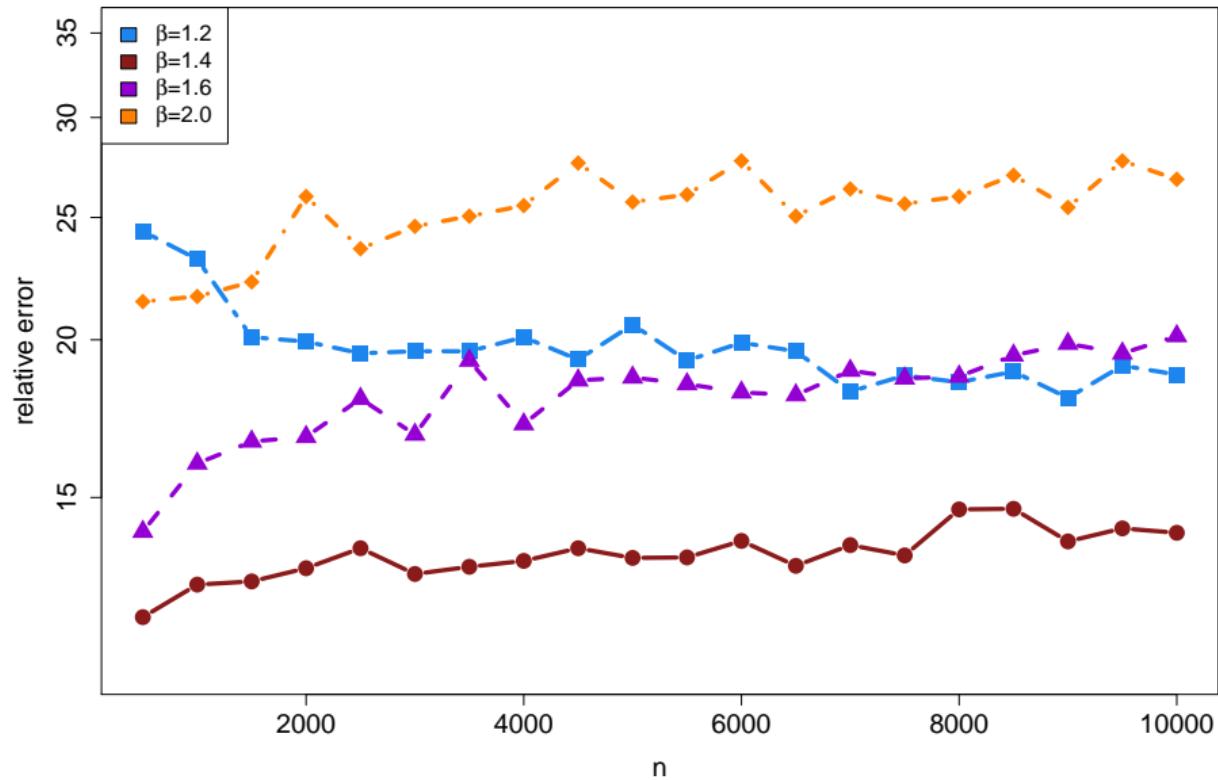
# Numerical Experiments for Reinsurance Example



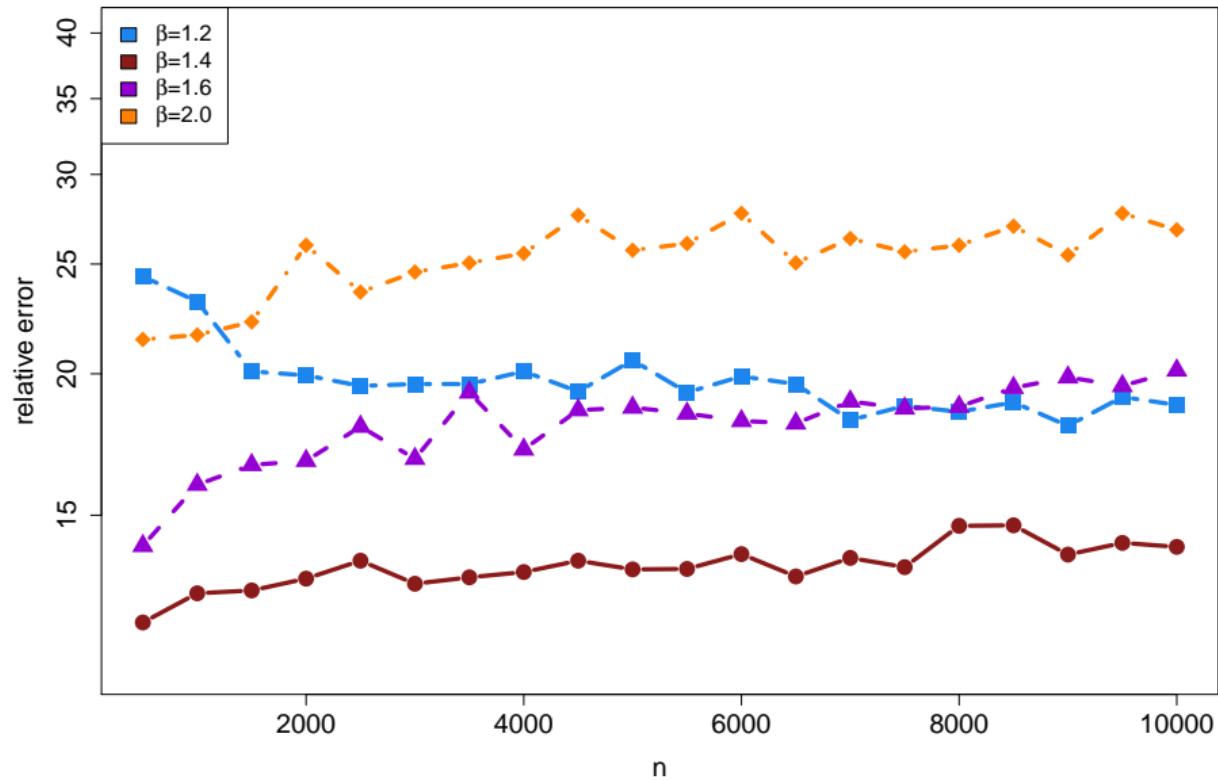
## Numerical Results for Reinsurance Example



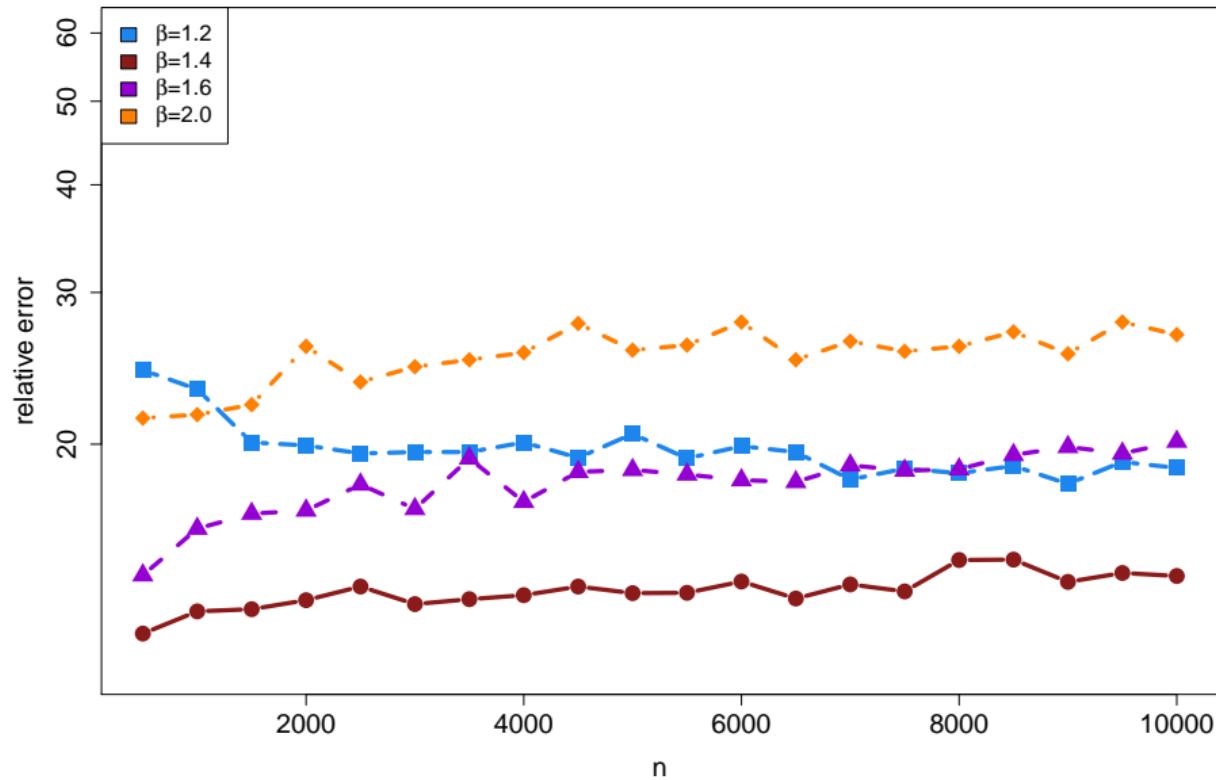
## Numerical Results for Reinsurance Example



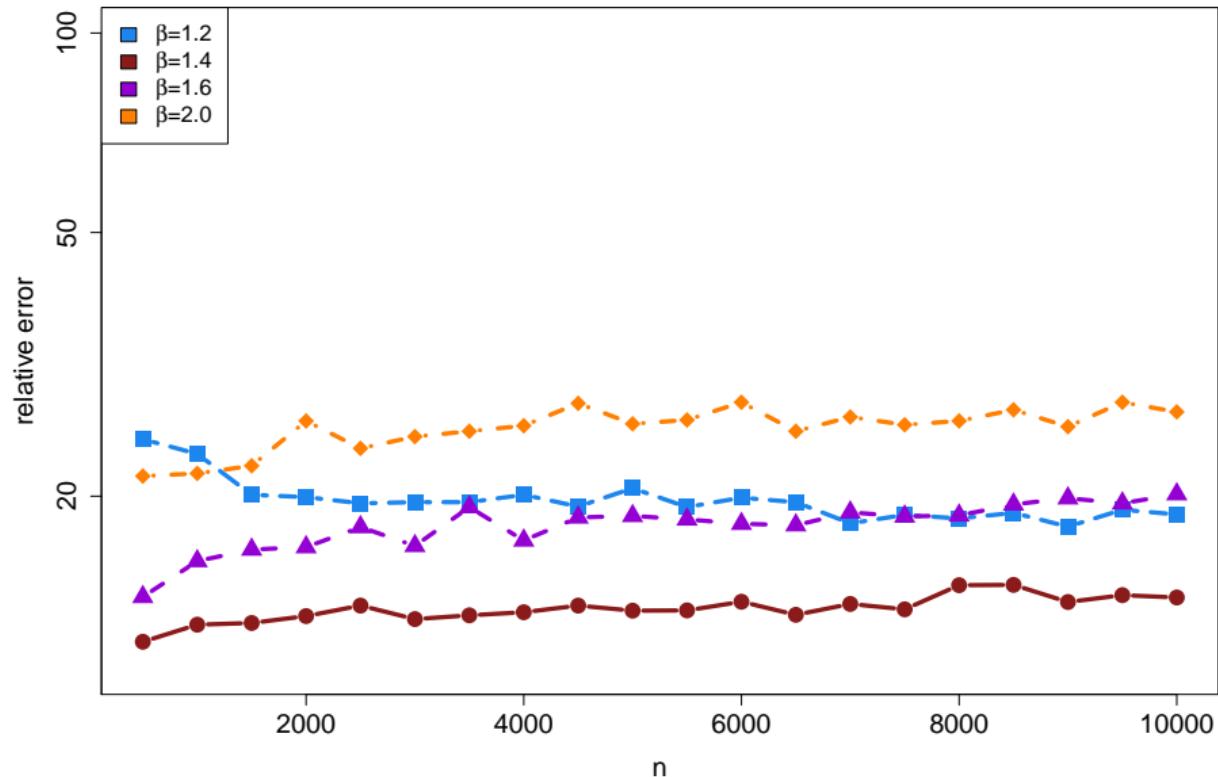
## Numerical Results for Reinsurance Example



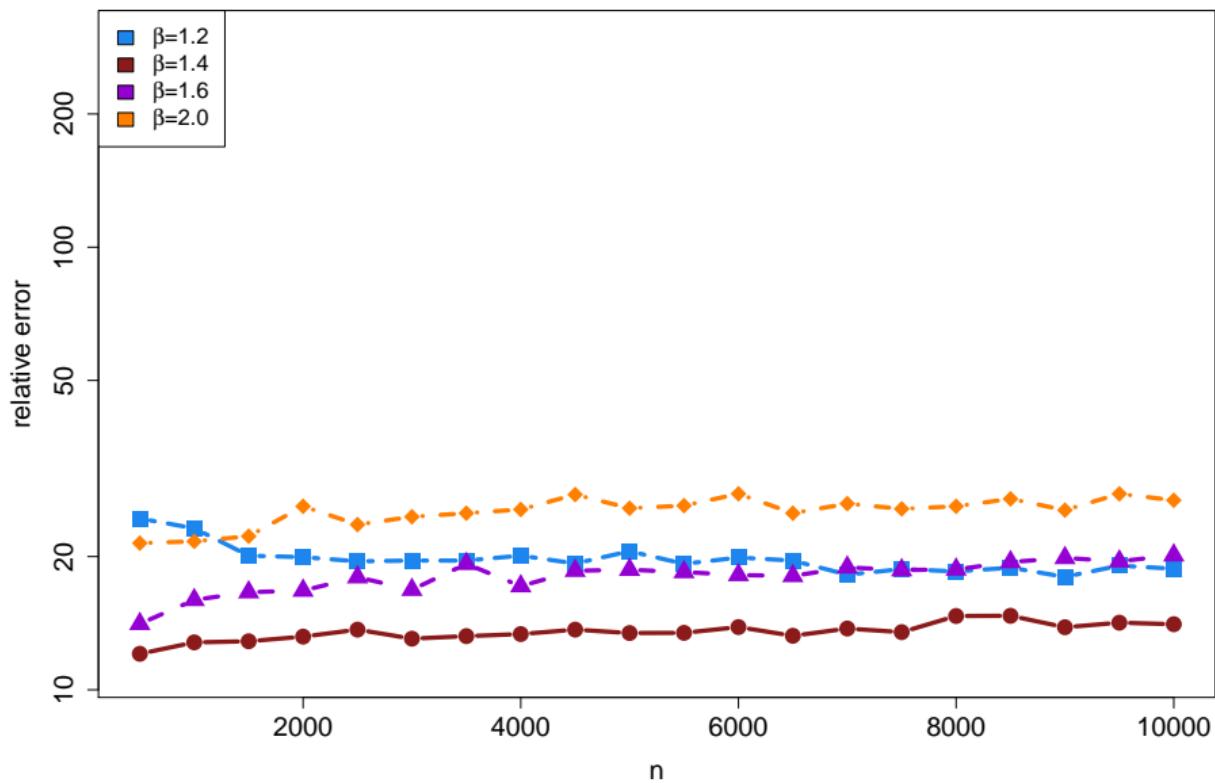
## Numerical Results for Reinsurance Example



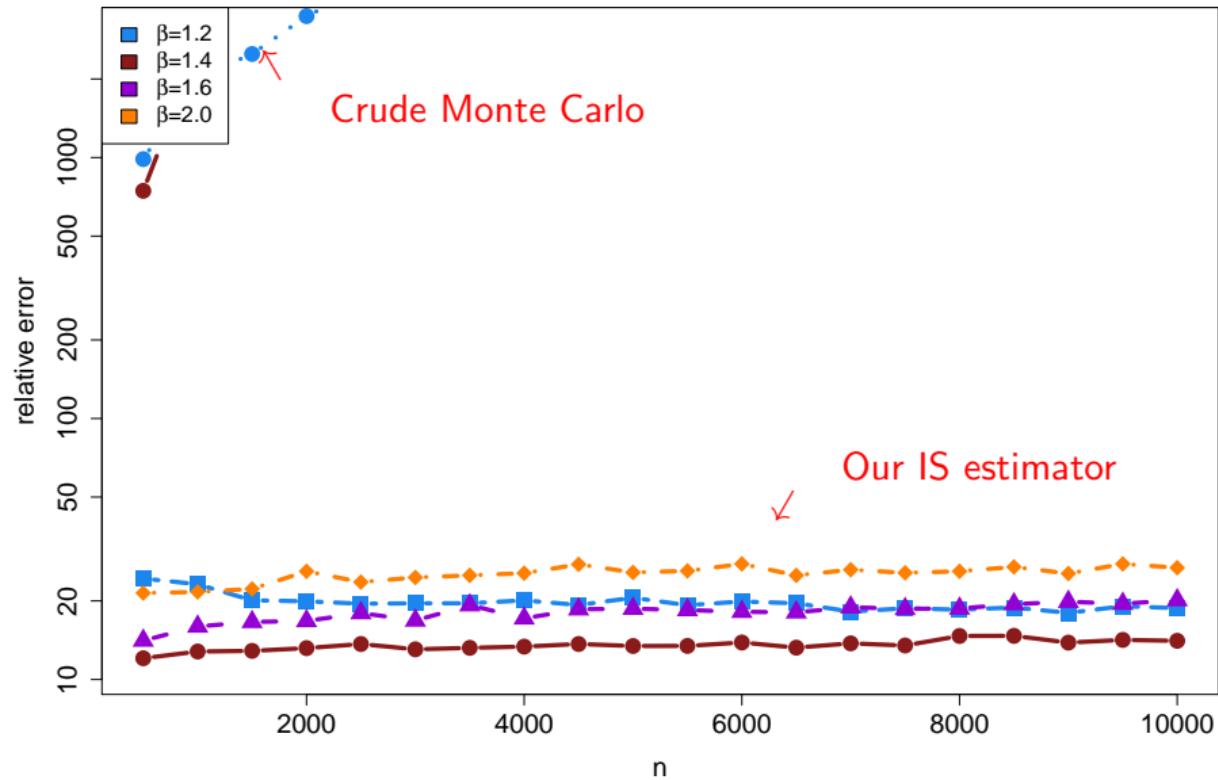
## Numerical Results for Reinsurance Example



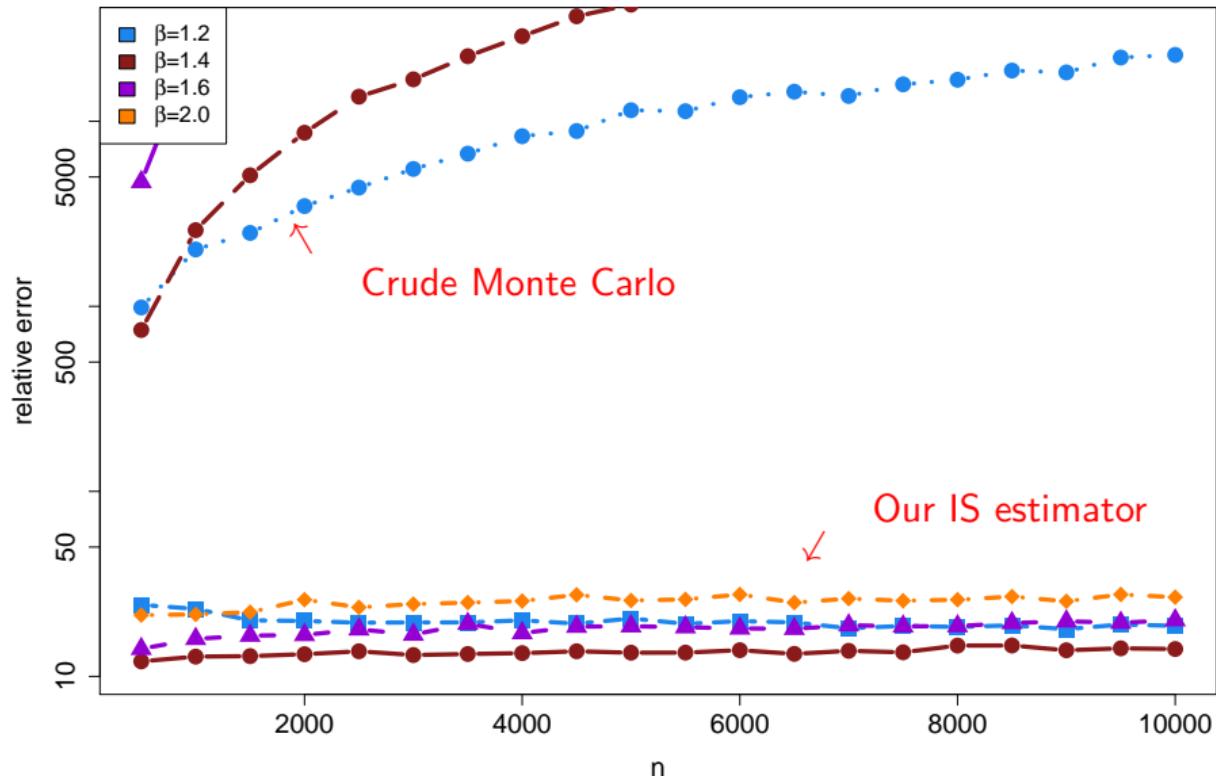
## Numerical Results for Reinsurance Example



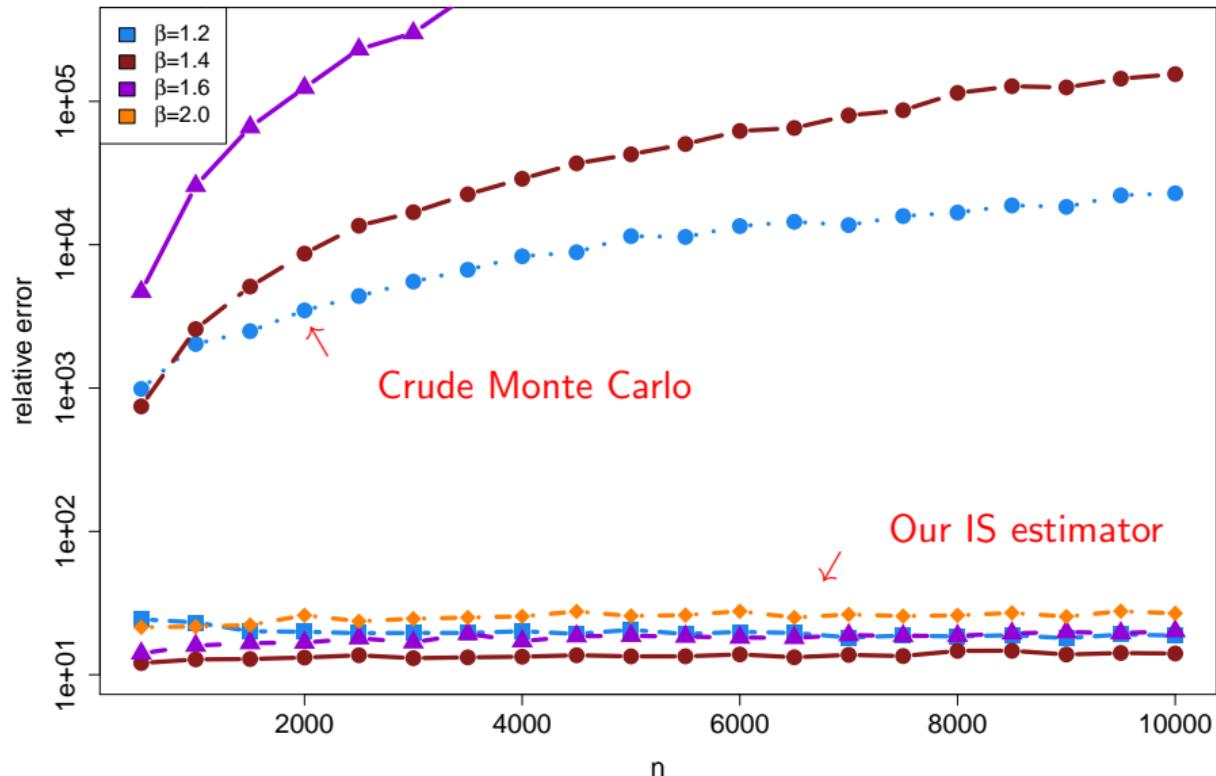
# Numerical Results for Reinsurance Example



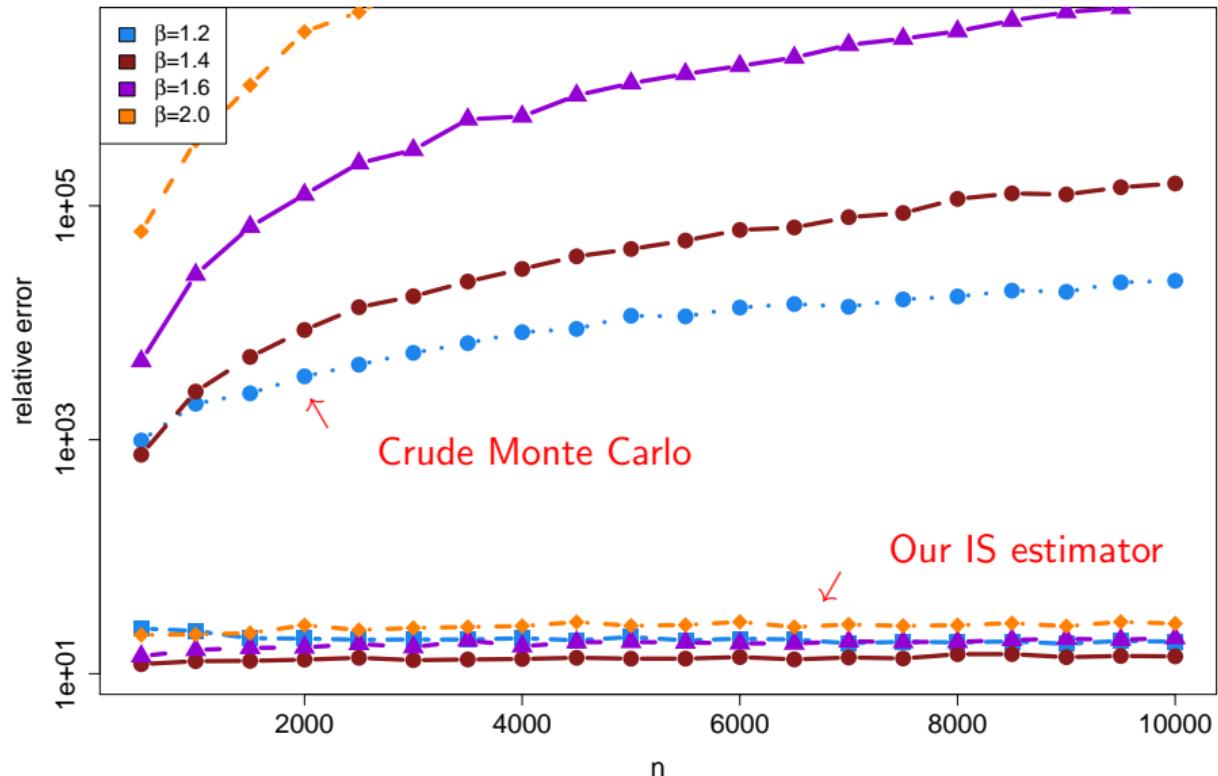
# Numerical Results for Reinsurance Example



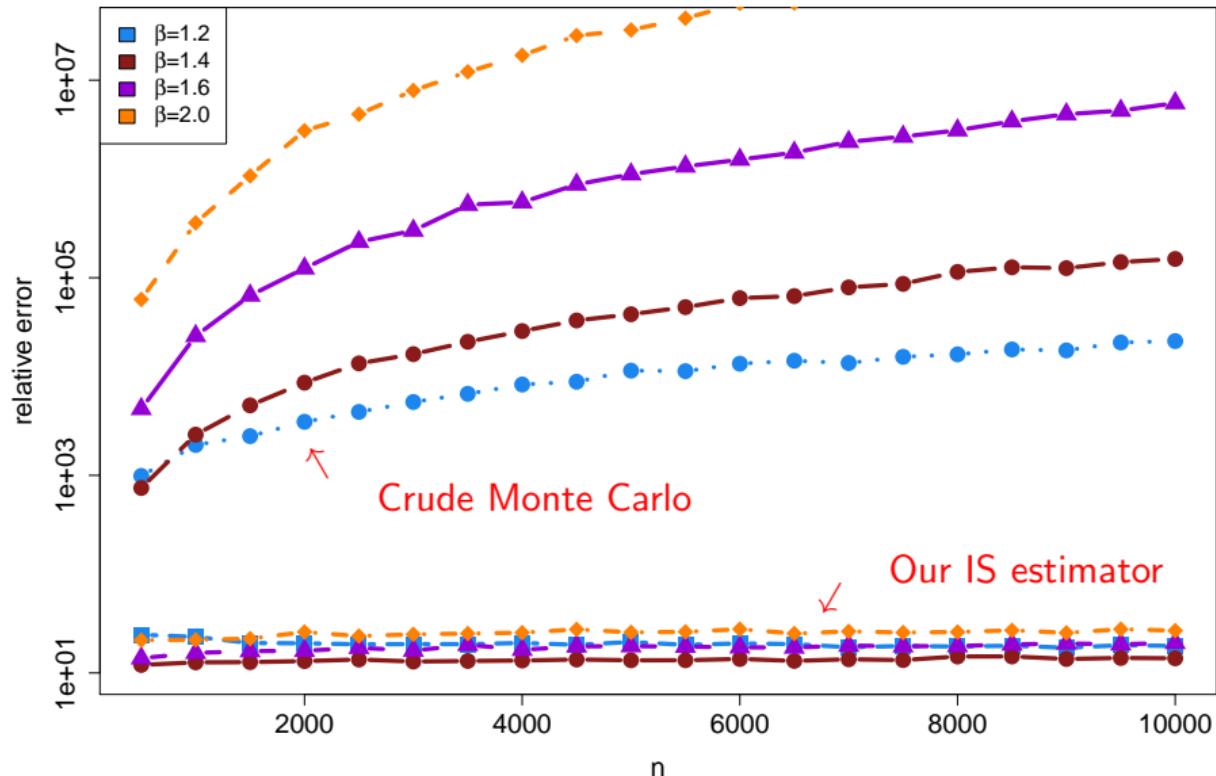
# Numerical Results for Reinsurance Example



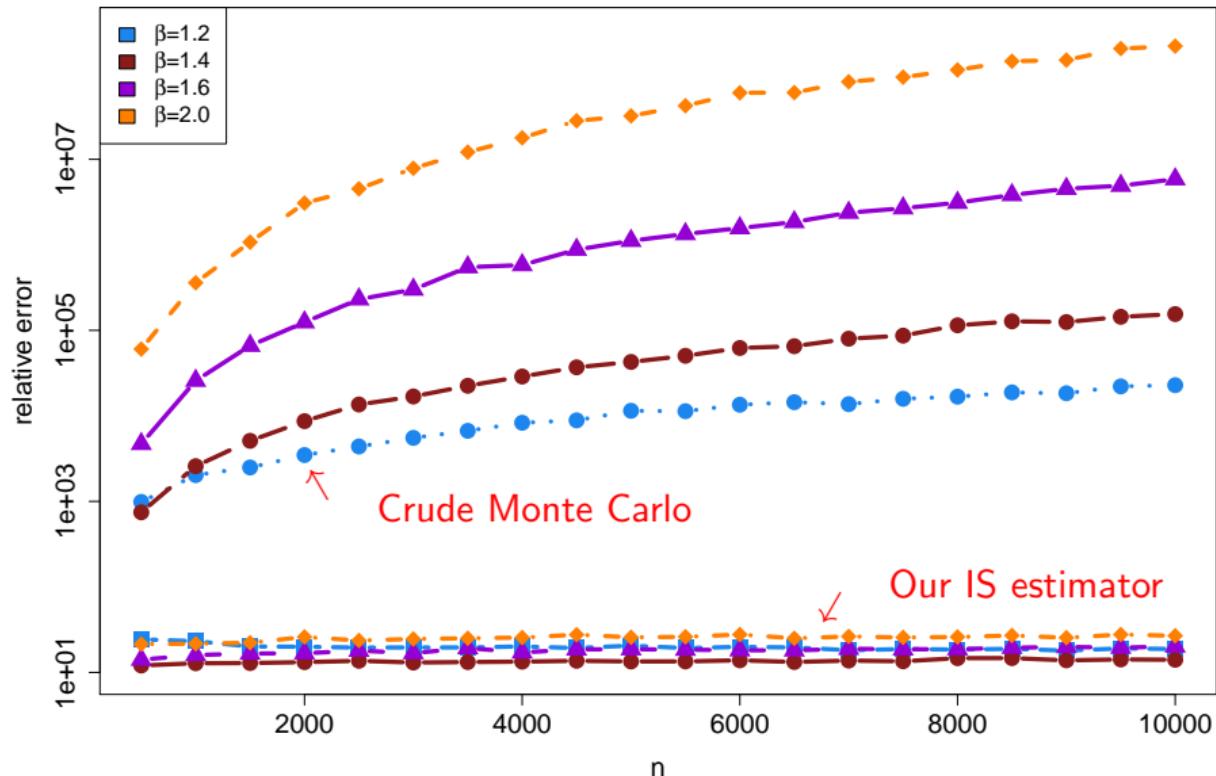
# Numerical Results for Reinsurance Example



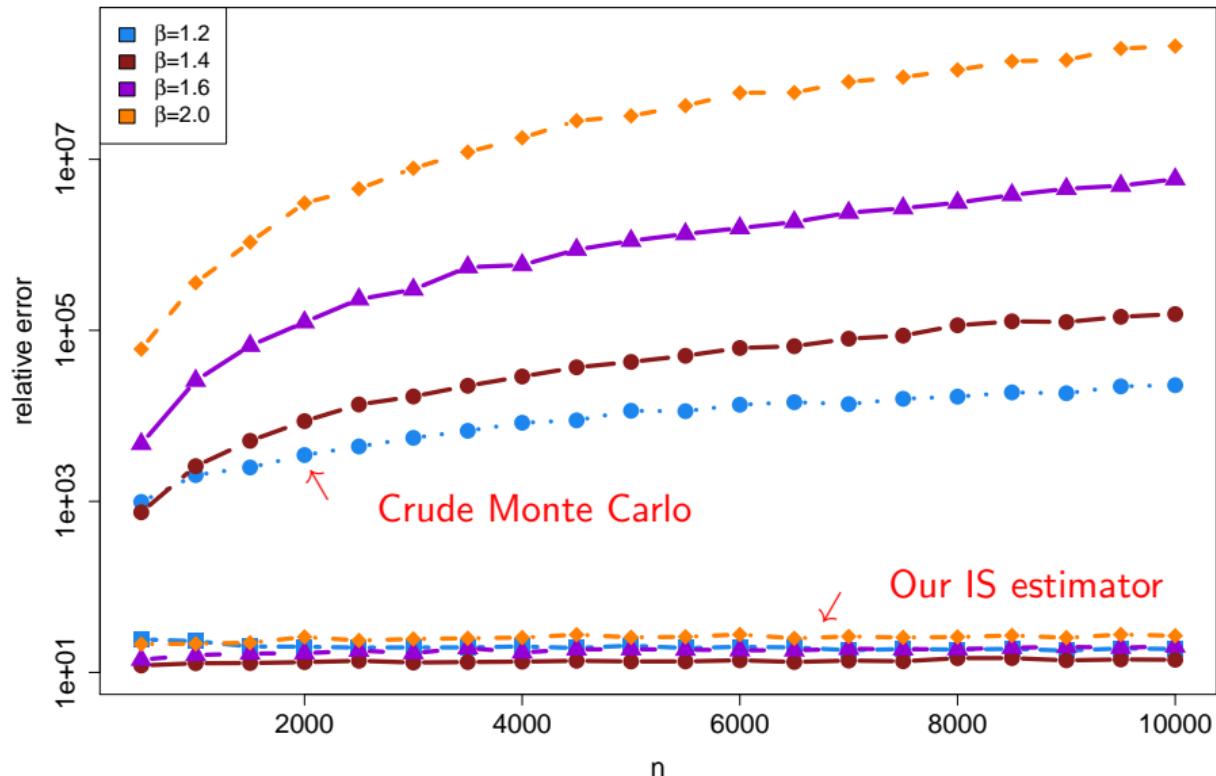
# Numerical Results for Reinsurance Example



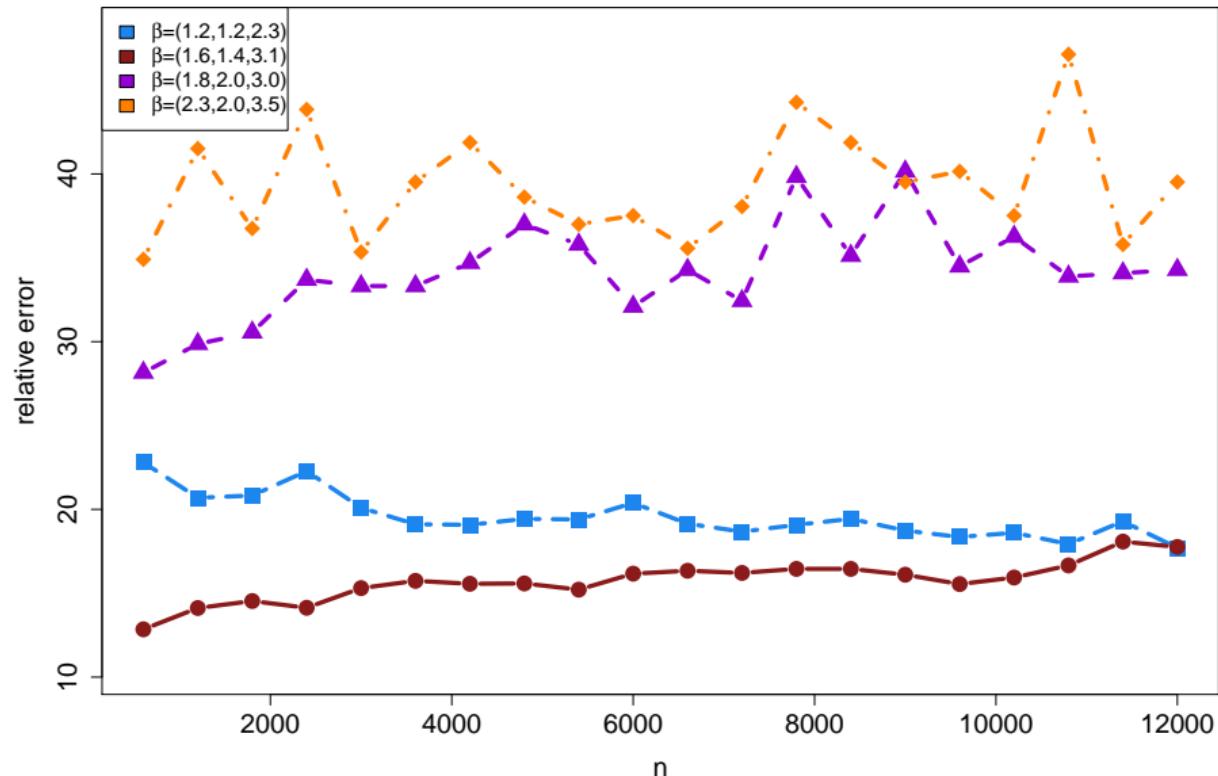
# Numerical Results for Reinsurance Example



# Numerical Results for Reinsurance Example



# Numerical Results for Fluid Network Example



## Large Deviations for Weibull Tails

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

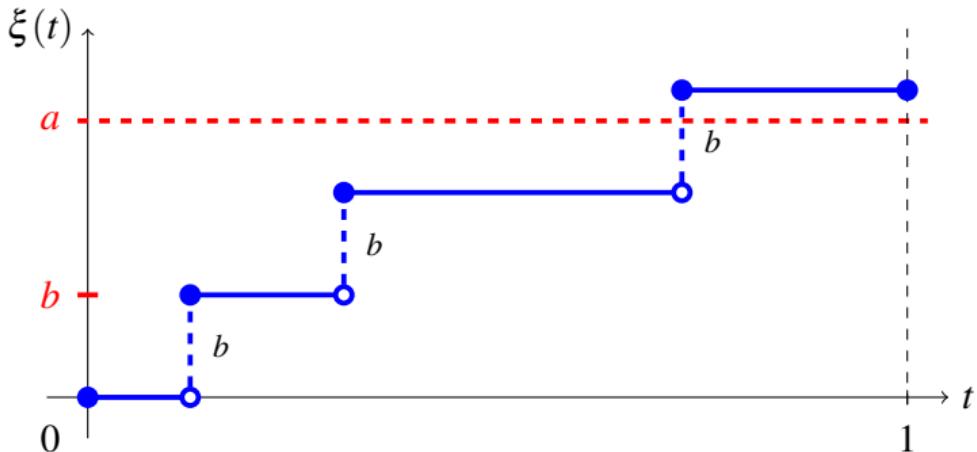
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $\mathbf{P}(\bar{S}_n \in A) \gtrsim \exp(-L(n)n^\alpha I(A^\circ))$ ,  $I(A^\circ) = \inf_{\xi \in A^\circ} I(\xi)$
- $I(\xi) = \begin{cases} \sum_{\{t: \xi(t) \neq \xi(t-)\}} (\xi(t) - \xi(t-))^\alpha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

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

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

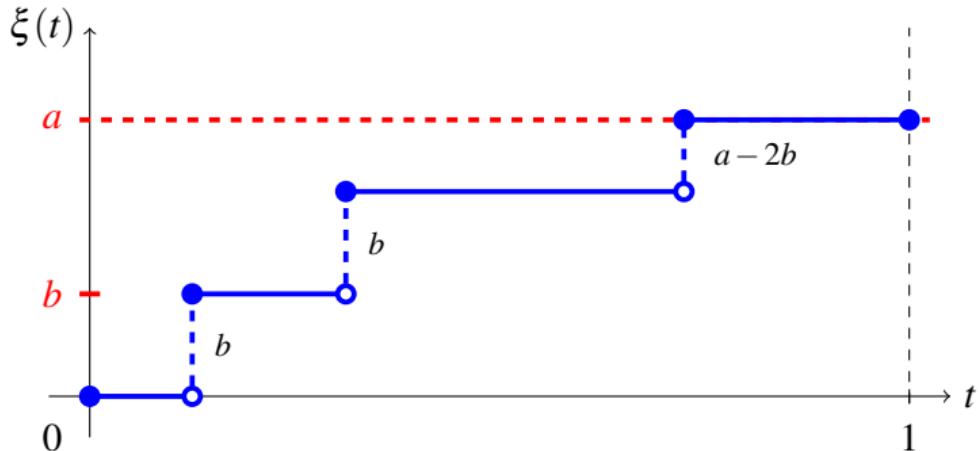
**BUT**

## $L_1$ topology is weak



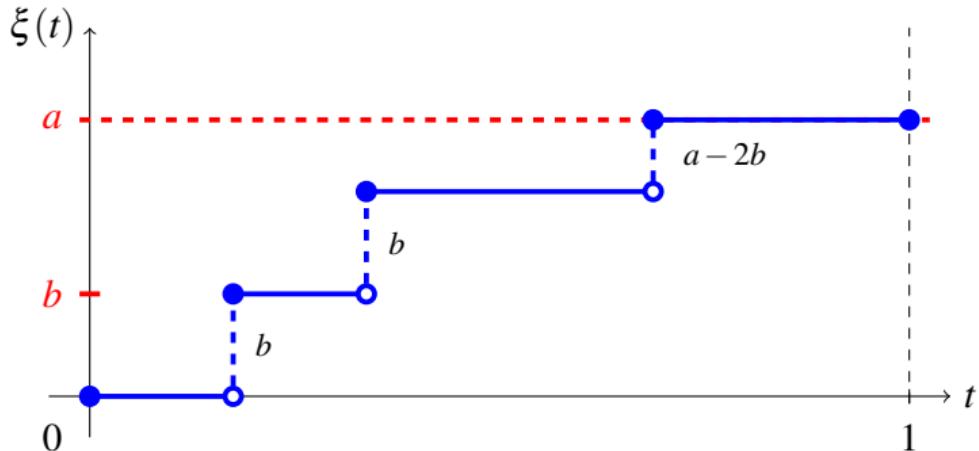
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + b^\alpha$

## $L_1$ topology is weak



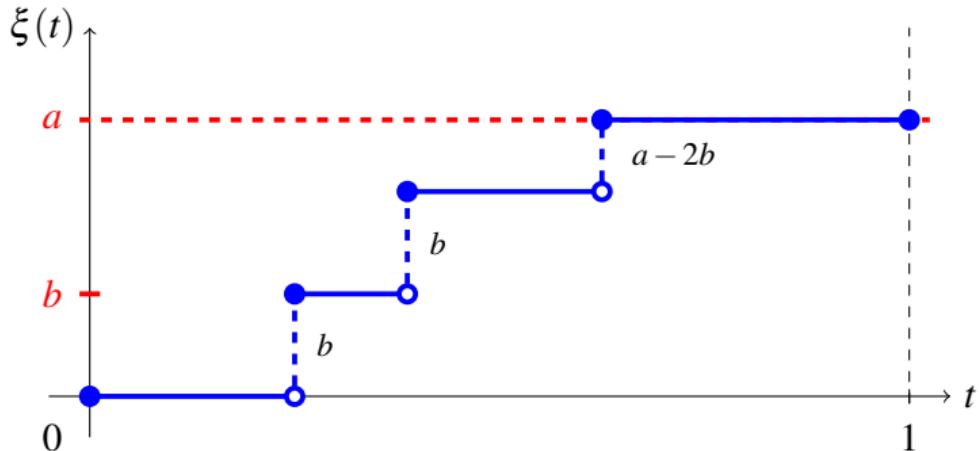
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



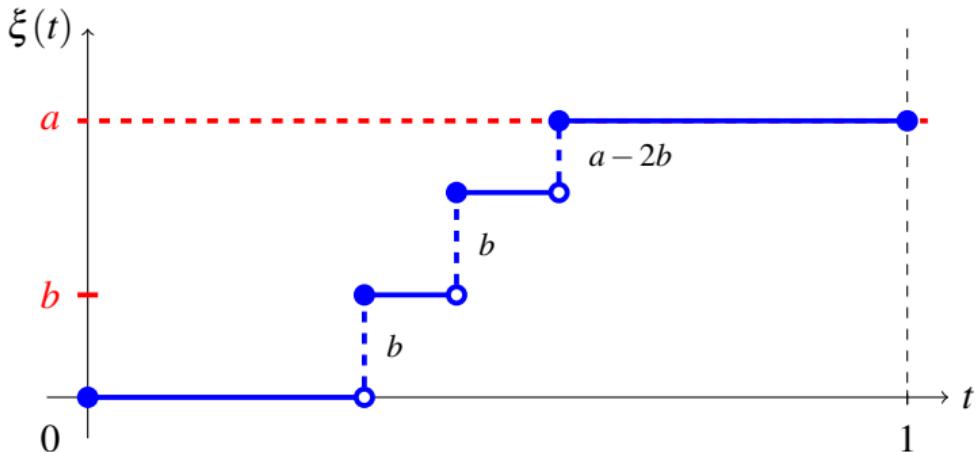
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a-2b)^\alpha$

## $L_1$ topology is weak



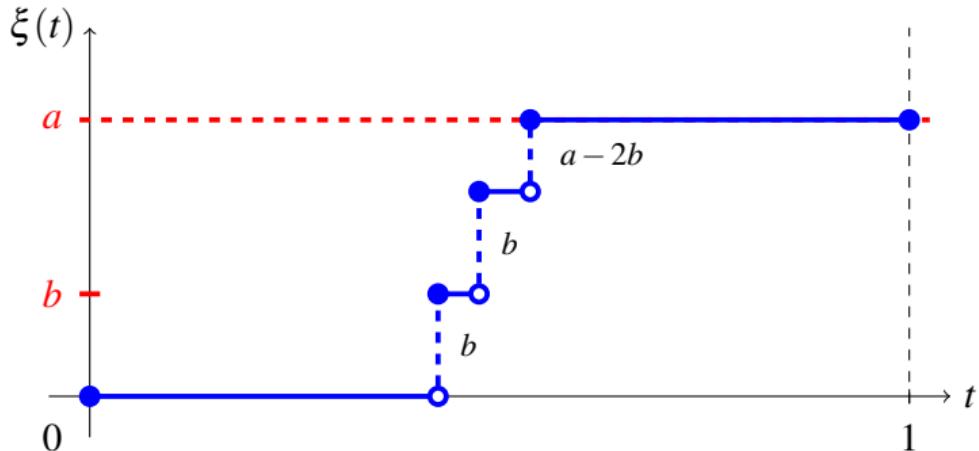
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



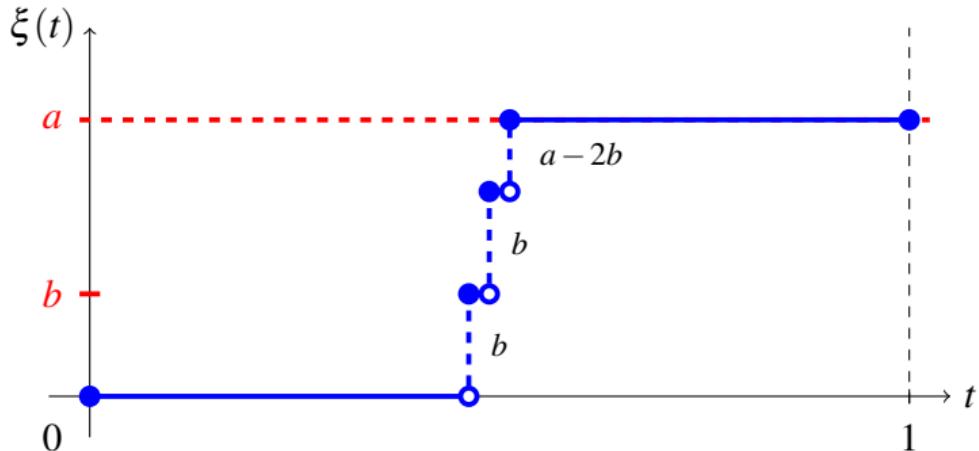
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



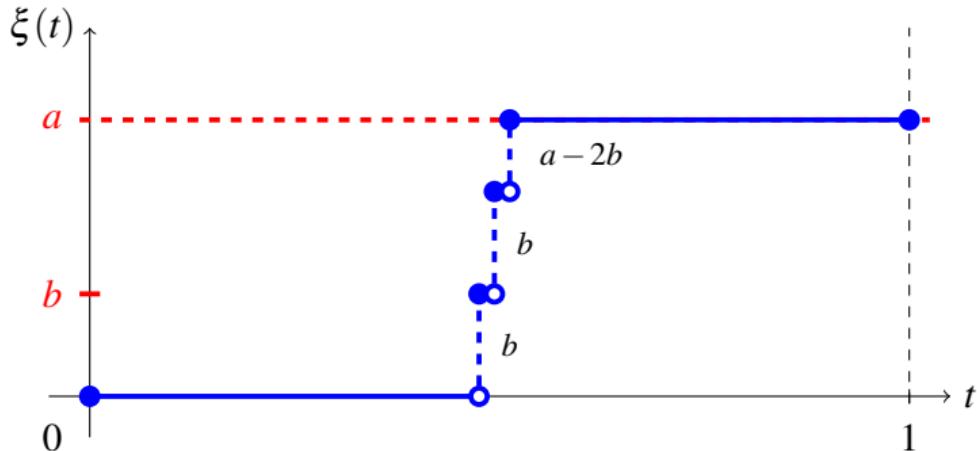
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



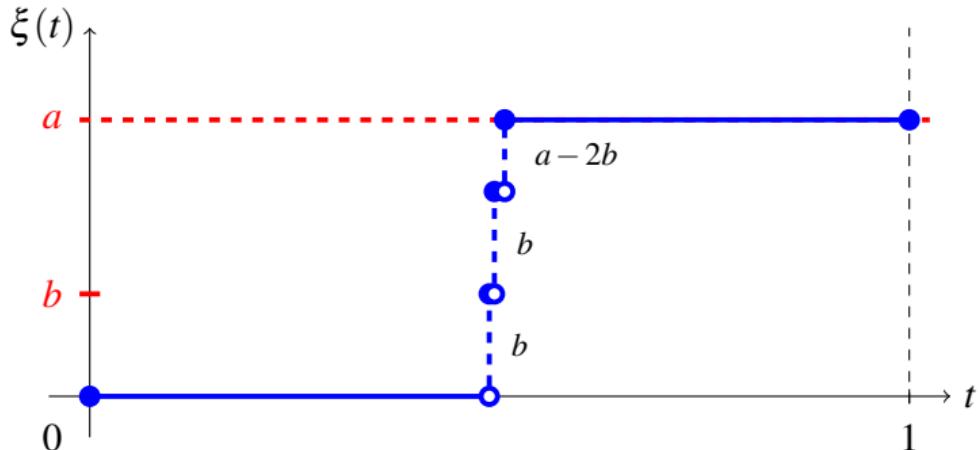
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



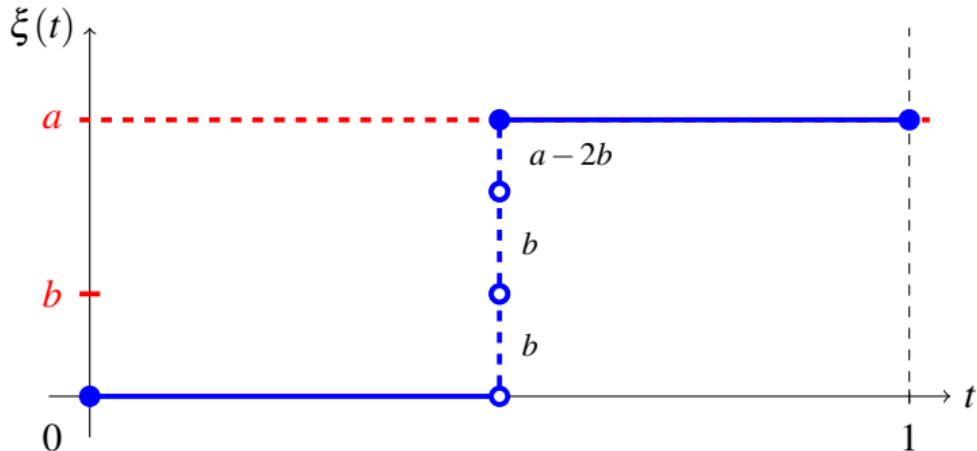
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



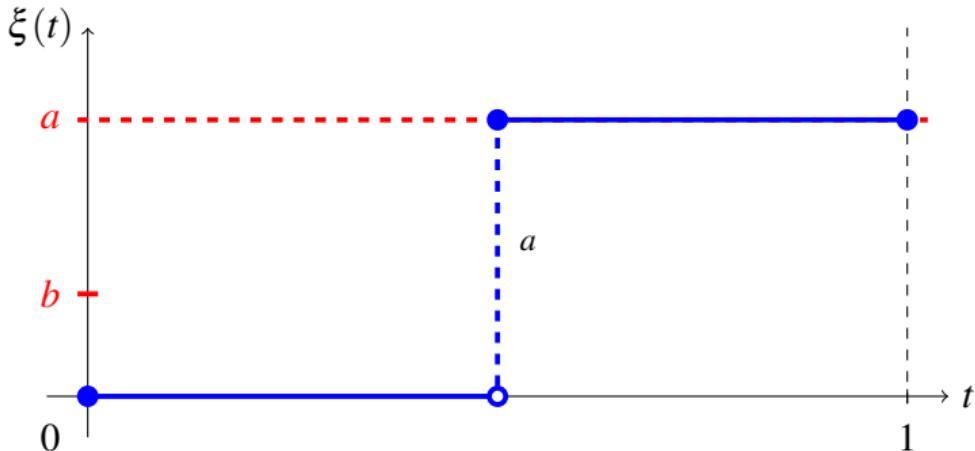
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a-2b)^\alpha$

## $L_1$ topology is weak



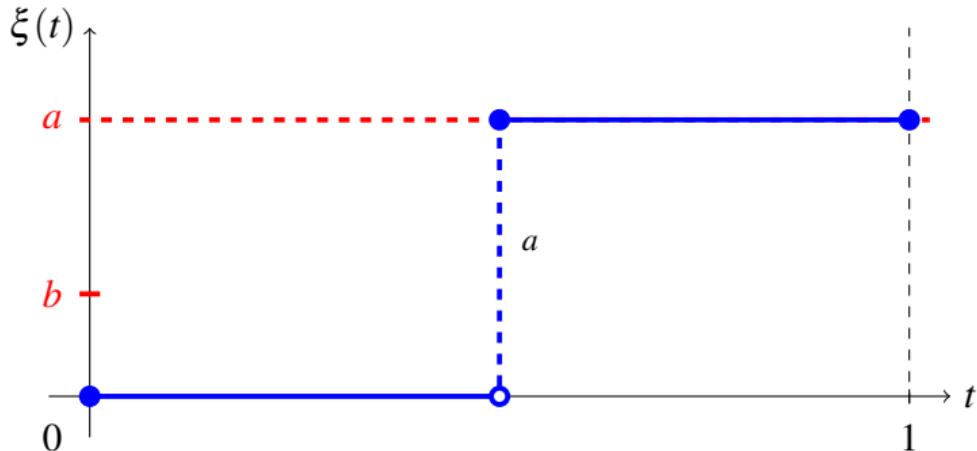
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-)), \quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



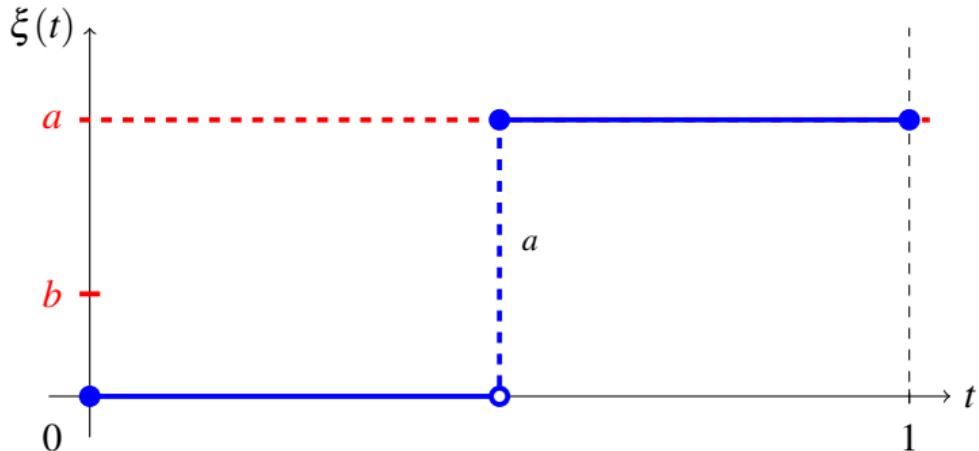
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-)), \quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) = b^\alpha + b^\alpha + (a - 2b)^\alpha$

## $L_1$ topology is weak



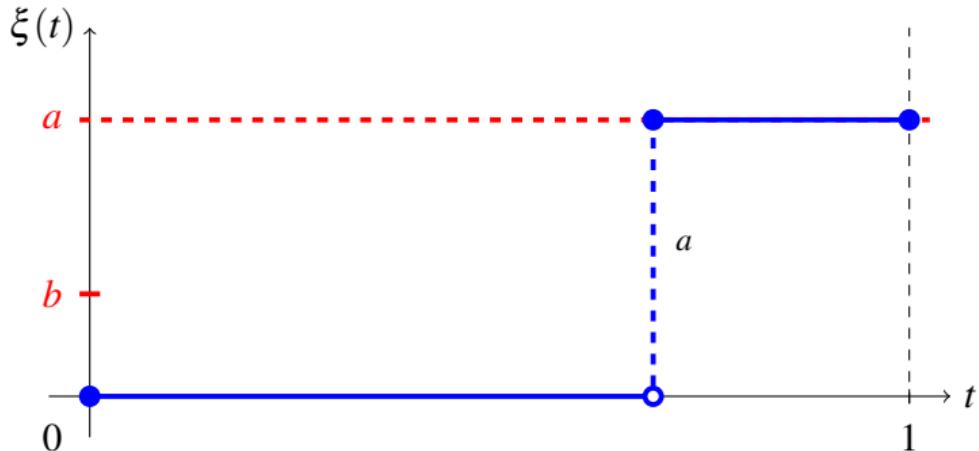
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-)), \quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \neq b^\alpha + b^\alpha + (a-2b)^\alpha$

## $L_1$ topology is weak



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

## $L_1$ topology is weak



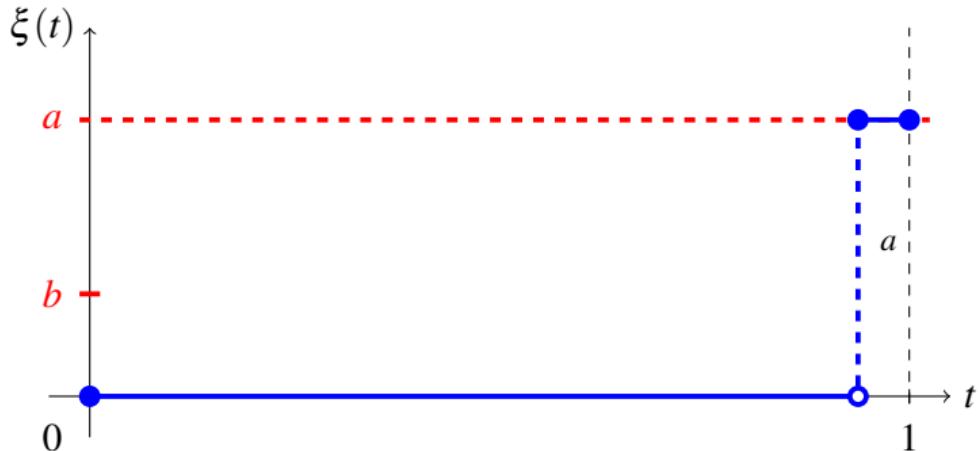
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-)), \quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b-2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



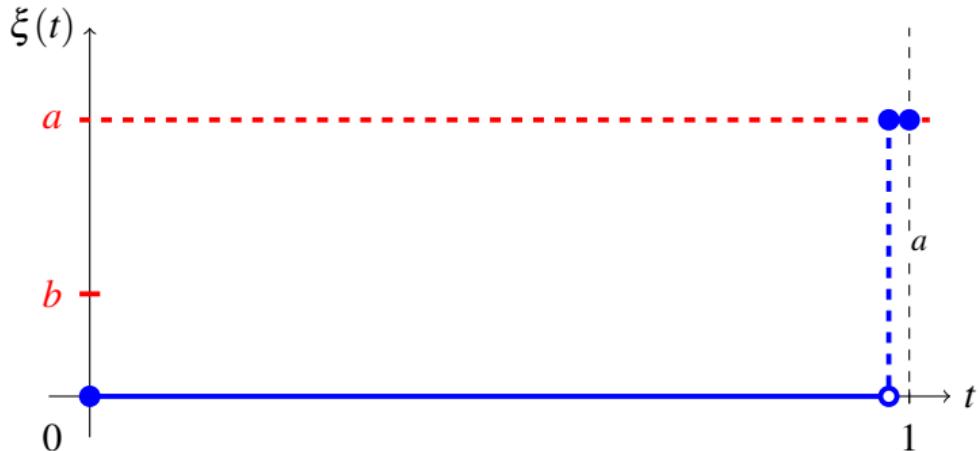
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b - 2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



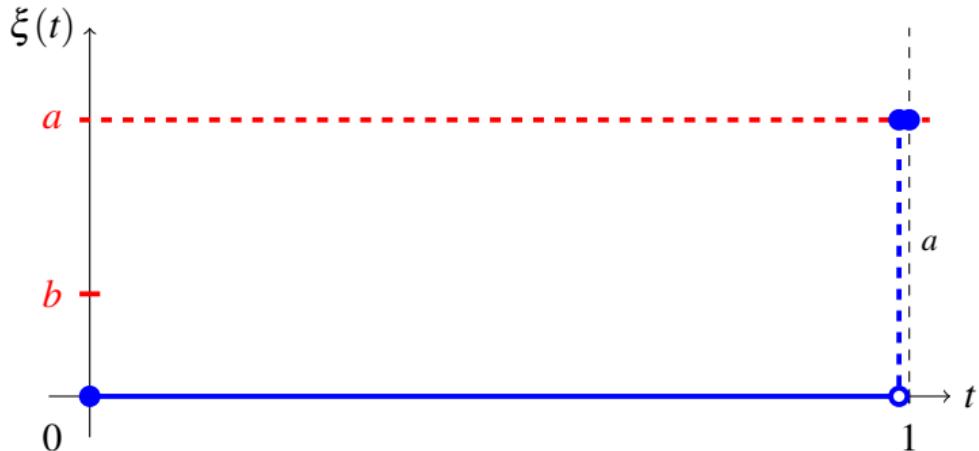
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b - 2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



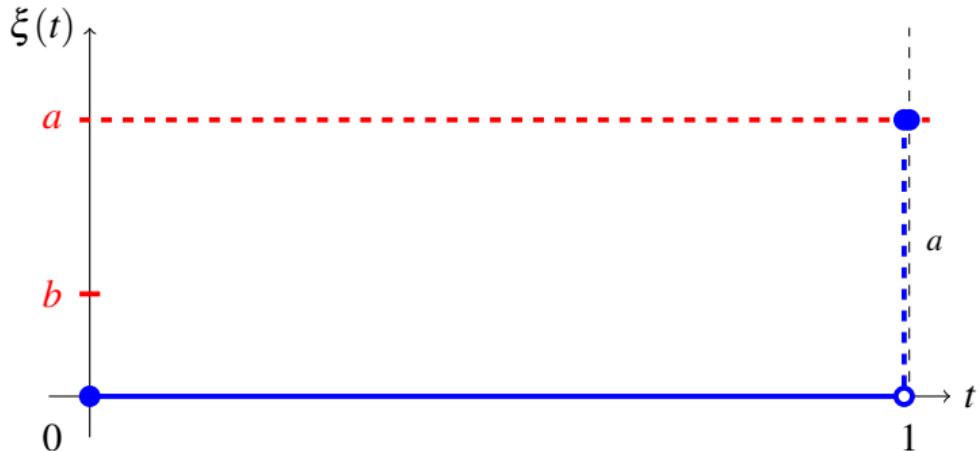
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-)), \quad I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b - 2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



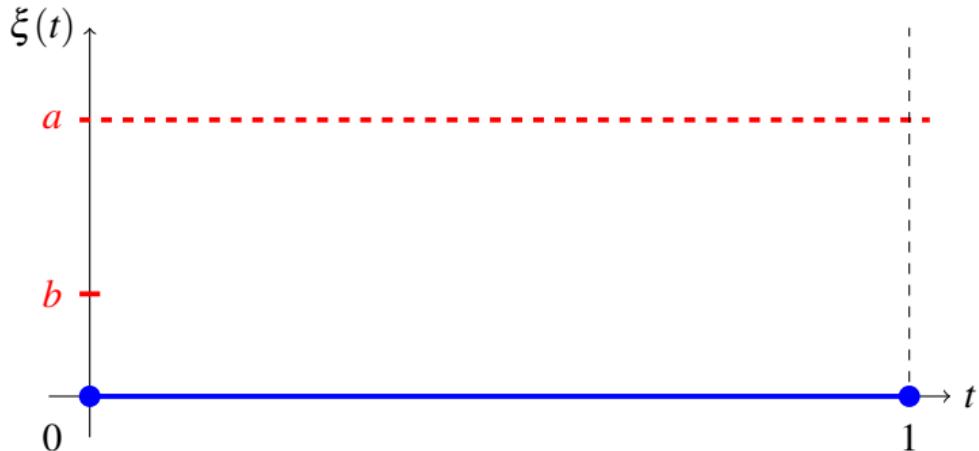
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b-2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



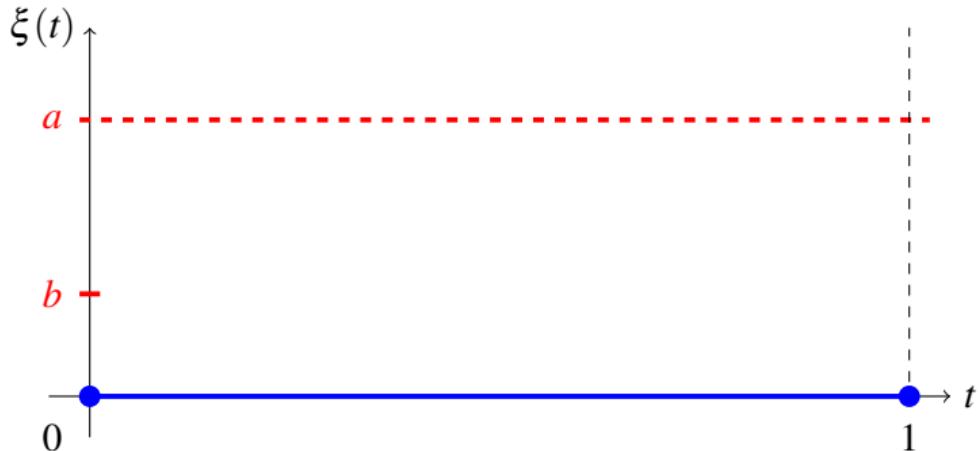
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b - 2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



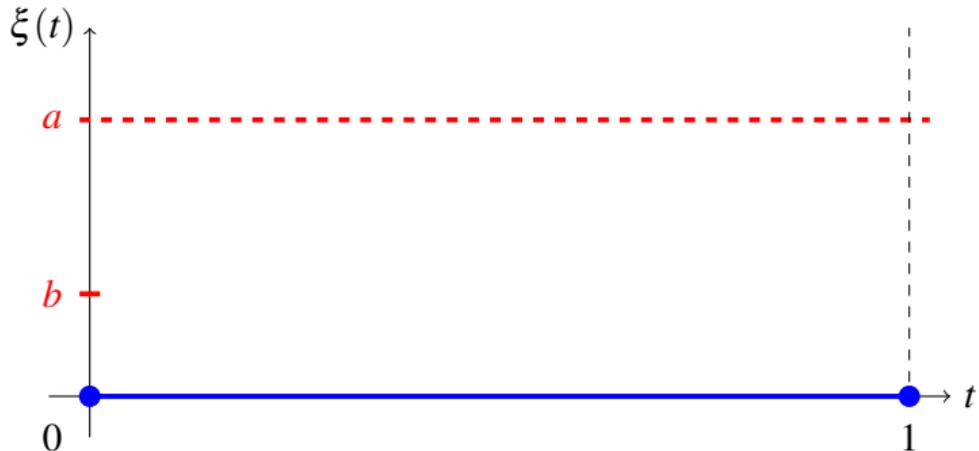
- $A = \{\xi \in \mathbb{D} : \sup_{t \in [0,1]} \xi(t) \geq a, \sup_{t \in [0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \lesssim \exp(-L(n)n^\alpha I(A^-))$ ,  $I(A^-) = \inf_{\xi \in A^-} I(\xi)$
- $I(A^-) \leq a^\alpha < b^\alpha + b^\alpha + (b-2a)^\alpha \iff a \mathbb{1}_{[1/2,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

## $L_1$ topology is weak



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

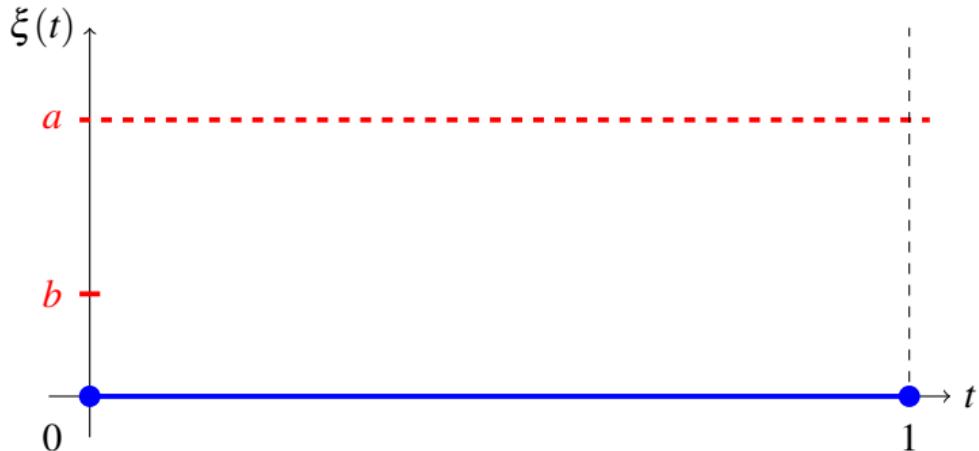
## $L_1$ topology is weak



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

**No information!**

## $L_1$ topology is weak



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

**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*

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

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

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

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

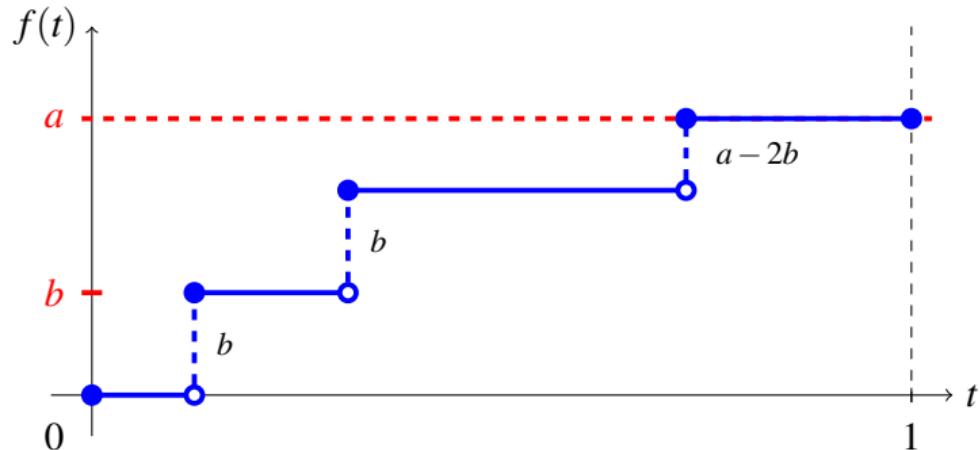
where

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

## Corollary

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

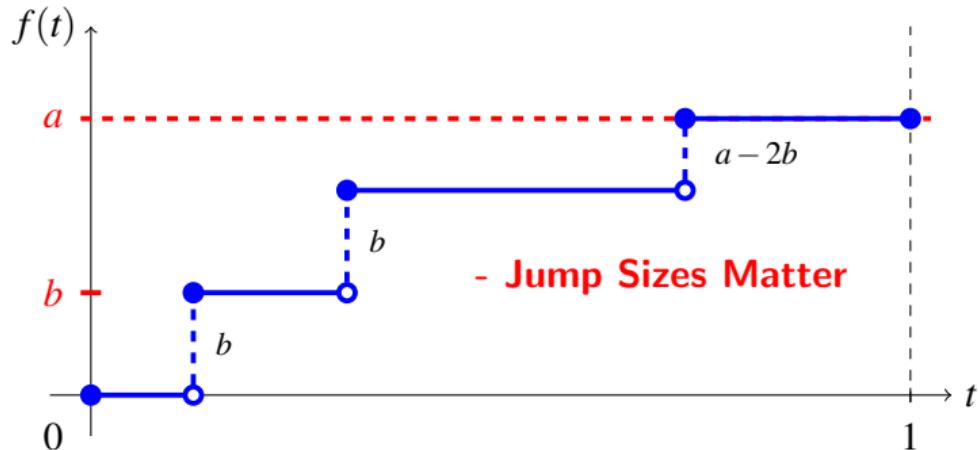
## Back to our reinsurance example



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

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

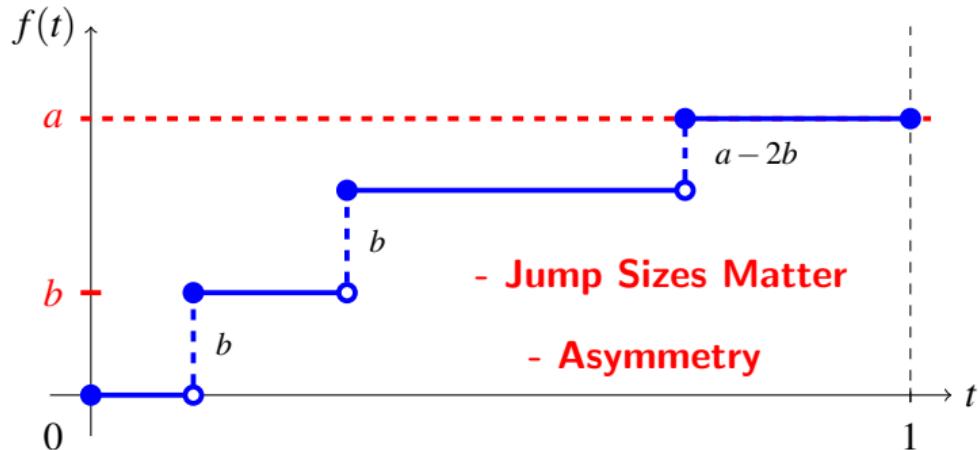
## Back to our reinsurance example



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

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

## Back to our reinsurance example

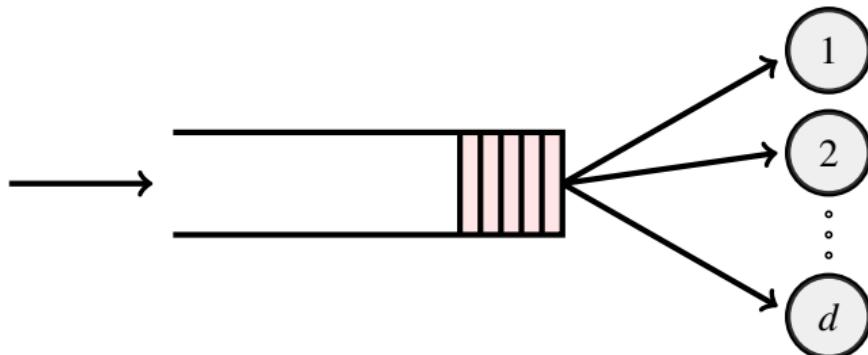


- $A = \{f \in \mathbb{D} : f \text{ crosses level } a \text{ on } [0, 1] \text{ & jump sizes } \leq b\}$
- $\mathbf{P}(\bar{S}_n \in A) \sim \exp(-n^\alpha I^*)$

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.,  $\mathbf{P}(S > x) \sim \exp(-x^\alpha)$ ,

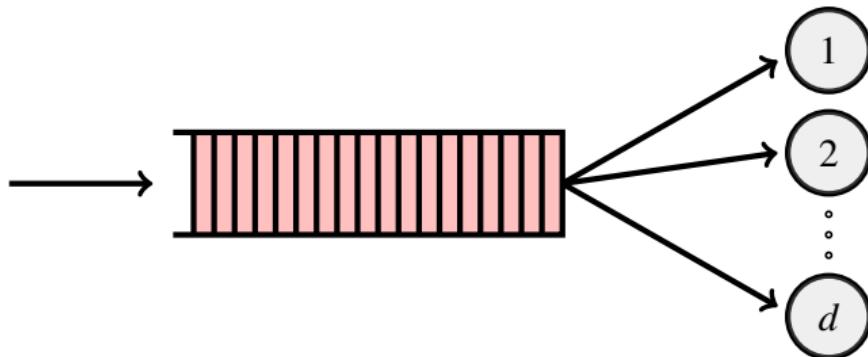


Tail Asymptotics? Most likely scenario?

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!

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$
- $A(t) = A_1 + A_2 + \dots + A_{\lfloor t \rfloor}$ ,

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

iid inter-arrival times

- $A(t) = \overset{\leftarrow}{A_1} + \overset{\leftarrow}{A_2} + \cdots + \overset{\downarrow}{A_{\lfloor t \rfloor}}$ ,

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$
  - $A(t) = \overset{\leftarrow}{A_1} + \overset{\leftarrow}{A_2} + \cdots + \overset{\downarrow}{A_{\lfloor t \rfloor}}$ , iid inter-arrival times
- $$S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$
  - $A(t) = A_1 + A_2 + \dots + A_{[t]}, \quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \dots + S_{[t]}^{(k)}$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$
  - $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$ ,  $\begin{matrix} \text{iid inter-arrival times} \\ \downarrow \\ \downarrow \\ \downarrow \end{matrix}$   $\begin{matrix} \text{iid service times for server } k \\ \downarrow \\ \downarrow \\ \downarrow \end{matrix}$
  - $M(t) = \inf\{s : A(s) > t\}$ ,

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$
  - $A(t) = A_1 + A_2 + \cdots + A_{[t]},$        $S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{[t]}^{(k)}$ 
    - iid inter-arrival times
    - iid service times for server  $k$
  - $M(t) = \inf\{s : A(s) > t\},$       inverse of  $A(\cdot)$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

iid inter-arrival times

- $$\bullet \quad A(t) = A_1 + A_2 + \cdots + A_{[t]},$$

iid service times for server  $k$

$$S^{(k)}(t) = \overset{\leftarrow}{S_1^{(k)}} + \overset{\leftarrow}{S_2^{(k)}} + \cdots + \overset{\downarrow}{S_{|t|}^{(k)}}$$

inverse of  $A(\cdot)$

- $$\bullet \quad M(t) = \inf\{s : A(s) > t\},$$

$$N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

- $A(t) = A_1 + A_2 + \dots + A_{[t]}$ ,      iid inter-arrival times
- $S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \dots + S_{[t]}^{(k)}$ ,      iid service times for server  $k$

$$\bullet M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

$$\bullet A(t) = \underbrace{A_1 + A_2 + \cdots + A_{[t]}}_{\text{iid inter-arrival times}}, \quad S^{(k)}(t) = \underbrace{S_1^{(k)} + S_2^{(k)} + \cdots + S_{[t]}^{(k)}}_{\text{iid service times for server } k}$$

$$\bullet M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$$

$$\bullet \quad \bar{M}_n(t) = M(nt)/n, \quad \quad \quad \bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

- $A(t) = A_1 + A_2 + \dots + A_{[t]}$ ,      iid inter-arrival times
- $S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \dots + S_{[t]}^{(k)}$ ,      iid service times for server  $k$

$$\bullet \quad M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$$

$$\bullet \quad \bar{M}_n(t) = M(nt)/n, \quad \bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

- $A(t) = A_1 + A_2 + \dots + A_{[t]}$ ,      iid inter-arrival times
- $S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \dots + S_{[t]}^{(k)}$ ,      iid service times for server  $k$

$$\bullet M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$$

$$\bullet \bar{M}_n(t) = M(nt)/n, \quad \bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$$

- LDP for  $A$  and  $S^{(k)}$   $\implies$  LDP for  $\bar{M}_n$  and  $\bar{N}_n^{(k)}$  (in  $M'_1$  topology)

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

- $A(t) = A_1 + A_2 + \cdots + A_{[t]}$ , iid inter-arrival times

- $M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$

$$\bullet \bar{M}_n(t) = M(nt)/n, \quad \bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$$

“contraction principle”

- LDP for  $A$  and  $S^{(k)}$   $\implies$  LDP for  $\bar{M}_n$  and  $\bar{N}_n^{(k)}$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

$$\bullet A(t) = \underbrace{A_1 + A_2 + \cdots + A_{[t]}}_{\text{iid inter-arrival times}}, \quad S^{(k)}(t) = \underbrace{S_1^{(k)} + S_2^{(k)} + \cdots + S_{[t]}^{(k)}}_{\text{iid service times for server } k}$$

$$\bullet M(t) \stackrel{\text{inverse of } A(\cdot)}{=} \inf\{s : A(s) > t\}, \quad N^{(k)}(t) \stackrel{\text{inverse of } S^{(k)}(\cdot)}{=} \inf\{s : S^{(k)}(s) > t\}$$

- $\bar{M}_n(t) = M(nt)/n$ , scaled processes (fluid scale)  $\bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$

“contraction principle”

- LDP for  $A$  and  $S^{(k)}$   $\stackrel{\downarrow}{\implies}$  LDP for  $\bar{M}_n$  and  $\bar{N}_n^{(k)}$  (in  $M'_1$  topology)

$$\mathbf{P}(Q(\gamma n) > n) \approx \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left( \bar{M}_n(s) - \sum_{k=1}^d \bar{N}_n^{(k)}(s) \right) > 1\right)$$

## Our Sample Path LDP Answers the Question!

- $Q(t)$ : length of GI/GI/d queue at time  $t$

iid inter-arrival times

$$\bullet \ A(t) = \overset{\leftarrow}{A_1} + \overset{\leftarrow}{A_2} + \cdots + \overset{\leftarrow}{A_{|t|}},$$

iid service times for server  $k$

$$S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{|t|}^{(k)}$$

inverse of  $A(\cdot)$

- $M(t) = \inf\{s : A(s) > t\}$ ,

inverse of  $S^{(k)}(\cdot)$

$$N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$$

scaled processes (fluid scale)

$$\bullet \quad \bar{M}_n(t) = M(nt)/n,$$

$$\bar{N}_n^{(k)}(t) = N^{(k)}(nt)/n$$

## “contraction principle”

- LDP for  $A$  and  $S^{(k)}$   $\xrightarrow{\text{LDP}}$  LDP for  $\bar{M}_n$  and  $\bar{N}_n^{(k)}$

(in  $M'_1$  topology)

$$\mathbf{P}(Q(\gamma n) > n) \approx \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left( \bar{M}_n(s) - \sum_{k=1}^d \bar{N}_n^{(k)}(s) \right) > 1\right)$$

continuous functional of  $\bar{M}_n$  and  $\bar{N}_n^{(k)}$ 's in  $M'_1$  topology

## 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

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

# 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

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

where

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

# 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

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

where

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

- 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

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

where

- Nontrivial Tradeoff between  
Size and Number of Jumps!

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

- 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

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

where

$$c^* = \min \sum_{i=1}^d x_i^\alpha \quad \text{s.t.}$$

- Nontrivial Tradeoff between  
Size and Number of Jumps!

$$\lambda s - \sum_{i=1}^d (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma],$$

$$x_1, \dots, x_d \geq 0.$$

- Asymmetry in Jump Sizes!

- 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

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

where

$$c^* = \min \sum_{i=1}^d x_i^\alpha \quad \text{s.t.}$$

- Nontrivial Tradeoff between  
Size and Number of Jumps!

$$\lambda s - \sum_{i=1}^d (s - x_i)^+ \geq 1 \text{ for some } s \in [0, \gamma],$$

- Asymmetry in Jump Sizes!

$$x_1, \dots, x_d \geq 0.$$

- 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  
⇒ **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  
⇒ **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)

## **Back to Stochastic Gradient Descent**

## Stochastic Gradient Descent

Minimizing loss function  $f$ :

$$W_{k+1} = W_k - \eta (f'(W_k)) \quad k = 0, 1, 2, \dots$$

## Stochastic Gradient Descent

Minimizing loss function  $f$ :

$$W_{k+1} = W_k - \eta(\tilde{f}'(W_k)) \quad k = 0, 1, 2, \dots$$

# Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

# Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

## Stochastic Gradient Descent

Minimizing loss function  $f$ :

$$W^{\eta}_{k+1} = W^{\eta}_k - \eta (f'(W^{\eta}_k) + Z_k) \quad k = 0, 1, 2, \dots$$

## Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

## Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

Then

$$W^\eta(\cdot) \rightarrow w(\cdot) \quad \text{as} \quad \eta \rightarrow 0$$

# Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

Then

$$W^\eta(\cdot) \rightarrow w(\cdot) \quad \text{as} \quad \eta \rightarrow 0$$

where

$$dw(t) = -f'(w(t))dt$$

# Stochastic Gradient Descent

Minimizing loss function  $f$ :

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

Then

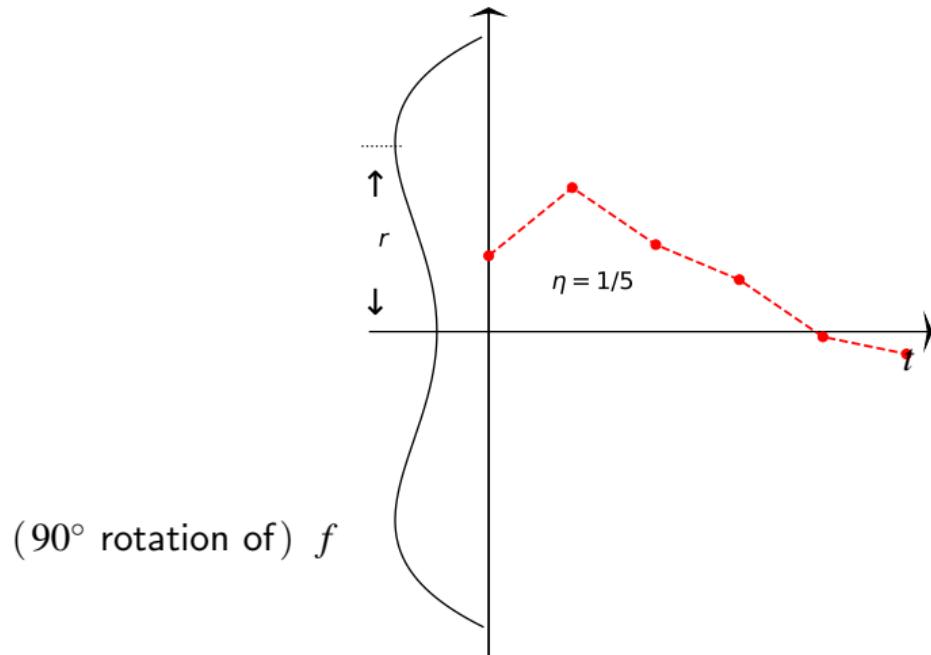
$$W^\eta(\cdot) \rightarrow w(\cdot) \quad \text{as} \quad \eta \rightarrow 0$$

↖  
Gradient Flow

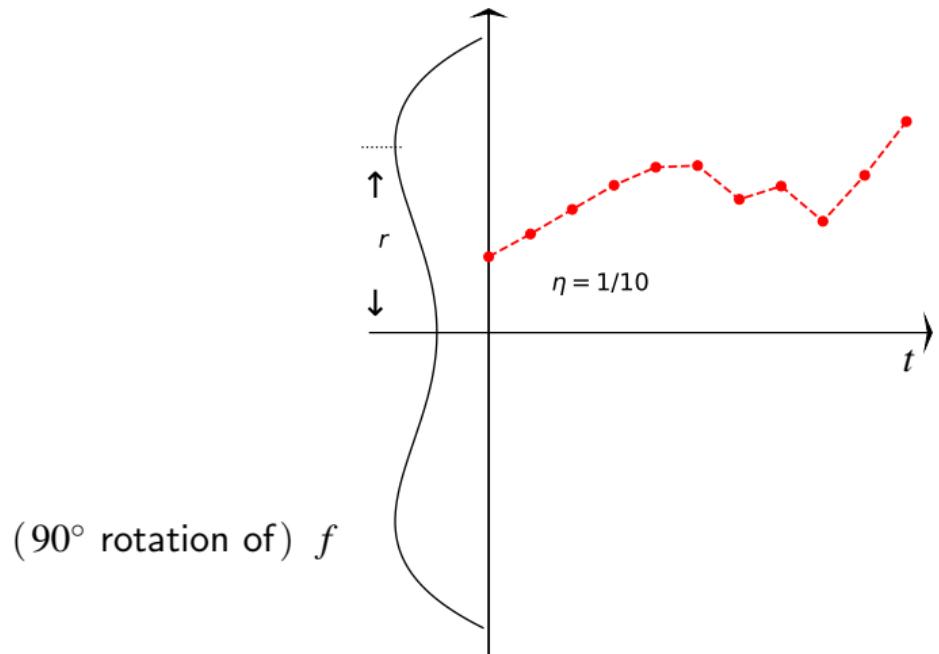
where

$$dw(t) = -f'(w(t))dt$$

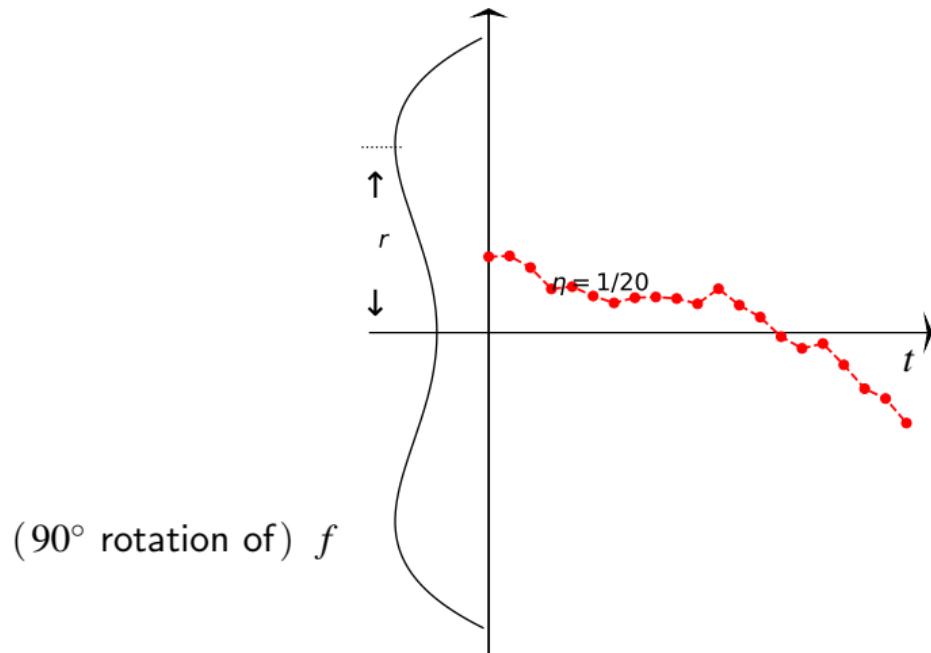
# Gradient Flow: Law of Large Numbers



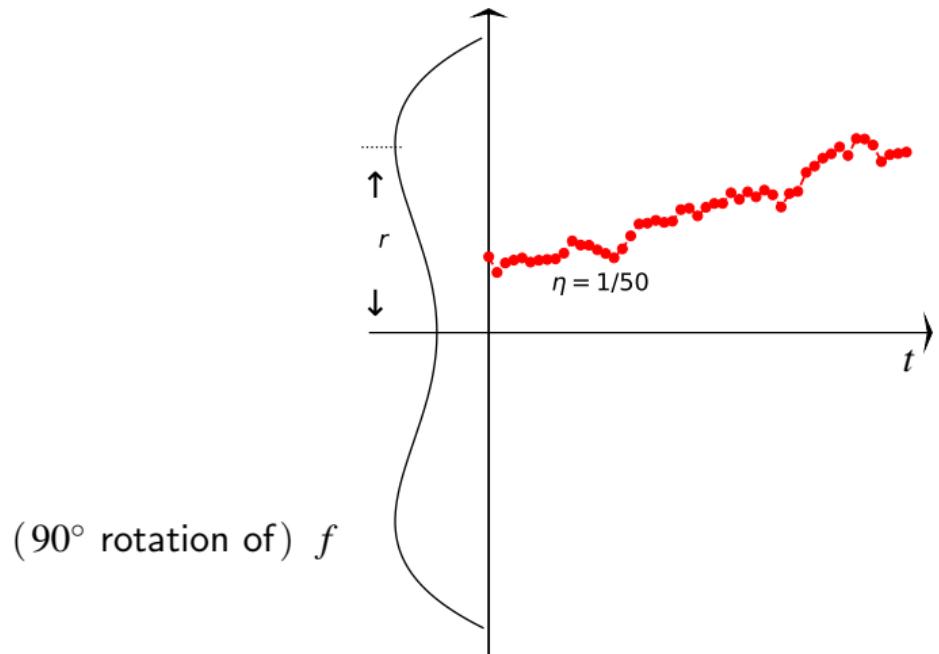
# Gradient Flow: Law of Large Numbers



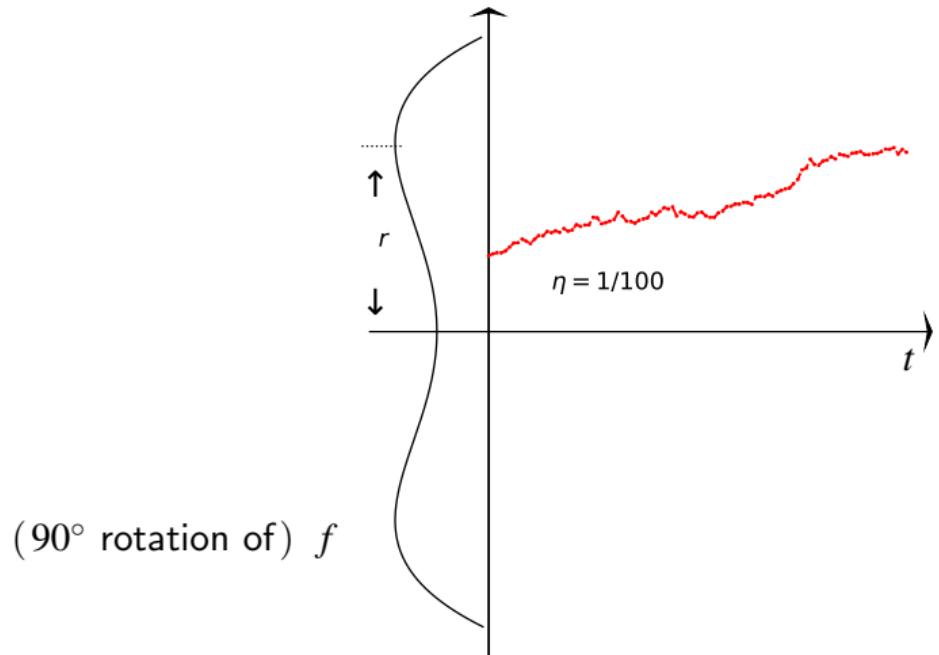
# Gradient Flow: Law of Large Numbers



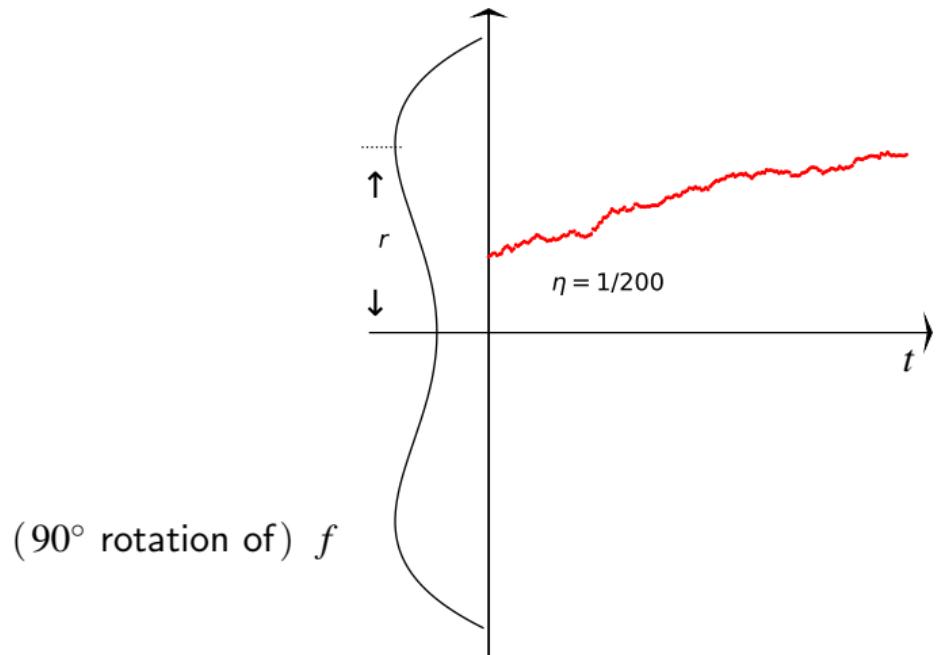
# Gradient Flow: Law of Large Numbers



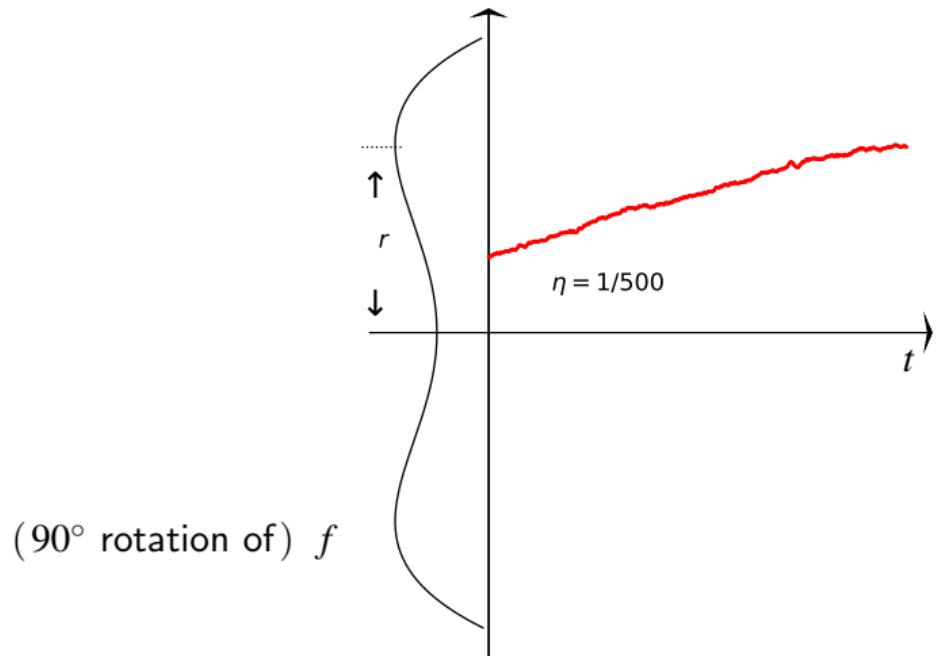
# Gradient Flow: Law of Large Numbers



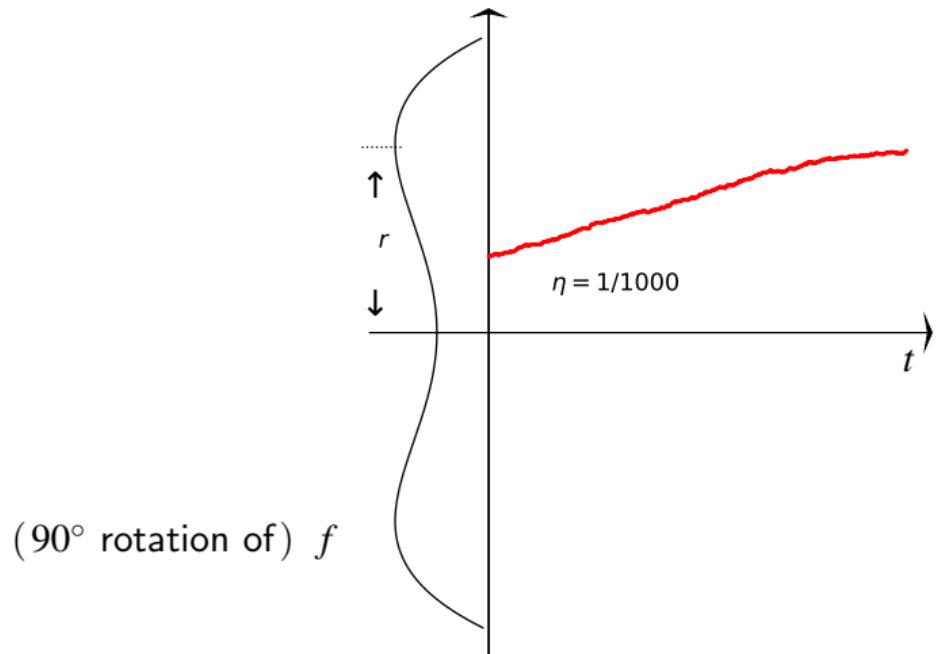
# Gradient Flow: Law of Large Numbers



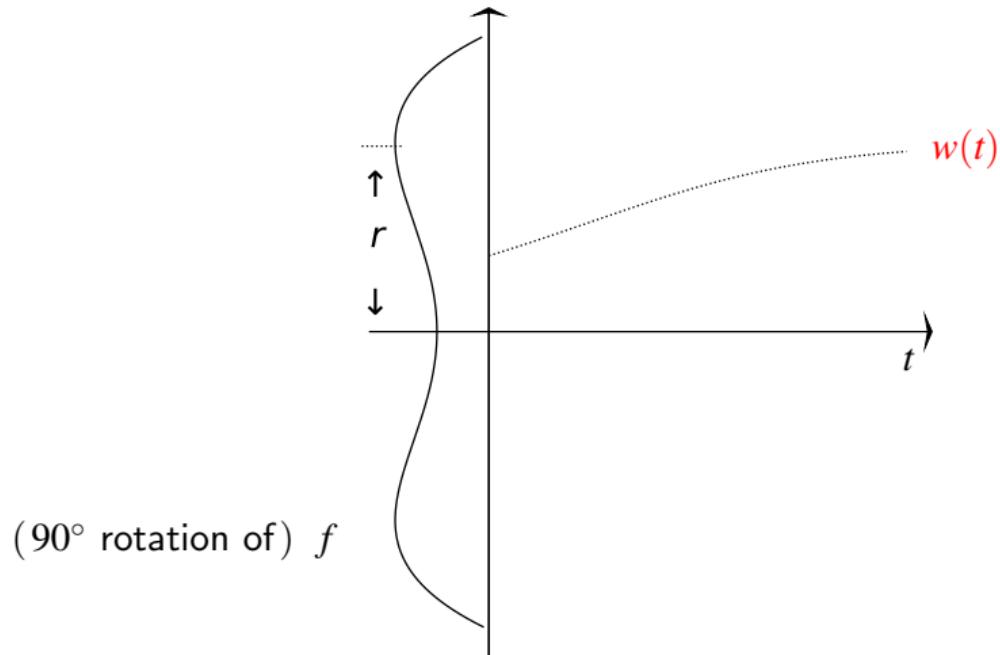
# Gradient Flow: Law of Large Numbers



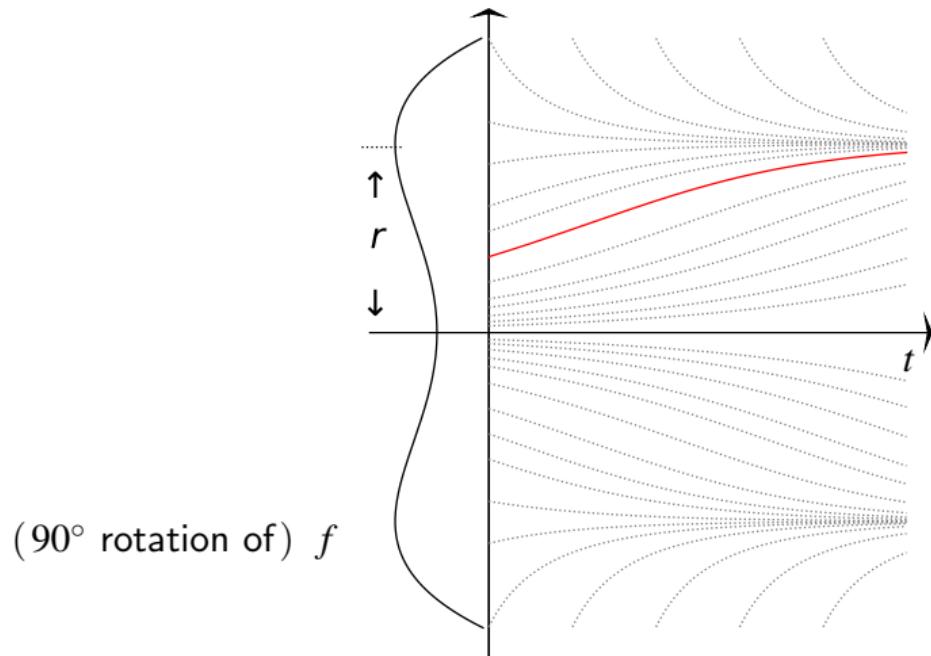
# Gradient Flow: Law of Large Numbers



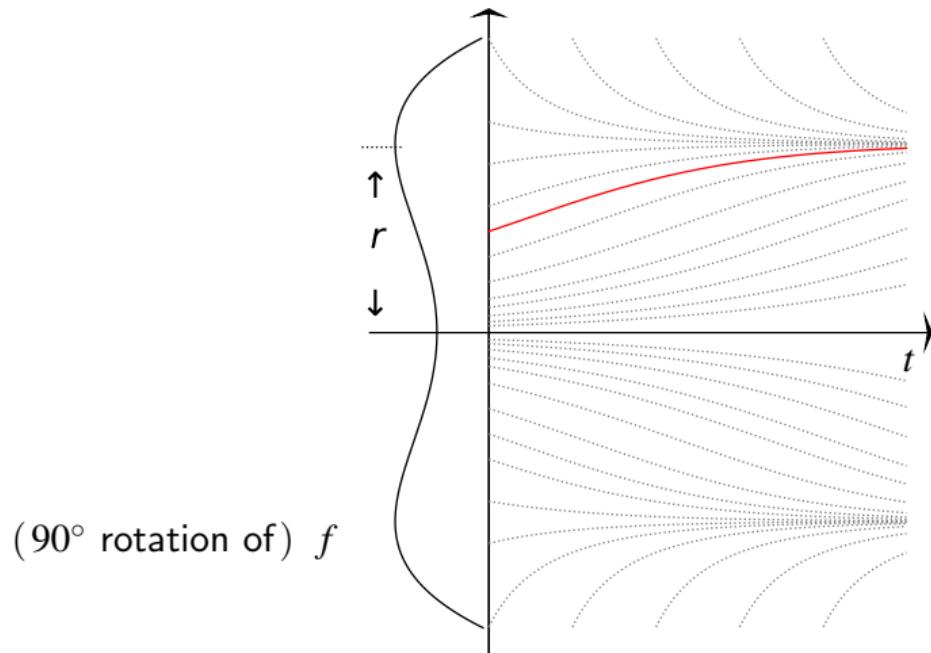
# Gradient Flow: Law of Large Numbers



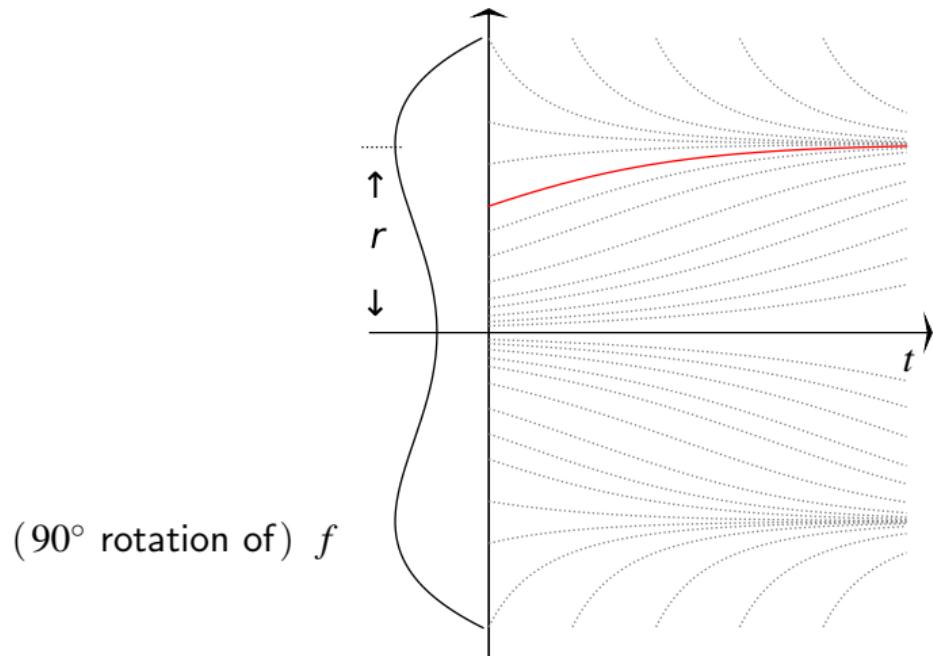
# Gradient Flow: Law of Large Numbers



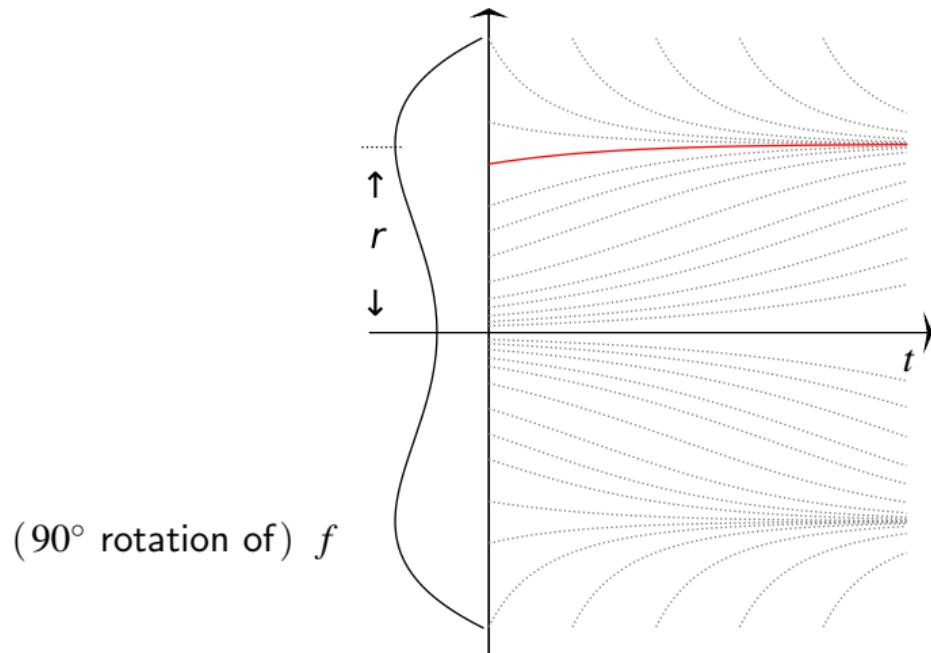
# Gradient Flow: Law of Large Numbers



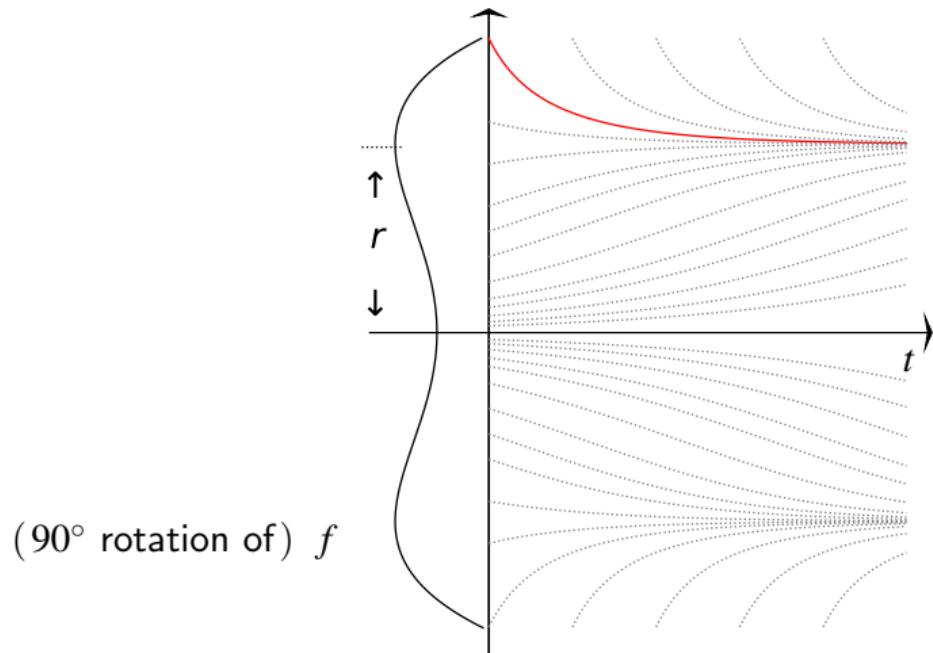
# Gradient Flow: Law of Large Numbers



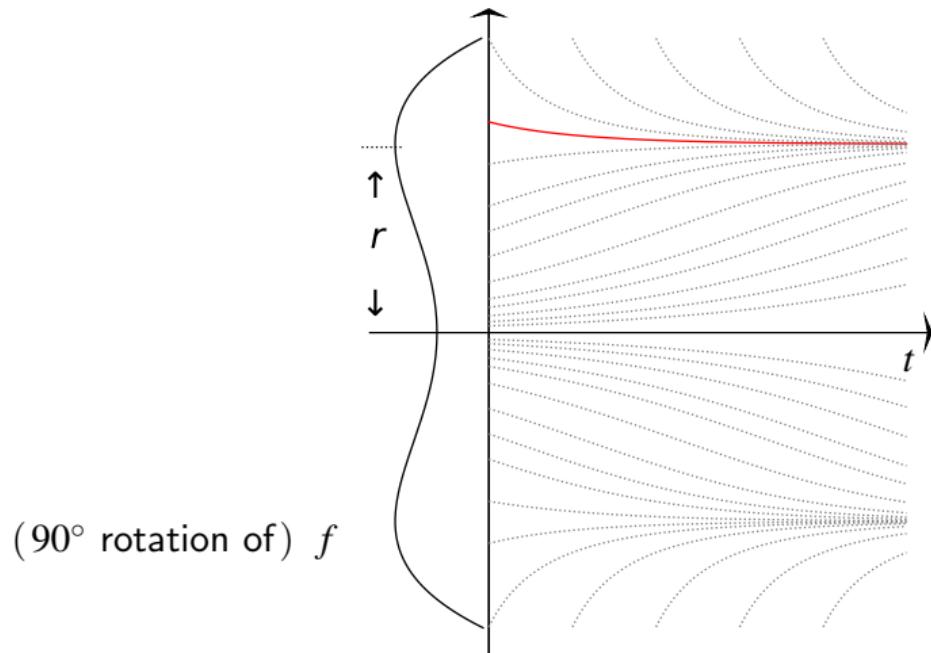
# Gradient Flow: Law of Large Numbers



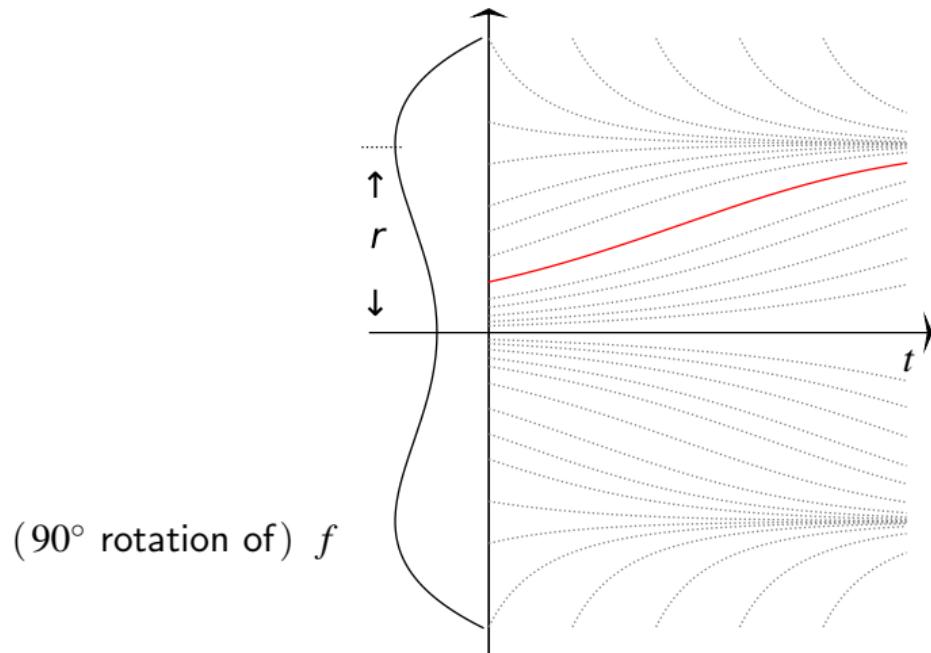
# Gradient Flow: Law of Large Numbers



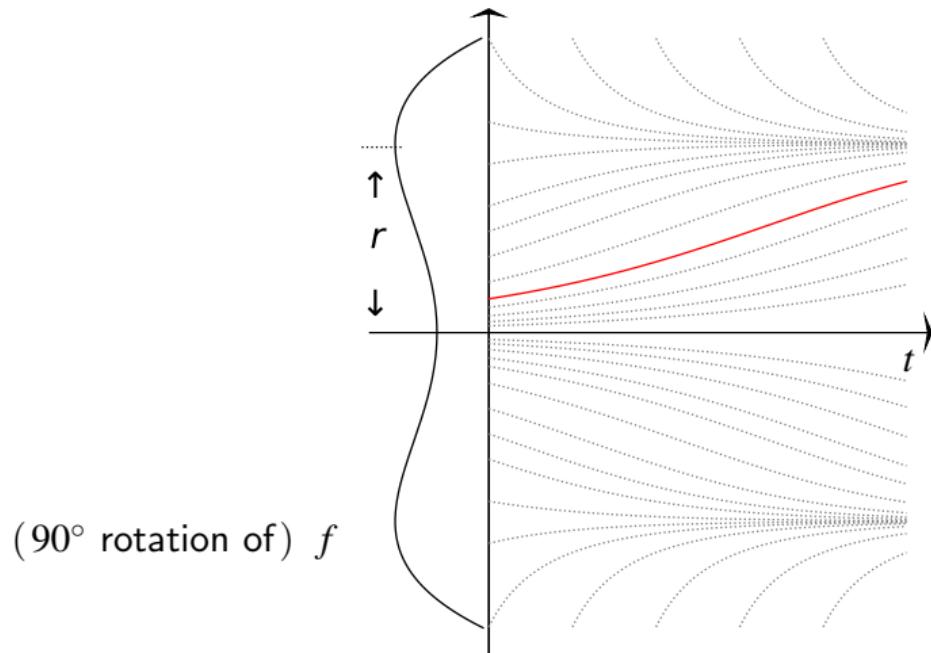
# Gradient Flow: Law of Large Numbers



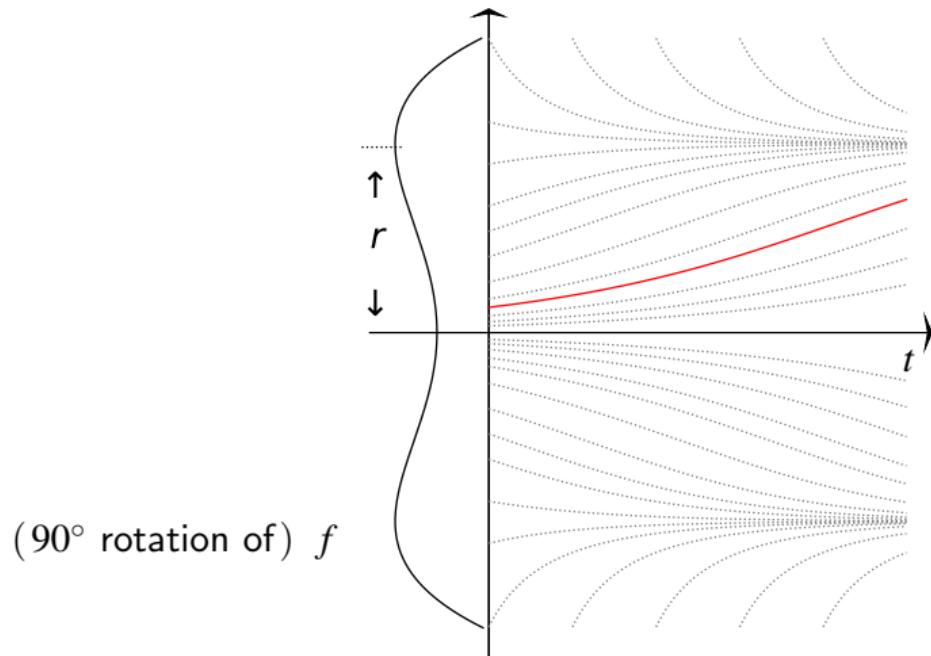
# Gradient Flow: Law of Large Numbers



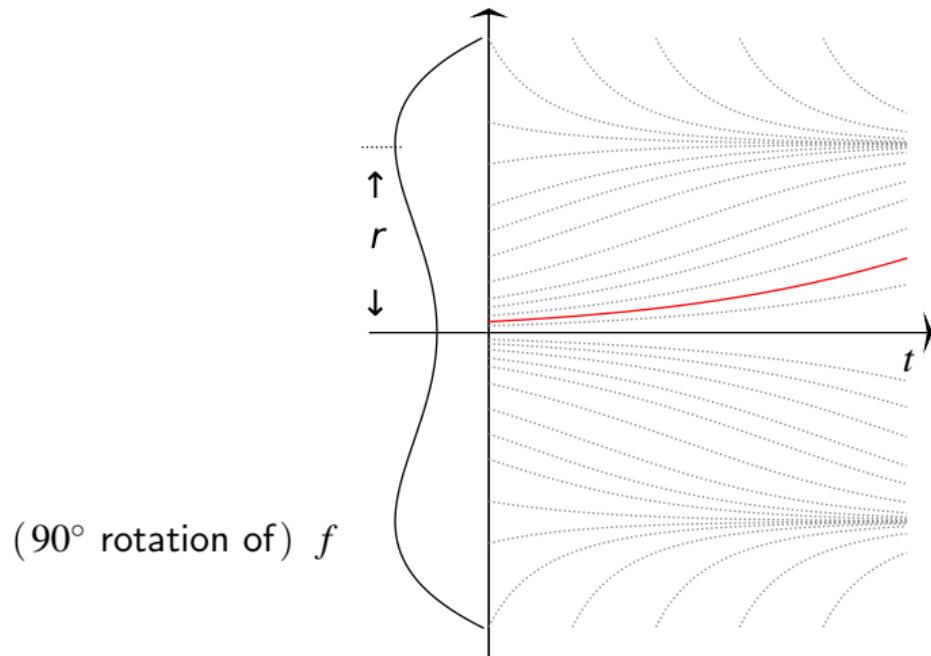
# Gradient Flow: Law of Large Numbers



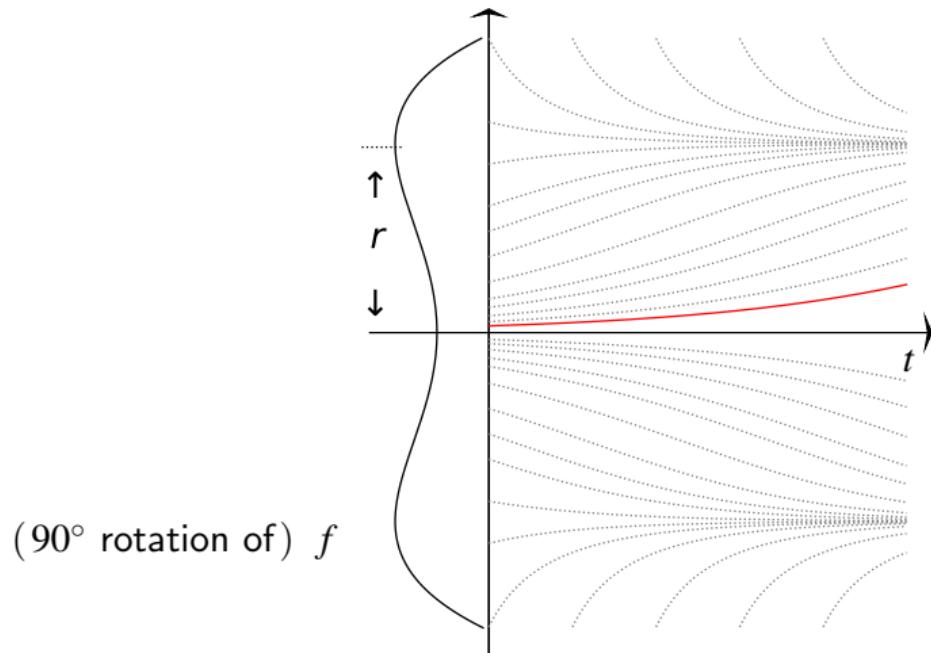
# Gradient Flow: Law of Large Numbers



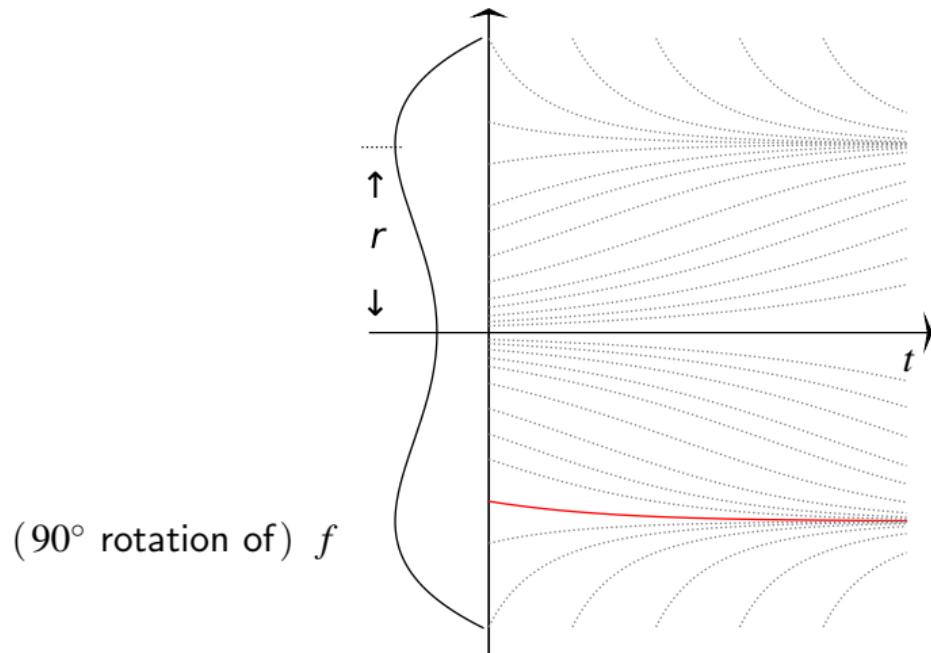
# Gradient Flow: Law of Large Numbers



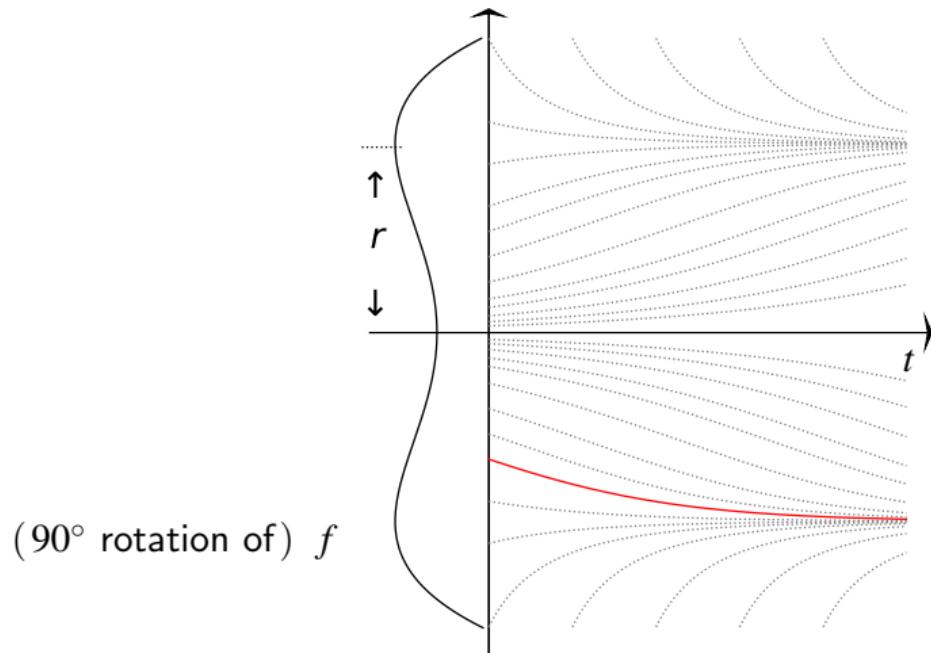
# Gradient Flow: Law of Large Numbers



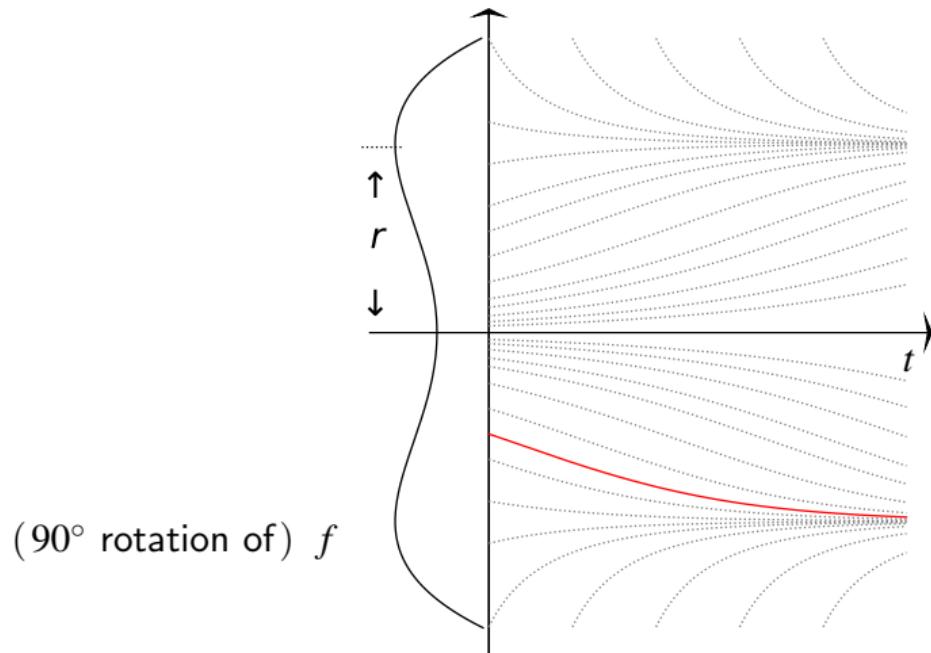
# Gradient Flow: Law of Large Numbers



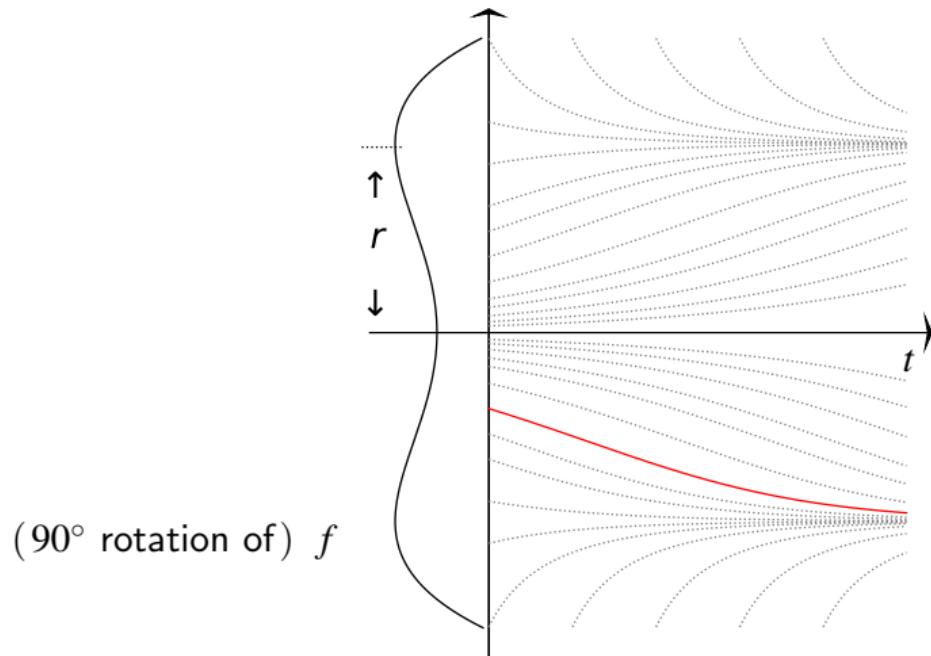
# Gradient Flow: Law of Large Numbers



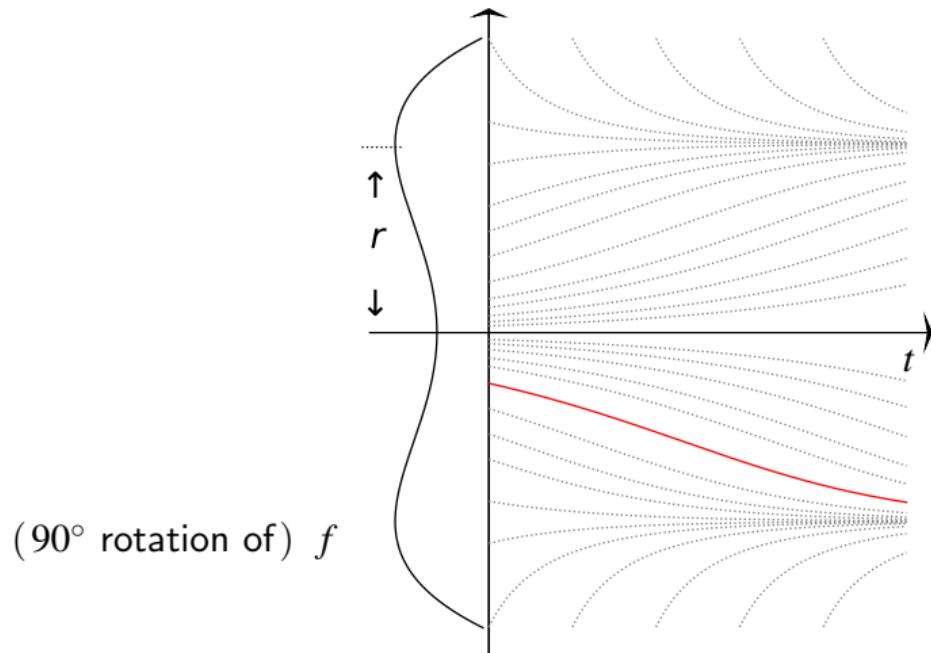
# Gradient Flow: Law of Large Numbers



# Gradient Flow: Law of Large Numbers



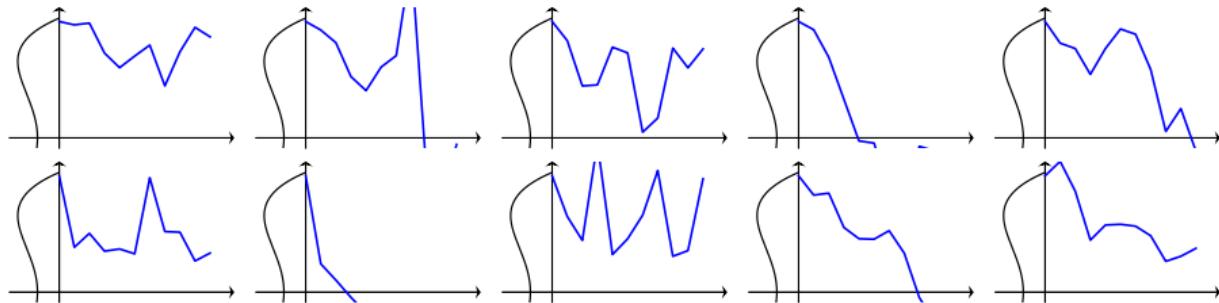
# Gradient Flow: Law of Large Numbers



## Typical Scenario

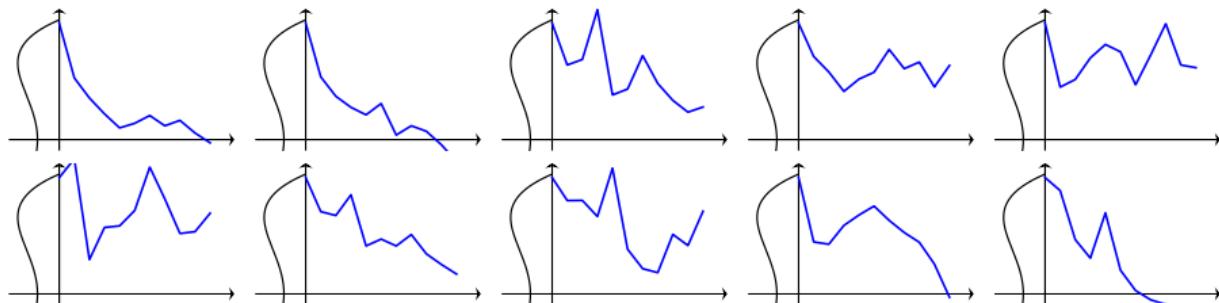
Trajectory of SGD  $W^\eta$ :

$\eta = 1/10$  & noises are **light-tailed**



Trajectory of SGD  $W^\eta$ :

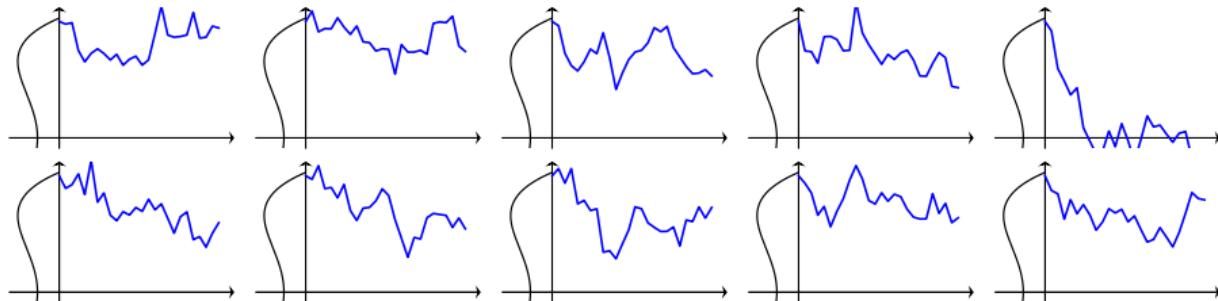
$\eta = 1/10$  & noises are **heavy-tailed**



## Typical Scenario

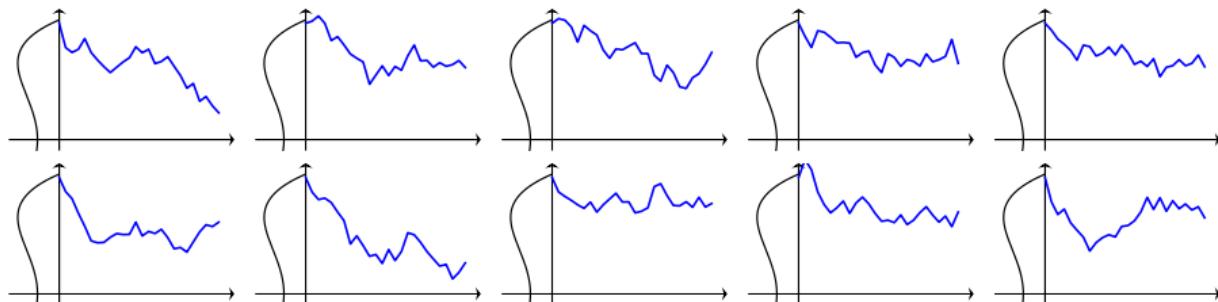
Trajectory of SGD  $W^\eta$ :

$\eta = 1/25$  & noises are **light-tailed**



Trajectory of SGD  $W^\eta$ :

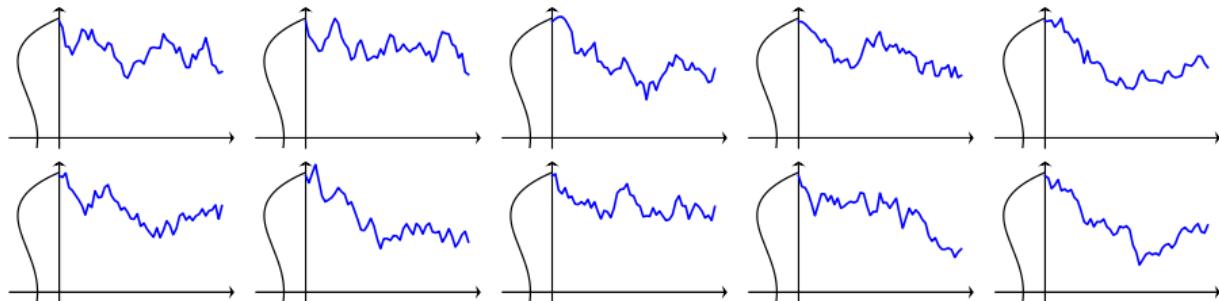
$\eta = 1/25$  & noises are **heavy-tailed**



## Typical Scenario

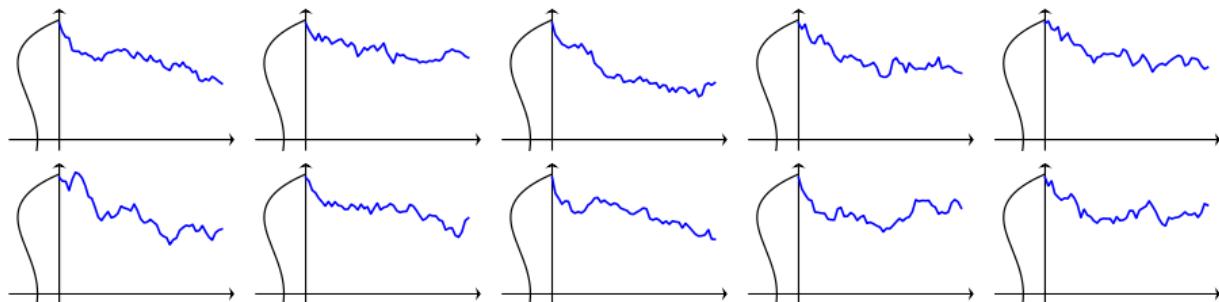
Trajectory of SGD  $W^\eta$ :

$\eta = 1/50$  & noises are **light-tailed**



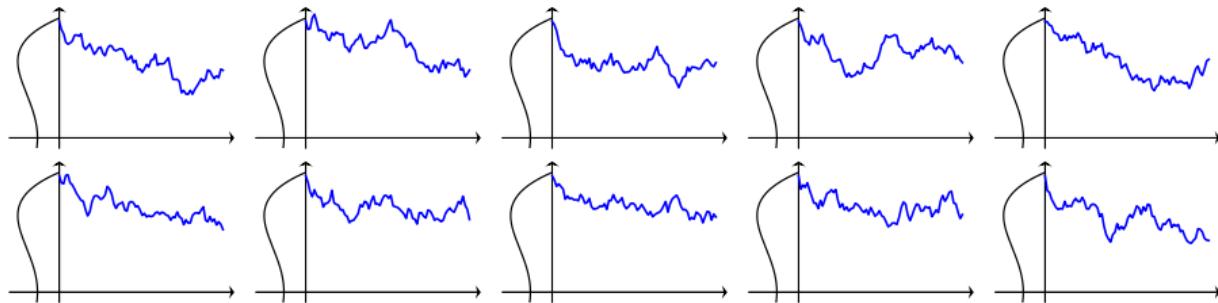
Trajectory of SGD  $W^\eta$ :

$\eta = 1/50$  & noises are **heavy-tailed**



## Typical Scenario

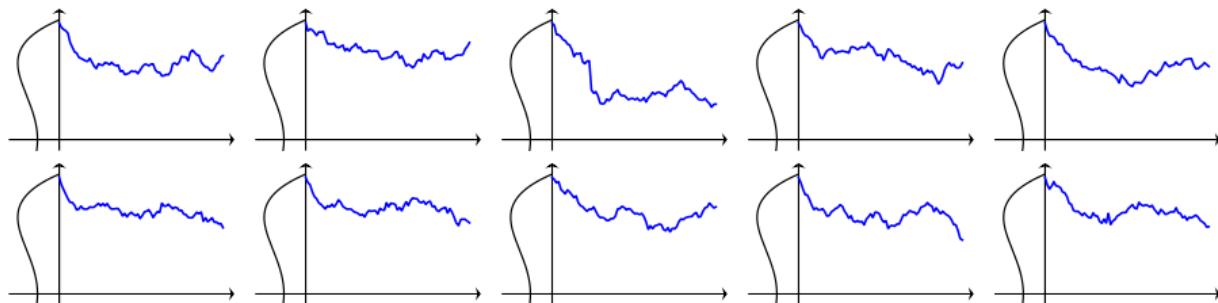
Trajectory of SGD  $W^\eta$ :



$\eta = 1/75$  & noises are **light-tailed**

Trajectory of SGD  $W^\eta$ :

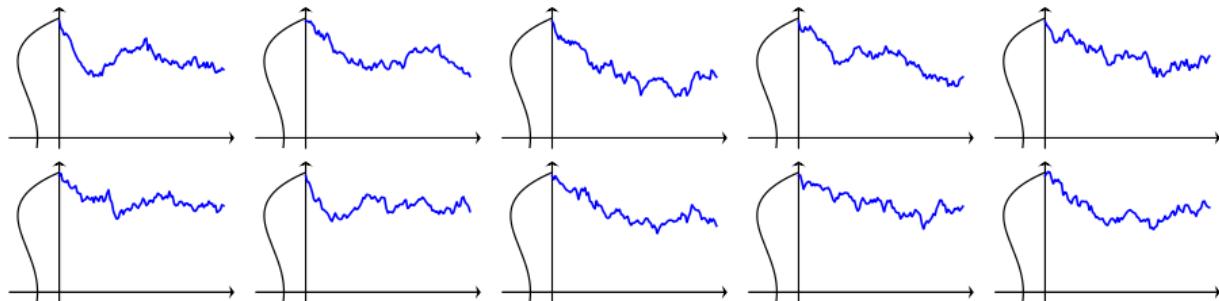
$\eta = 1/75$  & noises are **heavy-tailed**



## Typical Scenario

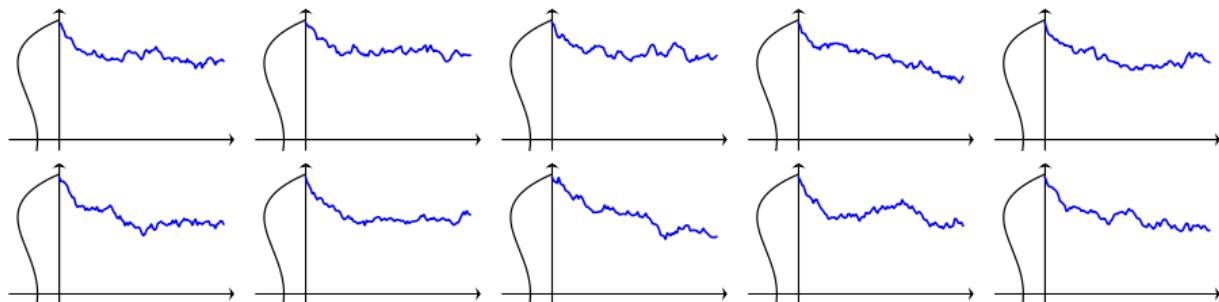
Trajectory of SGD  $W^\eta$ :

$\eta = 1/100$  & noises are **light-tailed**



Trajectory of SGD  $W^\eta$ :

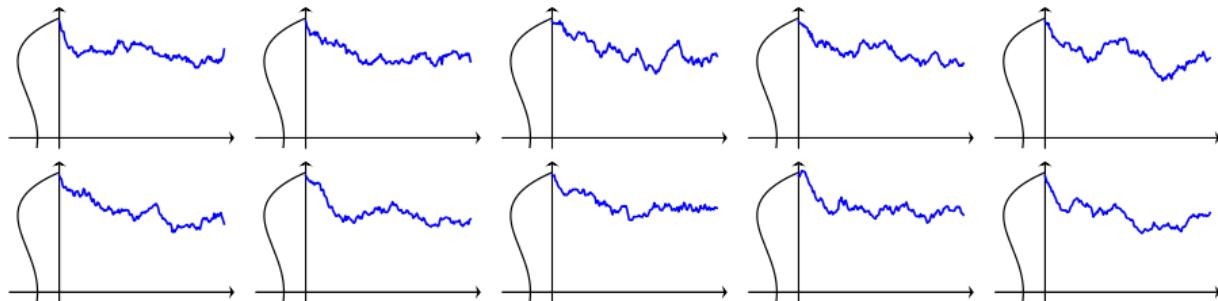
$\eta = 1/100$  & noises are **heavy-tailed**



## Typical Scenario

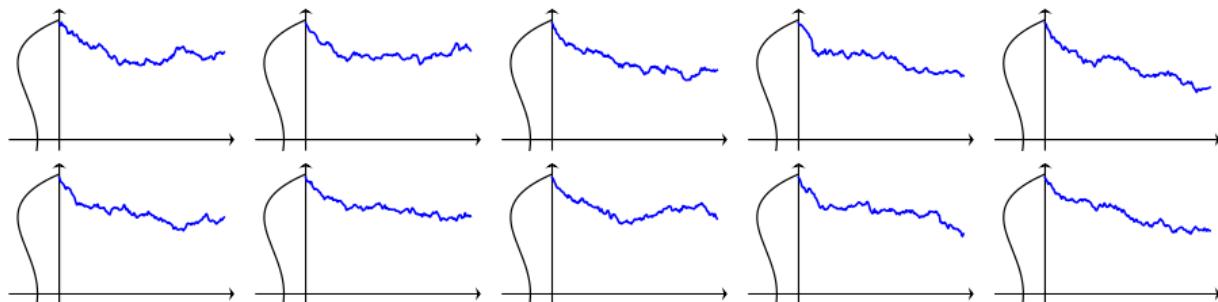
Trajectory of SGD  $W^\eta$ :

$\eta = 1/150$  & noises are **light-tailed**



Trajectory of SGD  $W^\eta$ :

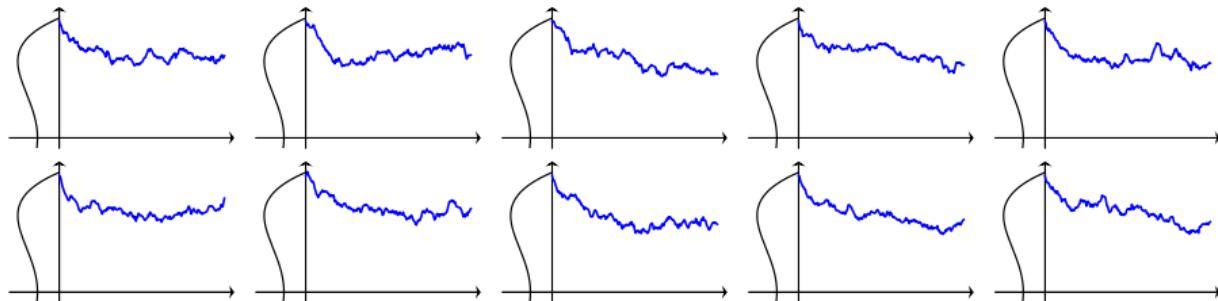
$\eta = 1/150$  & noises are **heavy-tailed**



## Typical Scenario

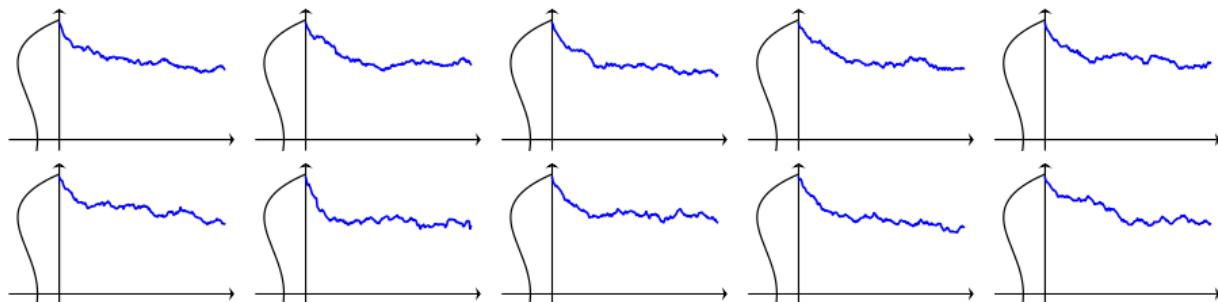
Trajectory of SGD  $W^\eta$ :

$\eta = 1/200$  & 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”  $A \subseteq \mathbb{D}$

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

- $\mathcal{J}(A)$ : 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”  $A \subseteq \mathbb{D}$

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

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

# Heavy-Tailed Large Deviations for SGD

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

gradient noise

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

For “general”  $A \subseteq \mathbb{D}$

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

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

# Heavy-Tailed Large Deviations for SGD

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

gradient noise

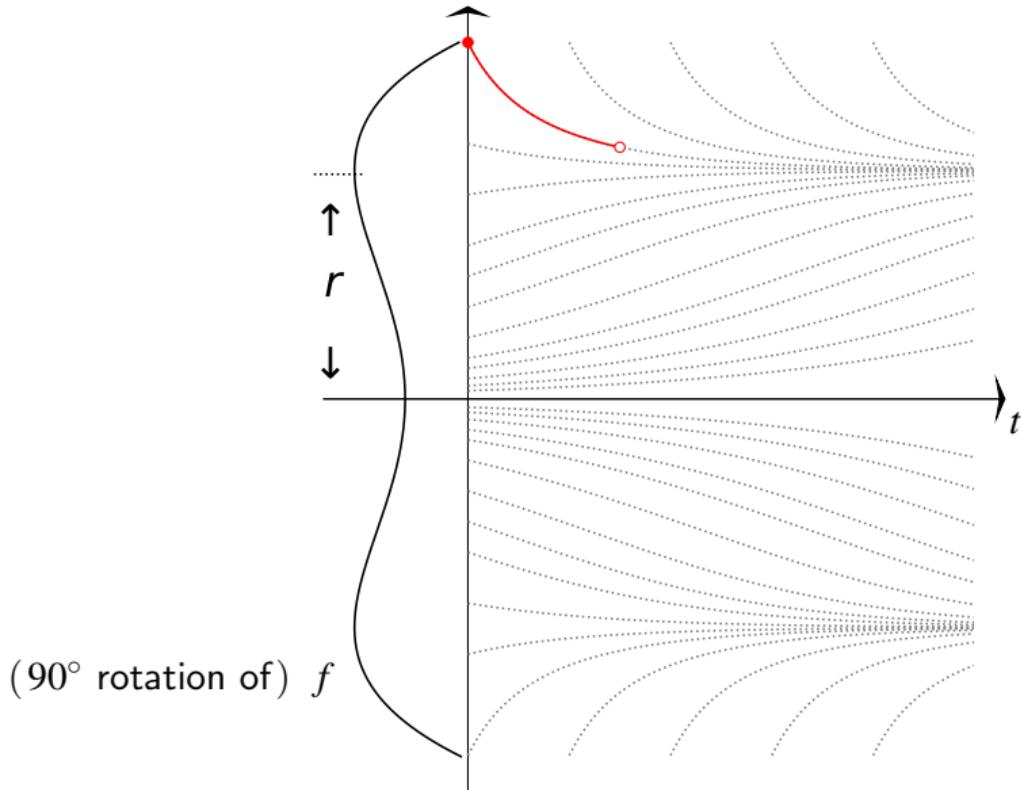
Theorem (Wang, Su, R., 2022+)

For “general”  $A \subseteq \mathbb{D}$

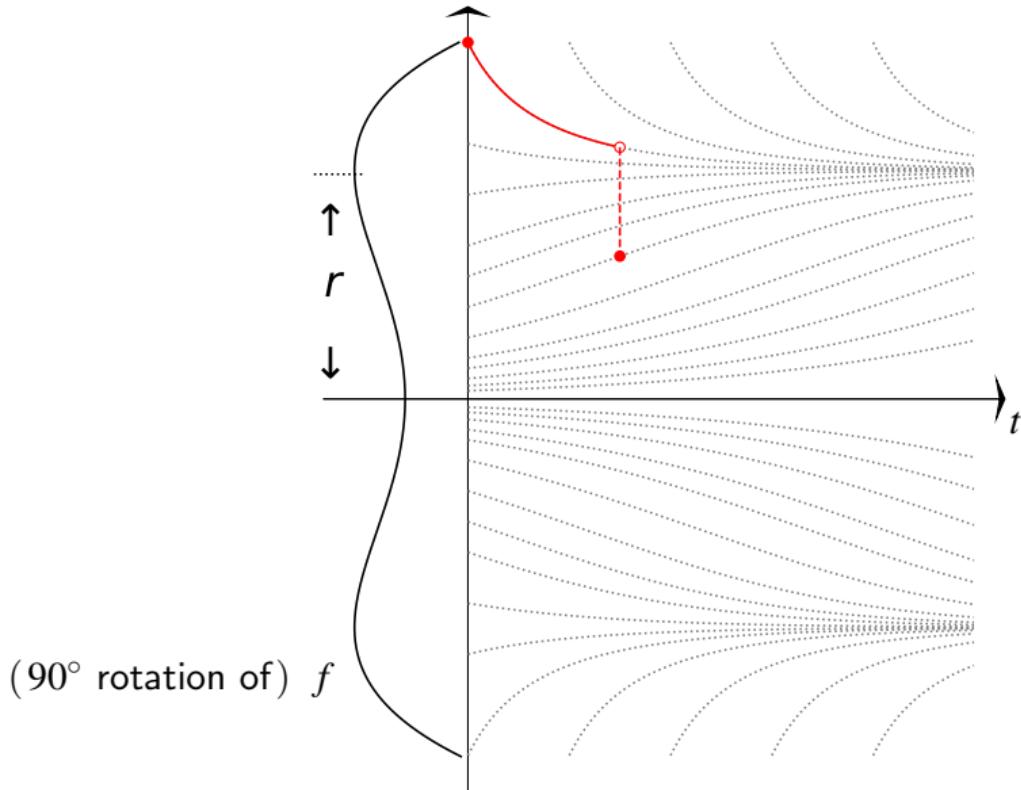
$$\mathbf{P}(W^\eta \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

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

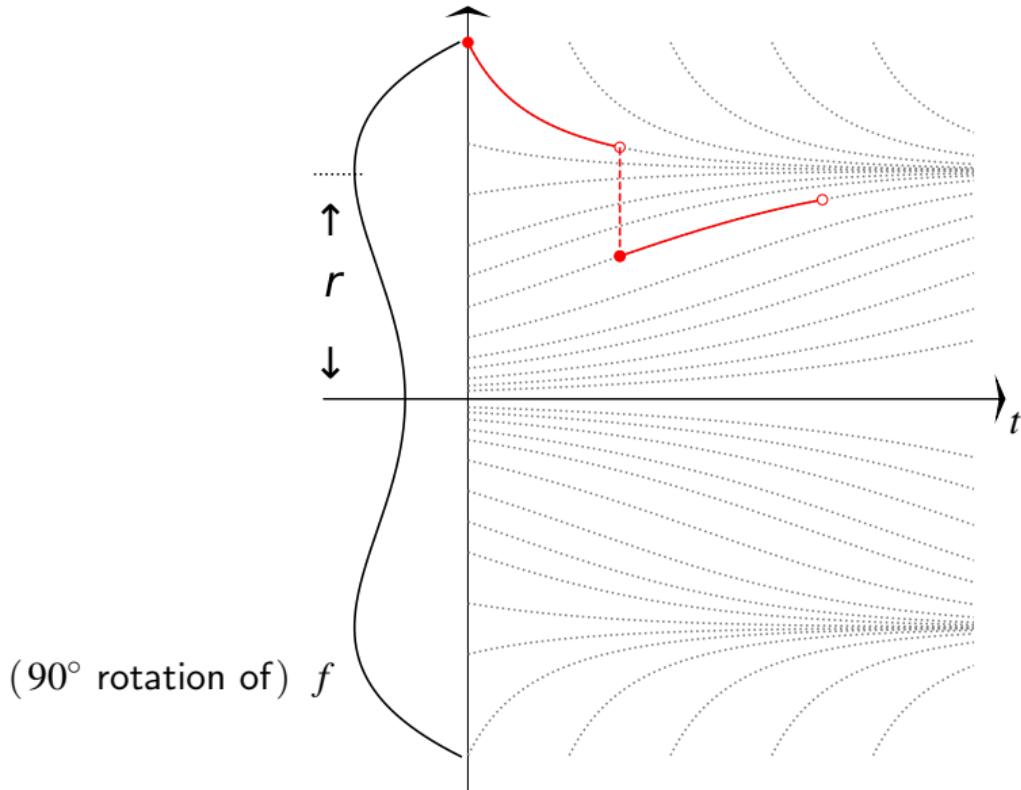
## Adding Jumps to $w(\cdot)$ : Piecewise Gradient Flow



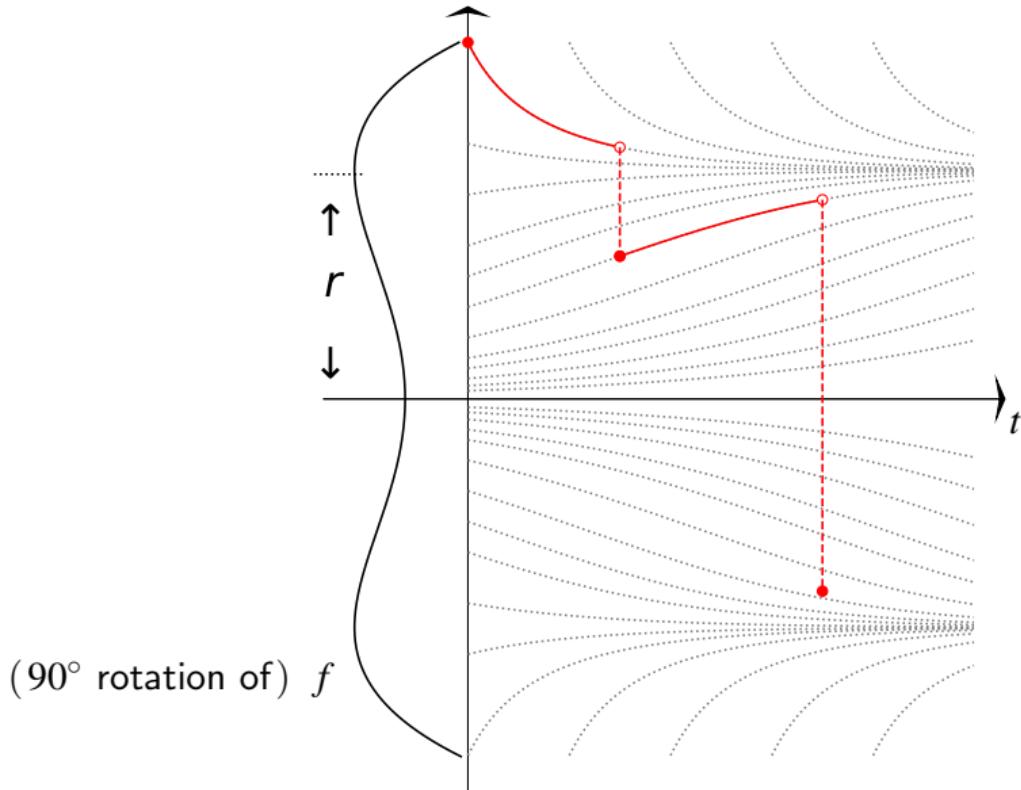
## Adding Jumps to $w(\cdot)$ : Piecewise Gradient Flow



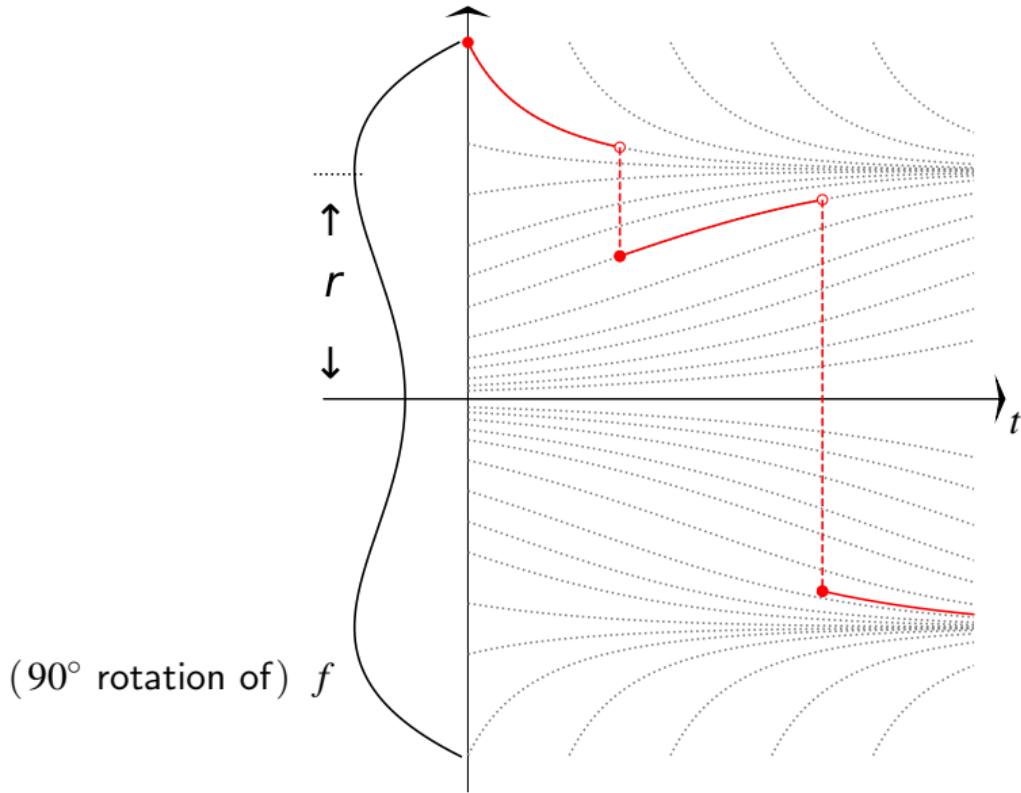
## Adding Jumps to $w(\cdot)$ : Piecewise Gradient Flow



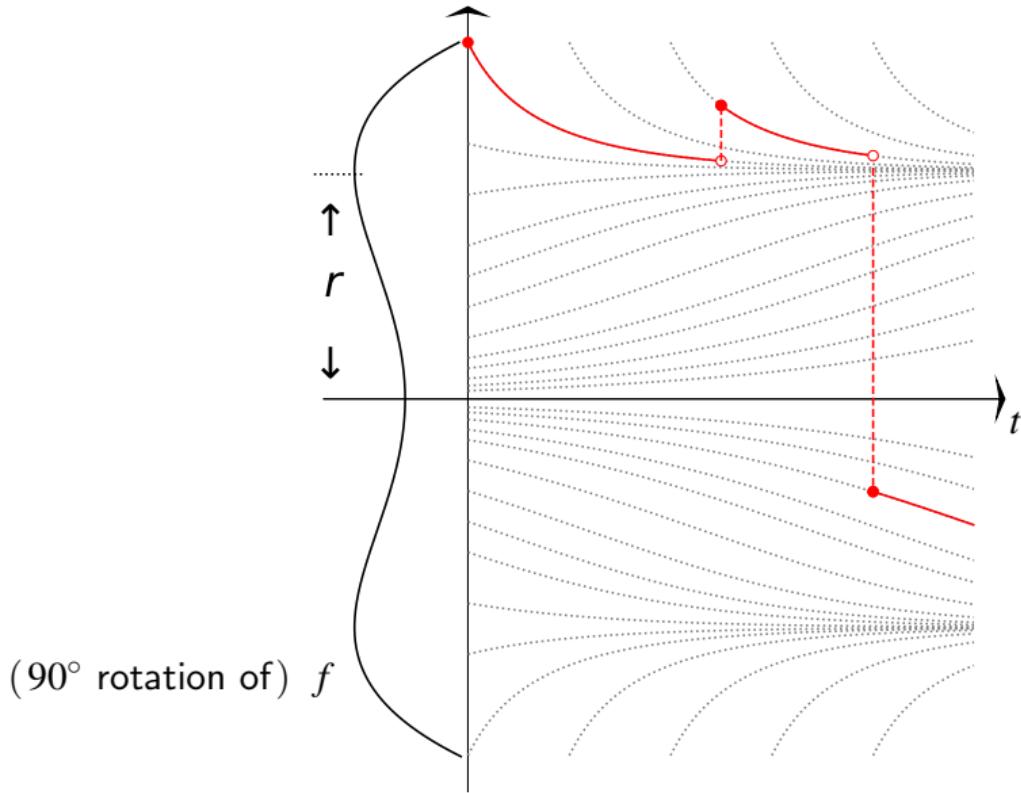
## Adding Jumps to $w(\cdot)$ : Piecewise Gradient Flow



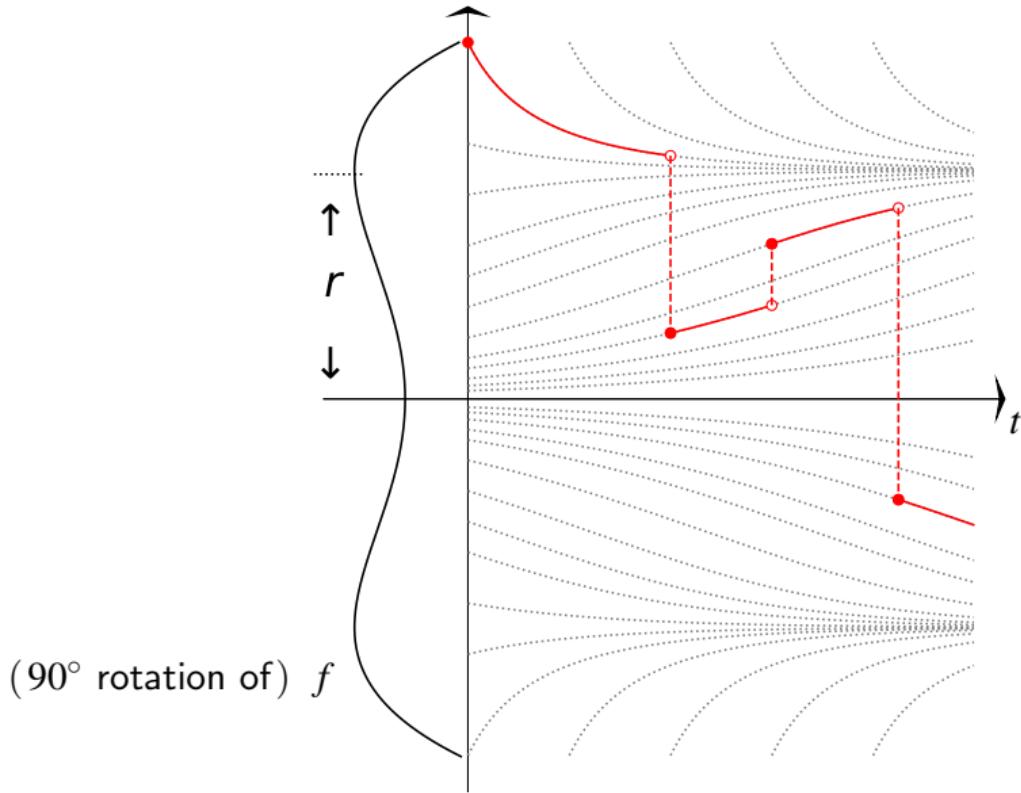
## Adding Jumps to $w(\cdot)$ : Piecewise Gradient Flow



## 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+)

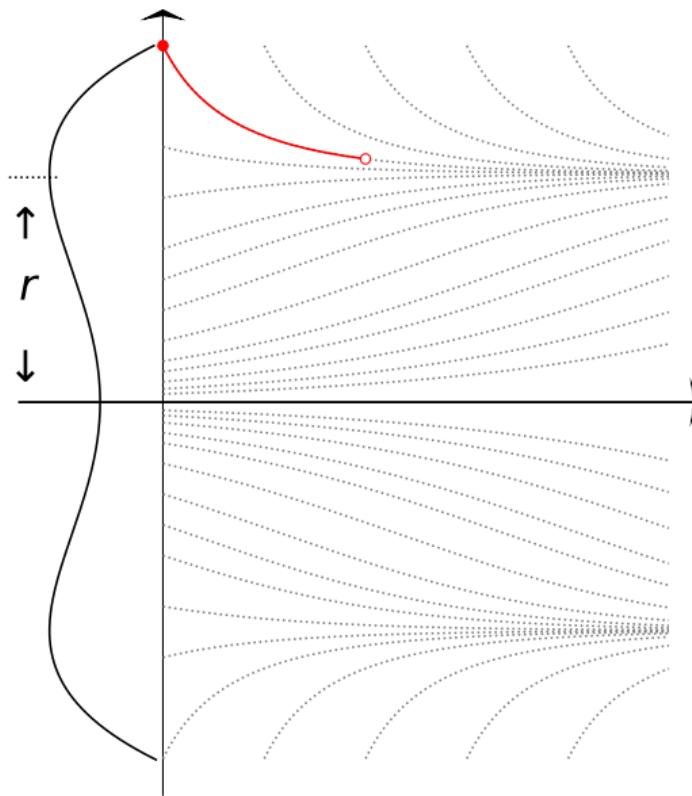
For “general”  $A \subseteq \mathbb{D}$

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

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

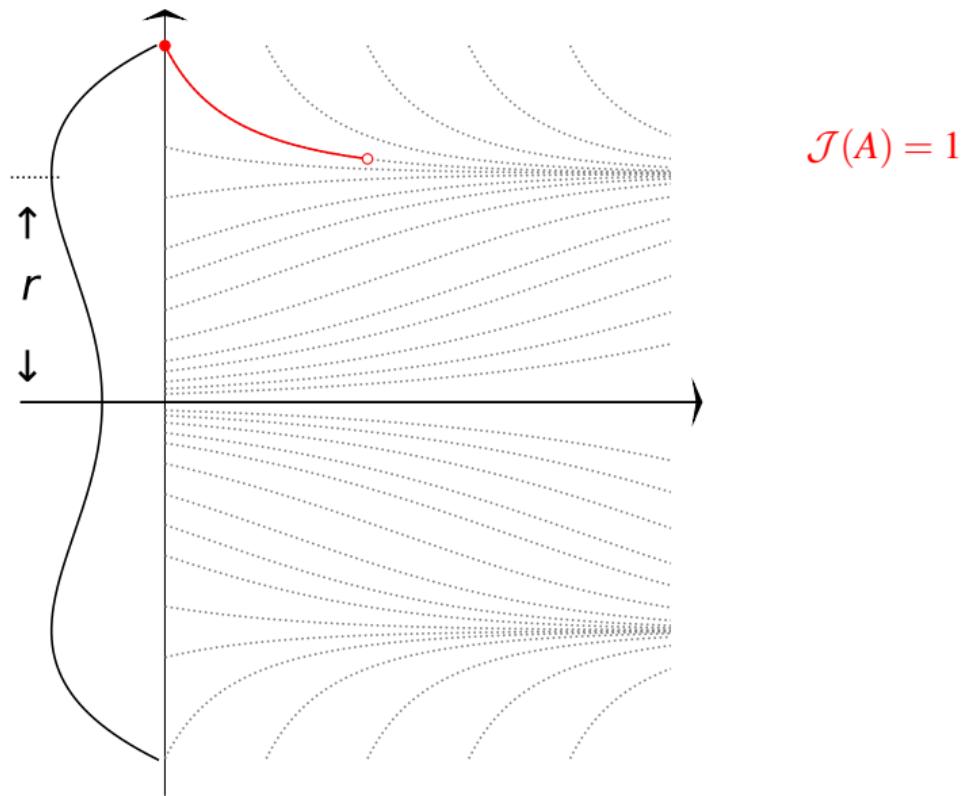
## Catastrophe Principle: Most Likely Escape Route

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



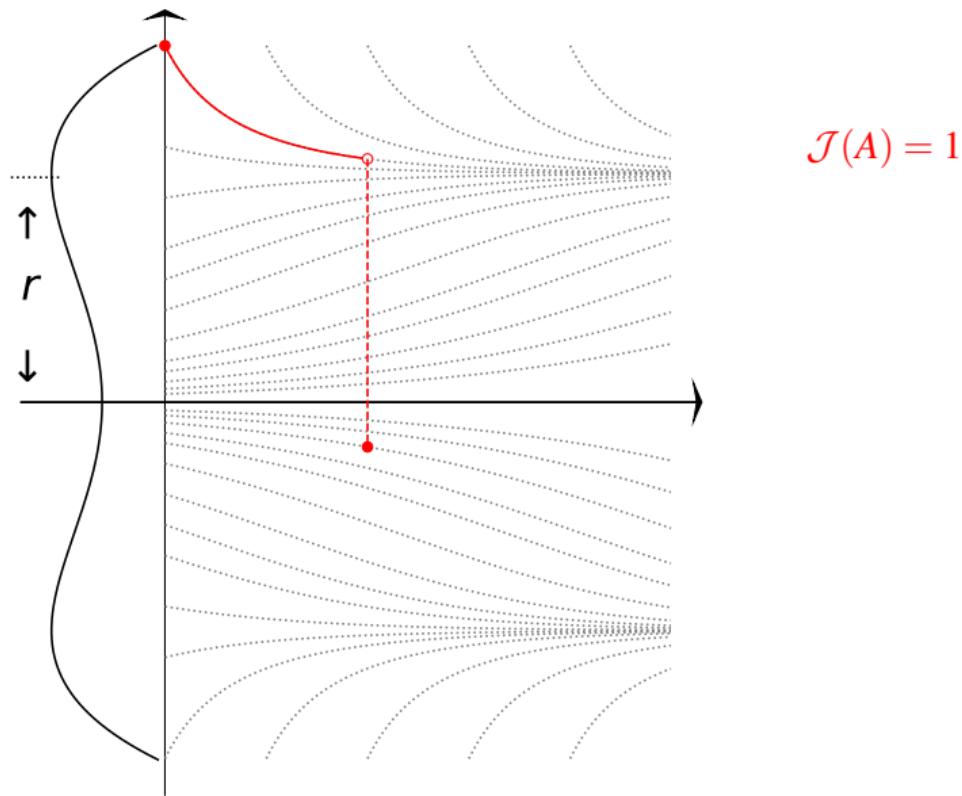
## Catastrophe Principle: Most Likely Escape Route

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



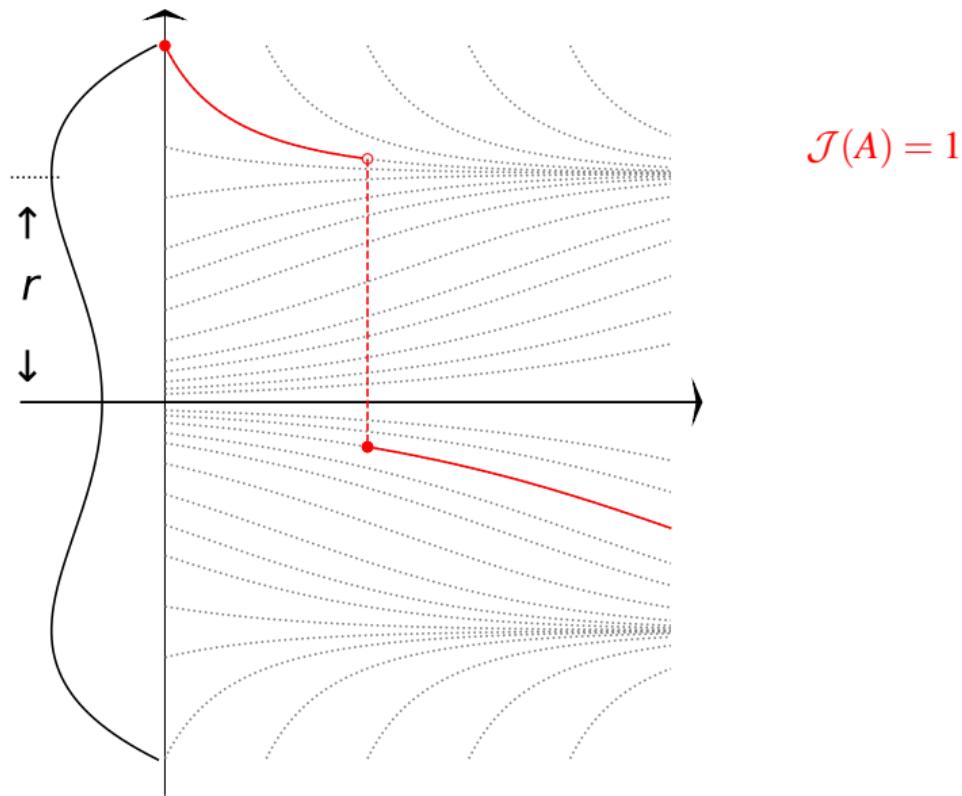
# Catastrophe Principle: Most Likely Escape Route

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



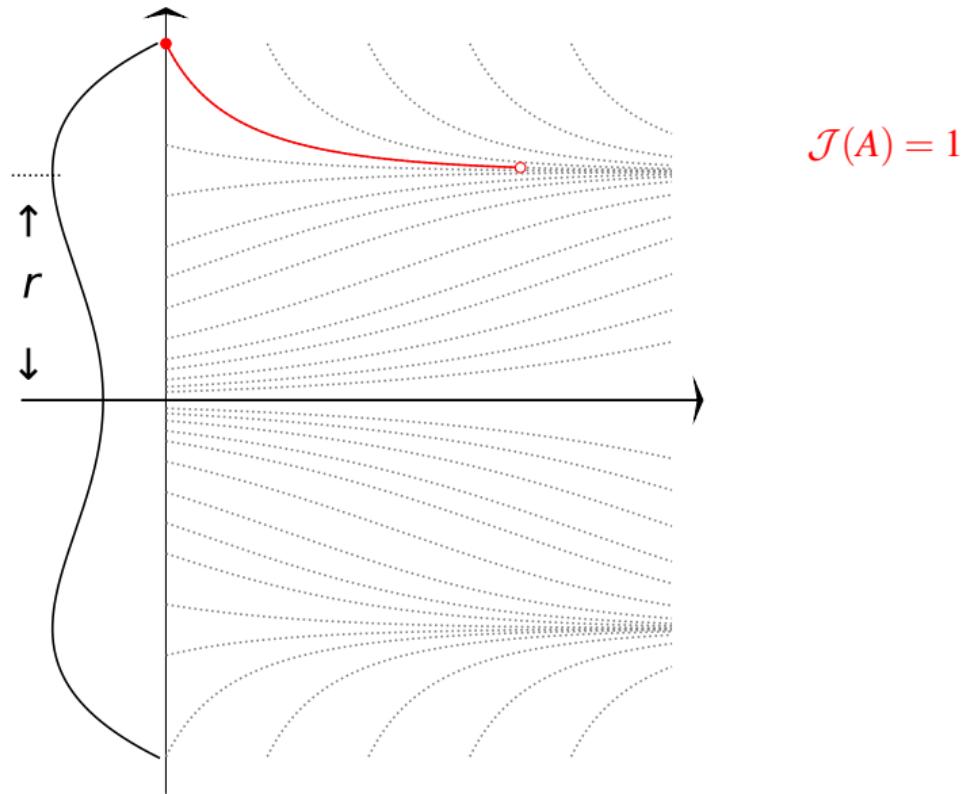
# Catastrophe Principle: Most Likely Escape Route

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



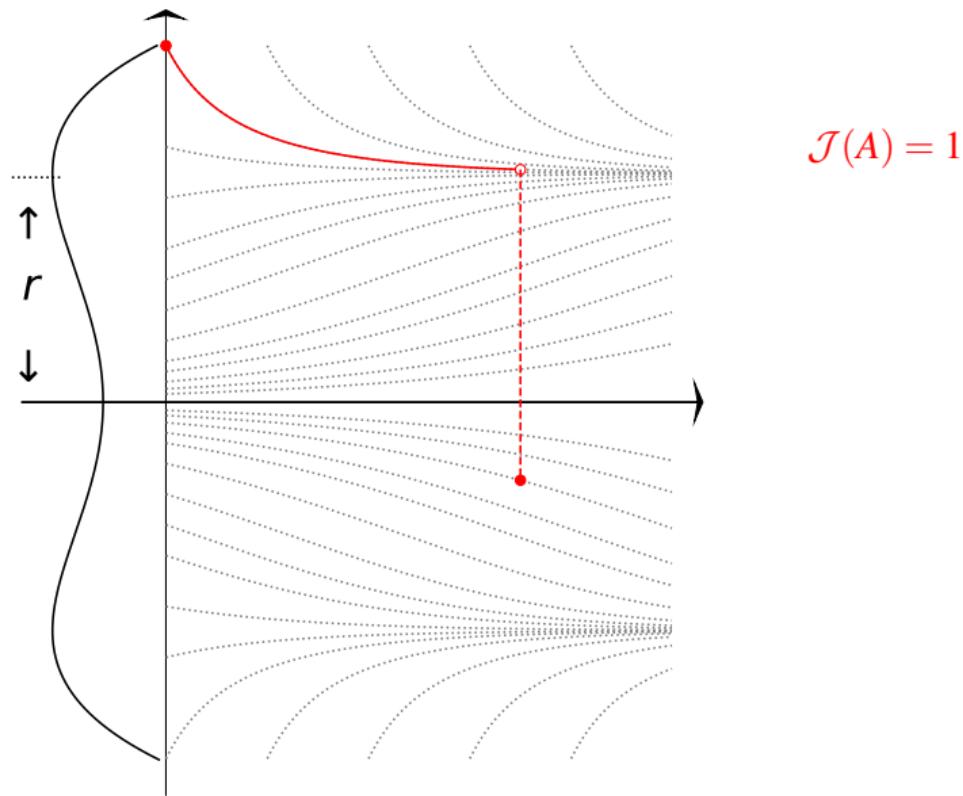
## Catastrophe Principle: Most Likely Escape Route

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



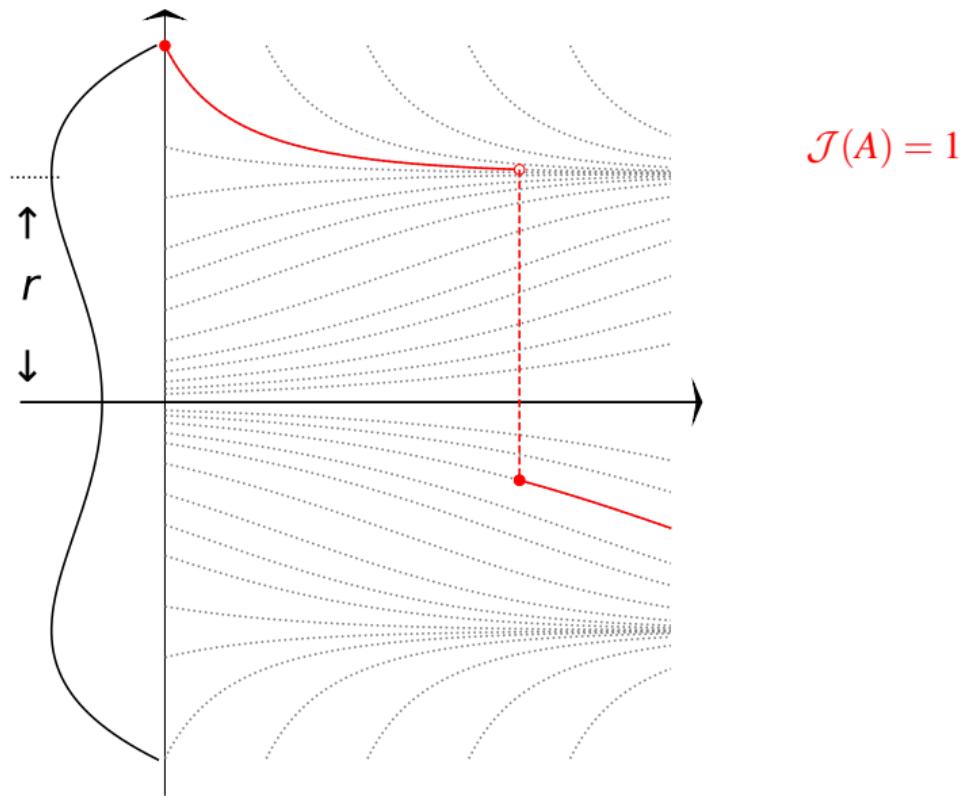
# Catastrophe Principle: Most Likely Escape Route

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



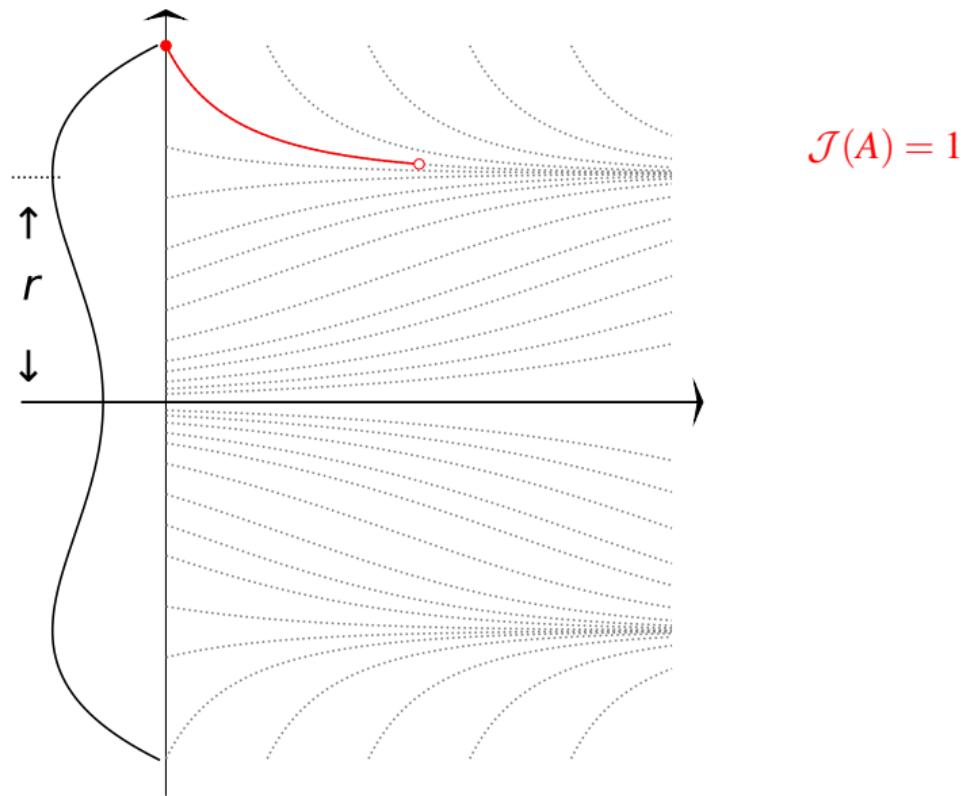
# Catastrophe Principle: Most Likely Escape Route

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



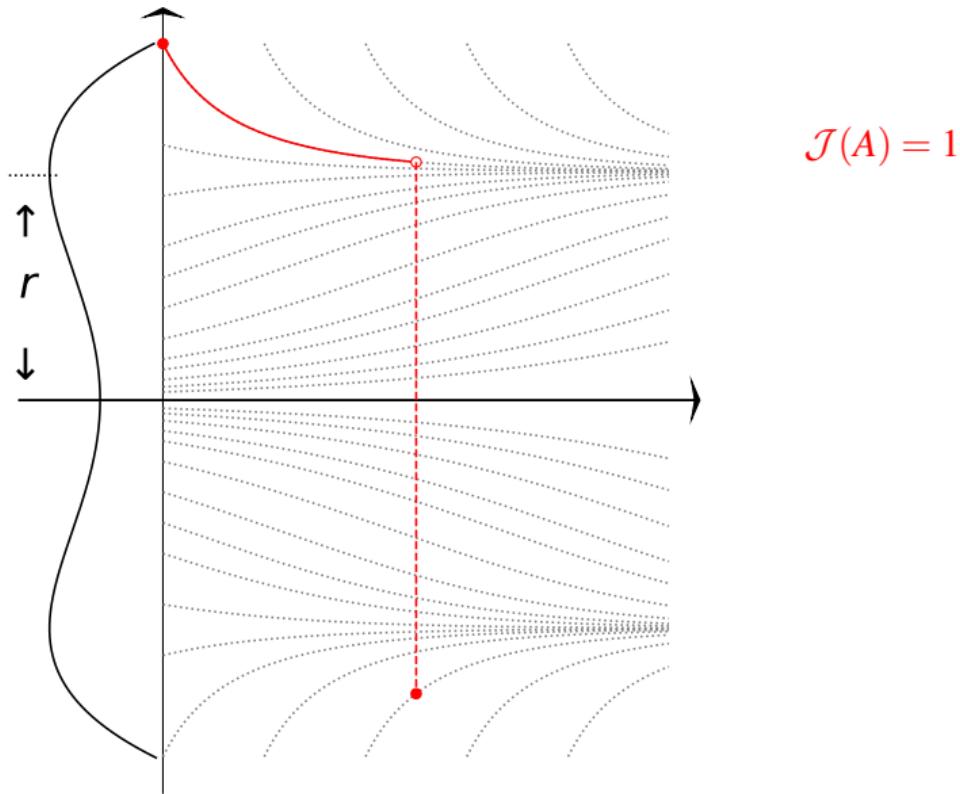
## Catastrophe Principle: Most Likely Escape Route

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



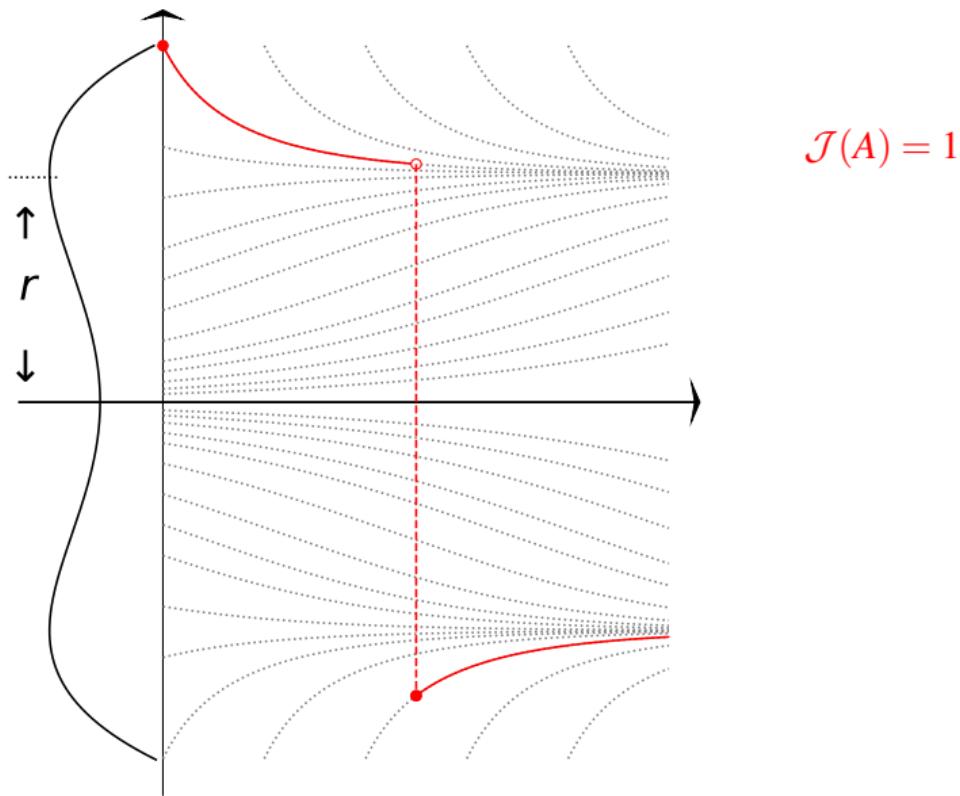
## Catastrophe Principle: Most Likely Escape Route

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



# Catastrophe Principle: Most Likely Escape Route

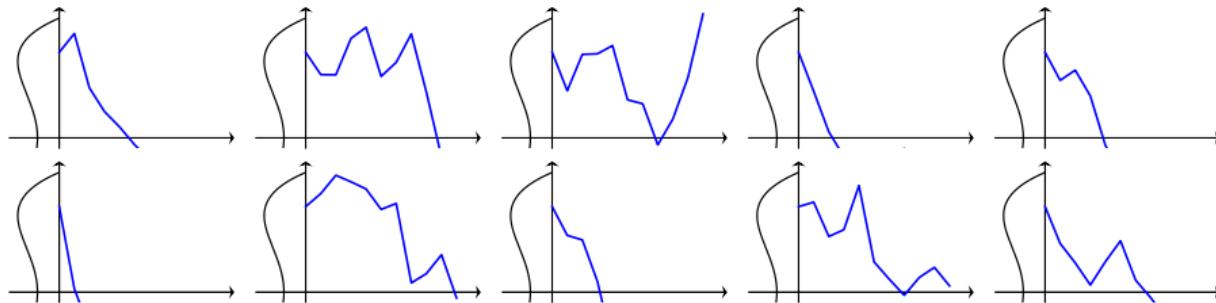
Most Likely Paths for  $\{W^\eta \text{ escapes the local minimum}\}$



# Catastrophe Principle Dictates SGD's Escape Route

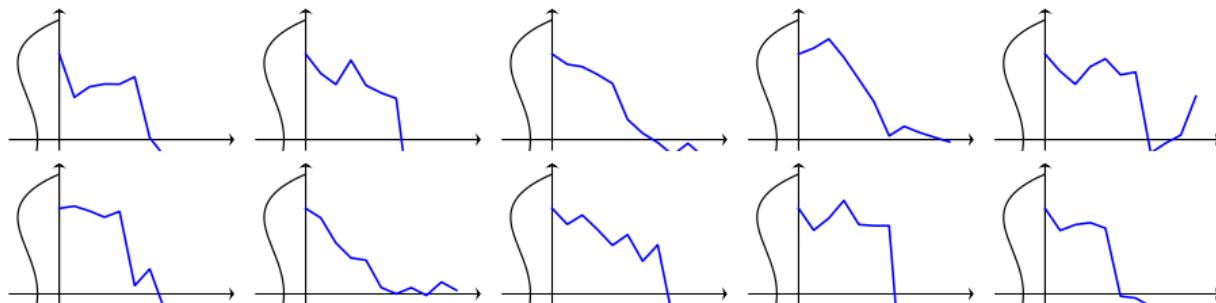
Trajectory of SGD  $X^\eta$  conditional on exit:

**light-tailed** noises with  $\eta = 1/10$



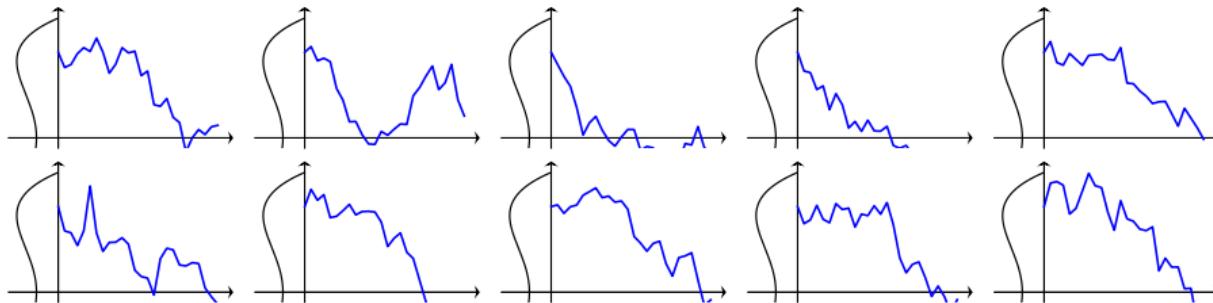
Trajectory of SGD  $X^\eta$  conditional on exit:

**heavy-tailed** noises with  $\eta = 1/10$



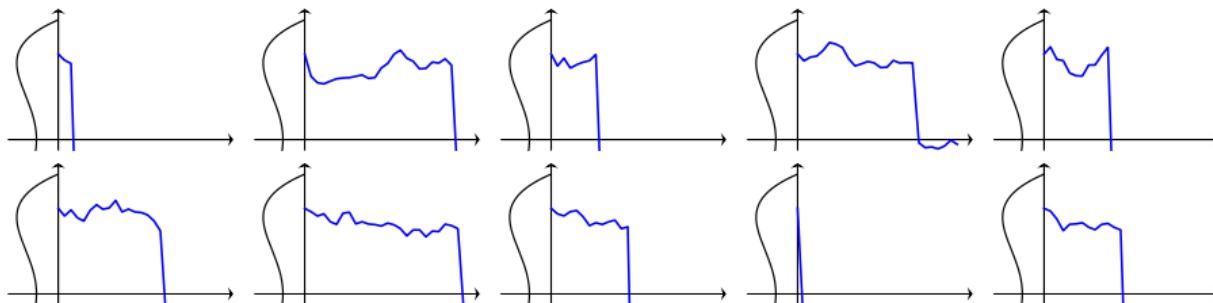
# Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/25$

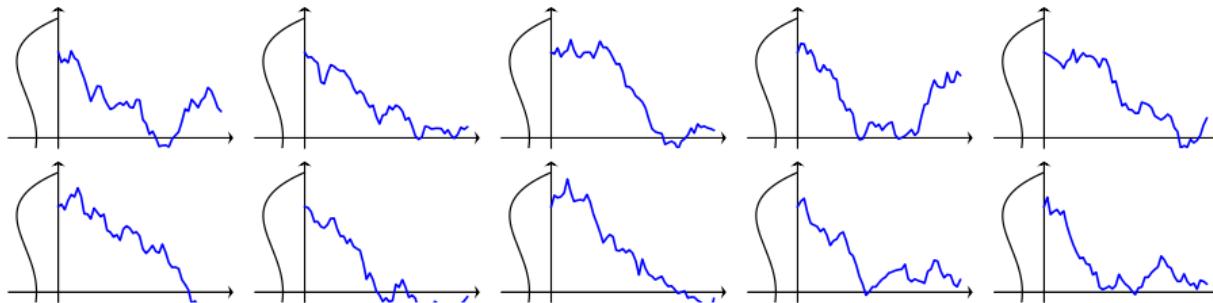
Trajectory of SGD  $X^\eta$  conditional on exit:



**heavy-tailed** noises with  $\eta = 1/25$

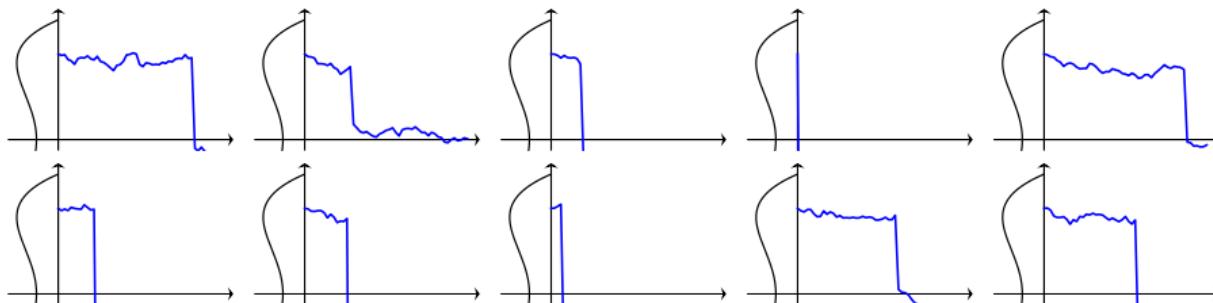
# Catastrophe Principle Dictates SGD's Escape Route

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/50$

Trajectory of SGD  $X^\eta$  conditional on exit:

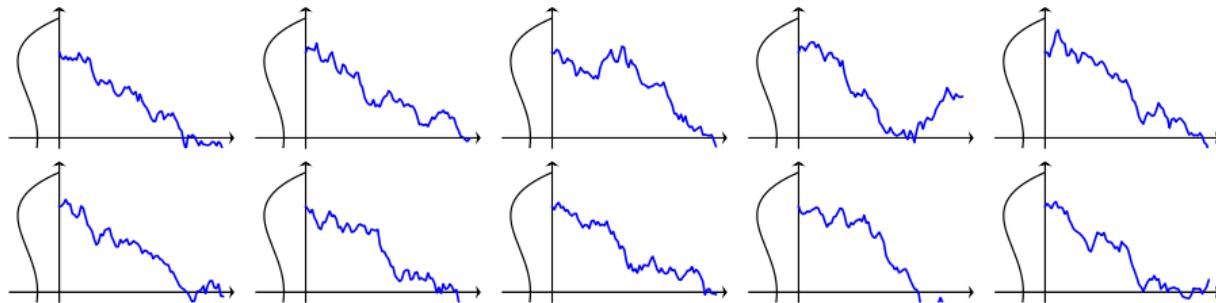


**heavy-tailed** noises with  $\eta = 1/50$

# Catastrophe Principle Dictates SGD's Escape Route

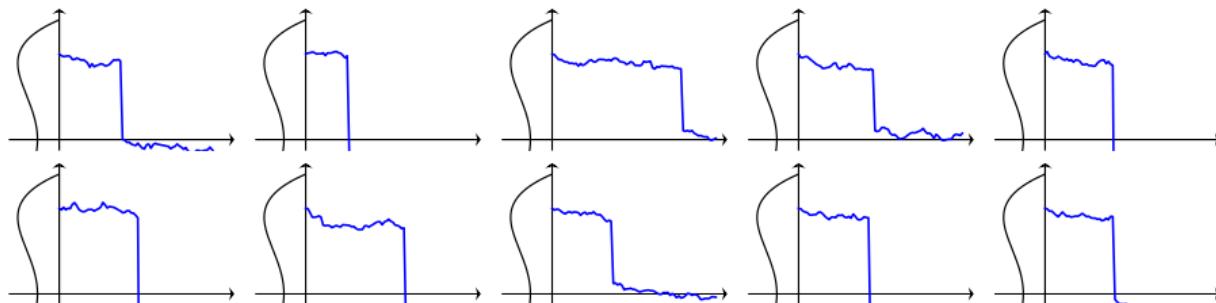
Trajectory of SGD  $X^\eta$  conditional on exit:

**light-tailed** noises with  $\eta = 1/75$



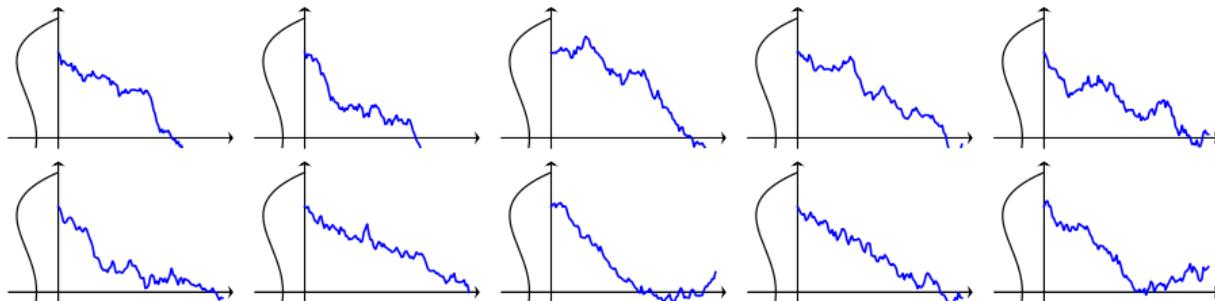
Trajectory of SGD  $X^\eta$  conditional on exit:

**heavy-tailed** noises with  $\eta = 1/75$

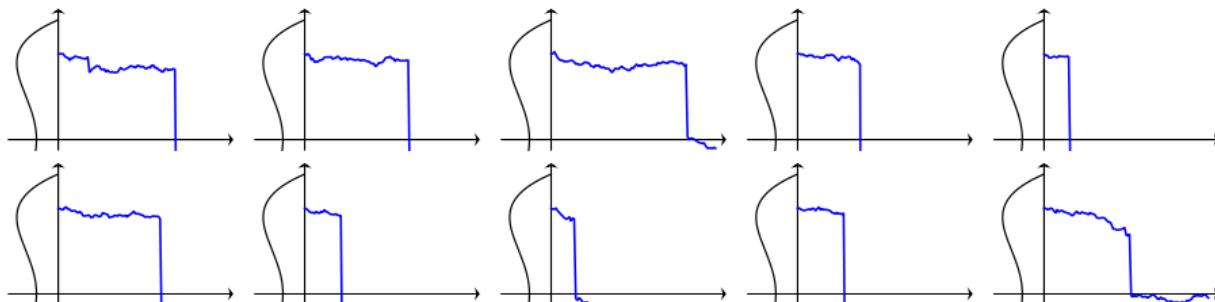


# Catastrophe Principle Dictates SGD's Escape Route

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

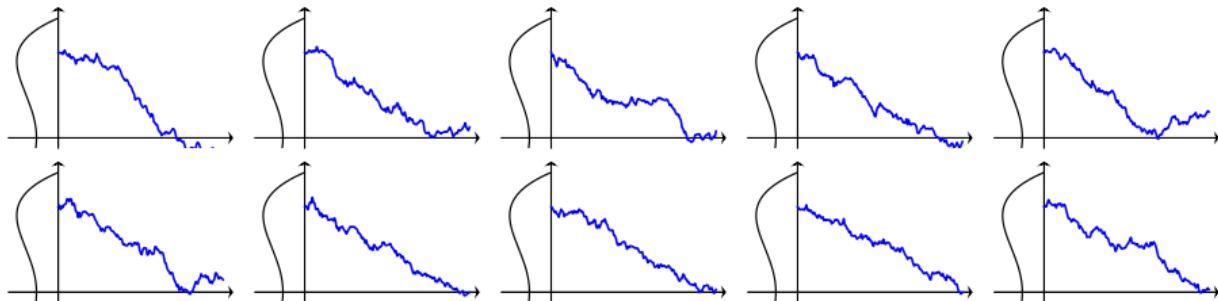


Trajectory of SGD  $X^\eta$  conditional on exit: **heavy-tailed** noises with  $\eta = 1/100$

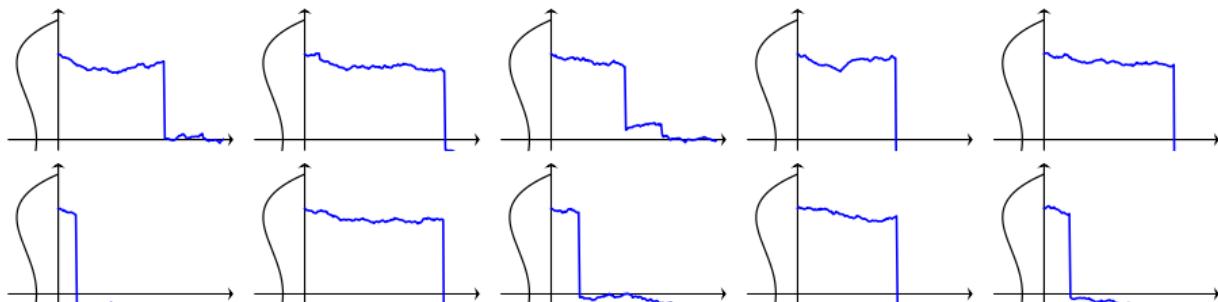


# Catastrophe Principle Dictates SGD's Escape Route

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

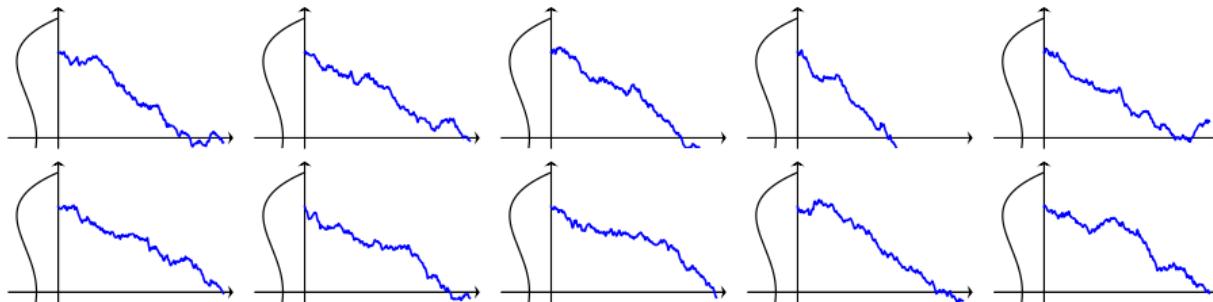


Trajectory of SGD  $X^\eta$  conditional on exit: **heavy-tailed** noises with  $\eta = 1/150$

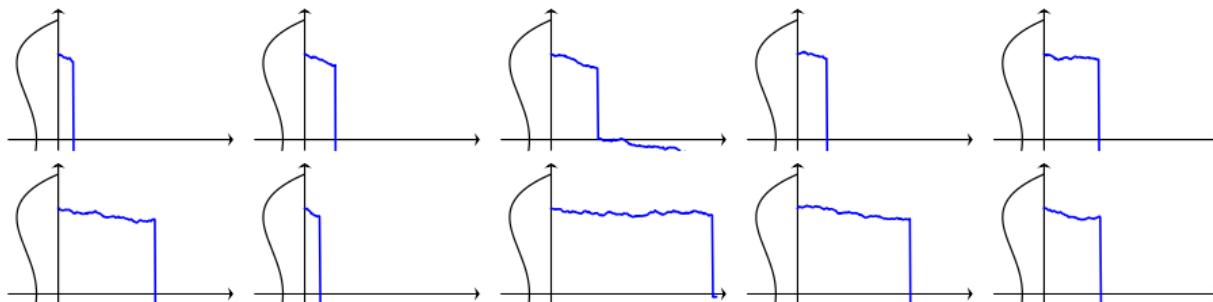


# Catastrophe Principle Dictates SGD's Escape Route

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

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

## Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

## Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

## Truncated Version of Stochastic Gradient Descent

### SGD with Gradient Clipping

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

where

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

Then, again,

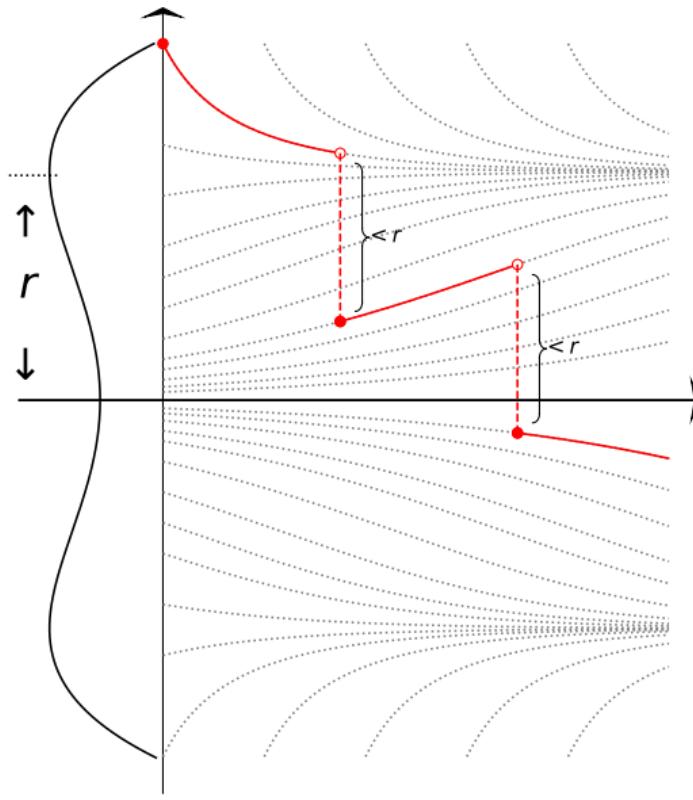
$$W^\eta(\cdot) \rightarrow w(\cdot) \quad \text{as} \quad \eta \rightarrow 0$$

where

$$dw(t) = -f'(w(t))dt.$$

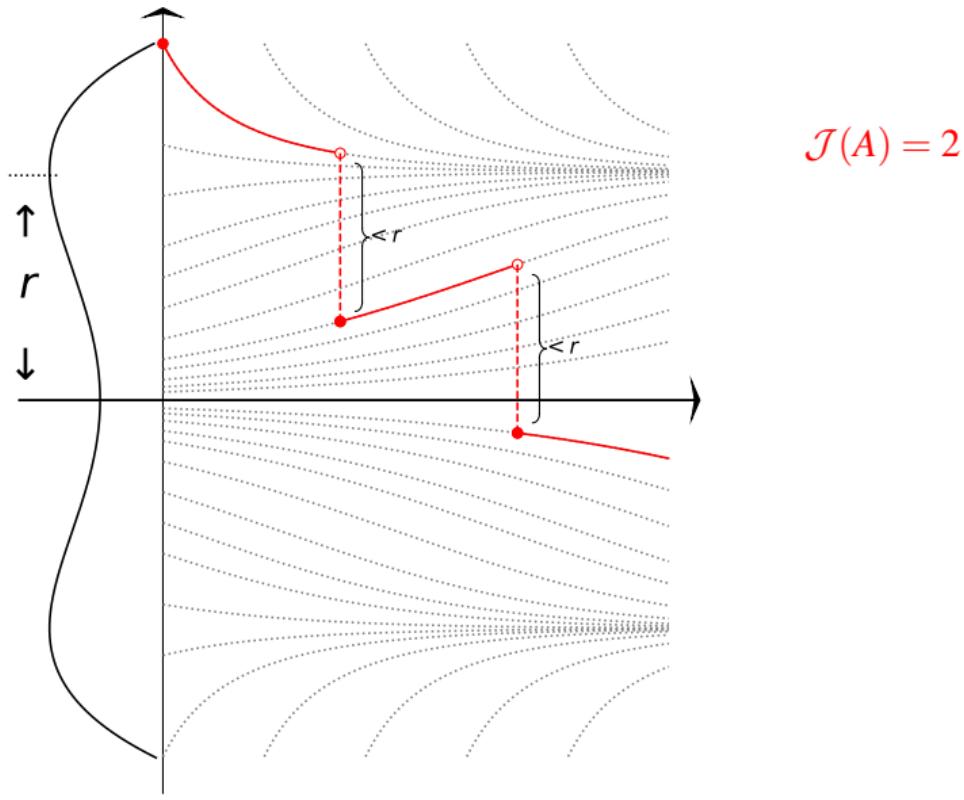
# How does $\mathcal{J}(A)$ change?

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



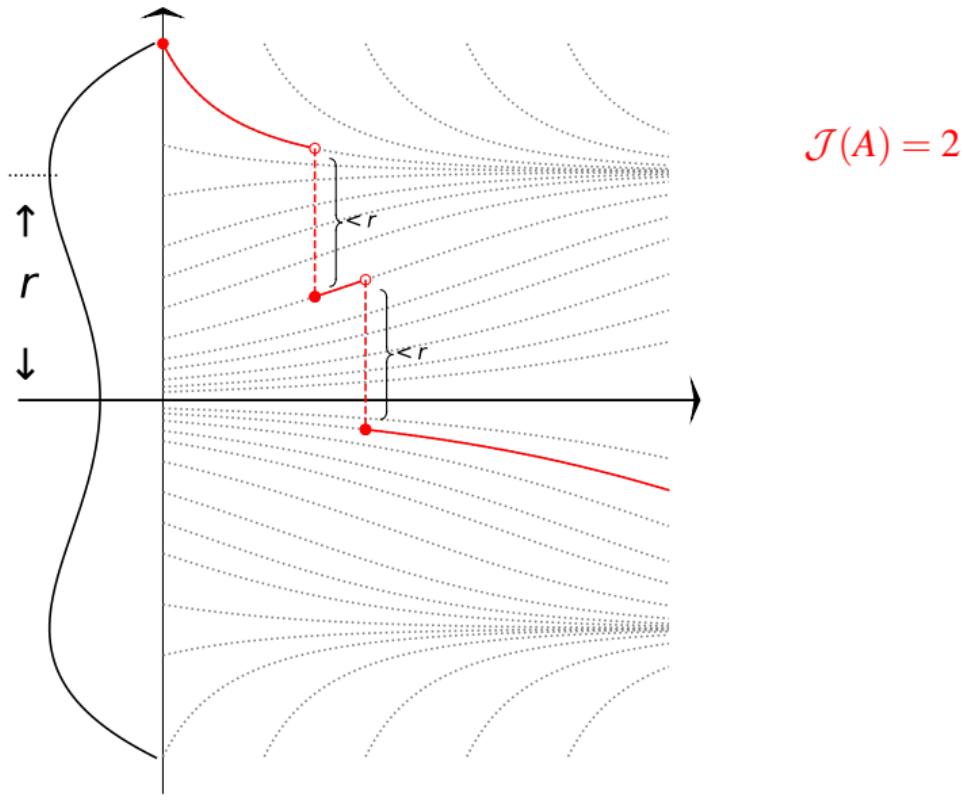
# How does $\mathcal{J}(A)$ change?

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



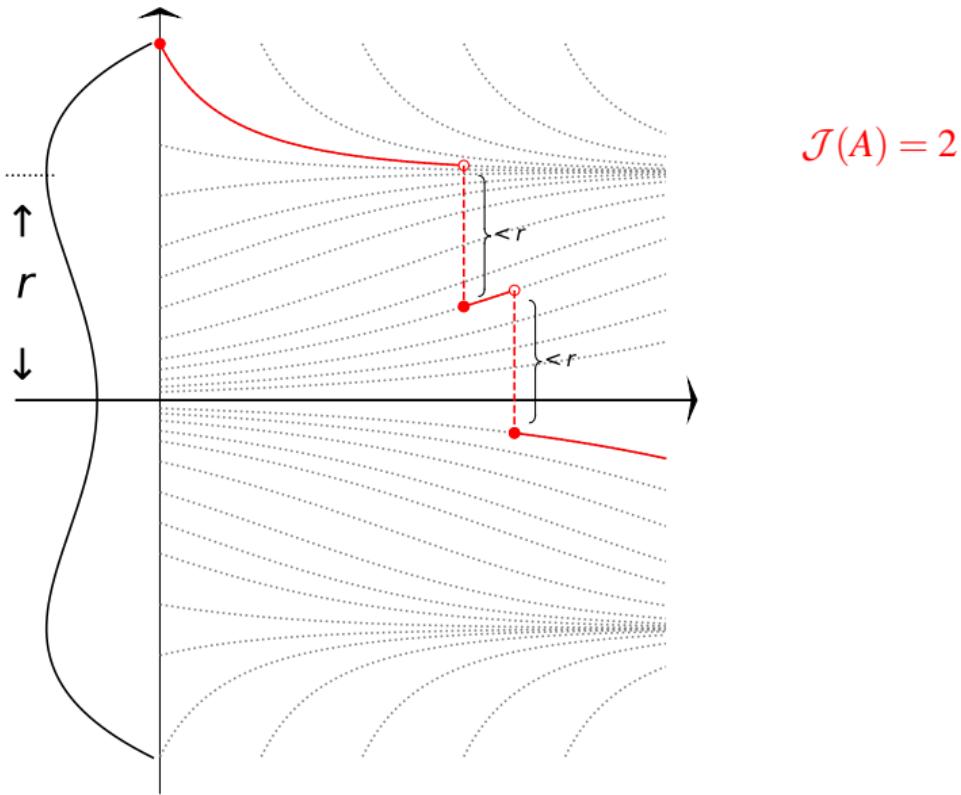
# How does $\mathcal{J}(A)$ change?

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



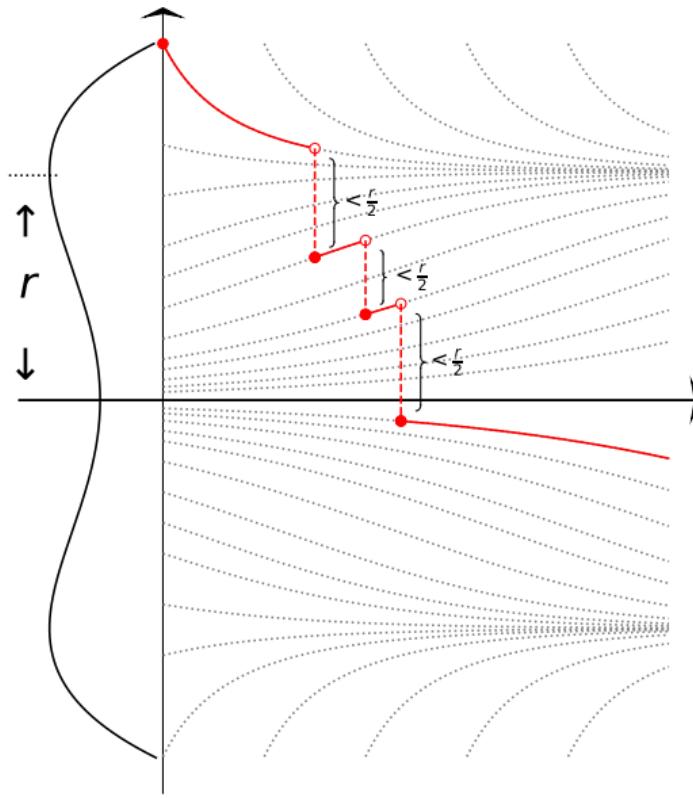
# How does $\mathcal{J}(A)$ change?

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



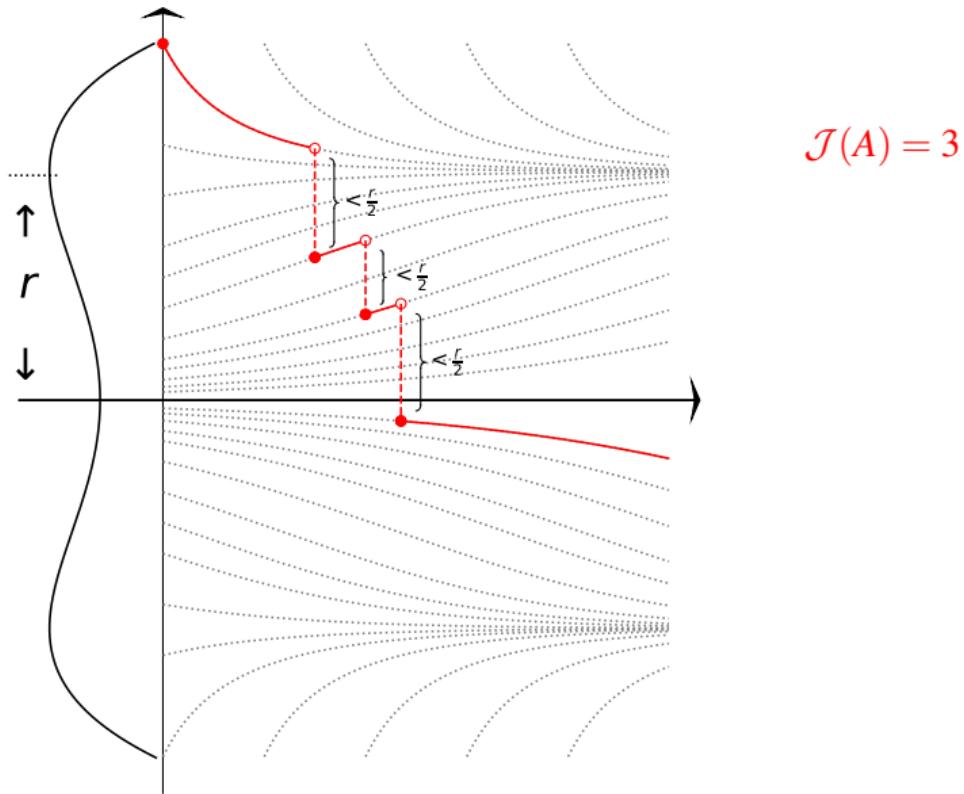
# How does $\mathcal{J}(A)$ change?

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



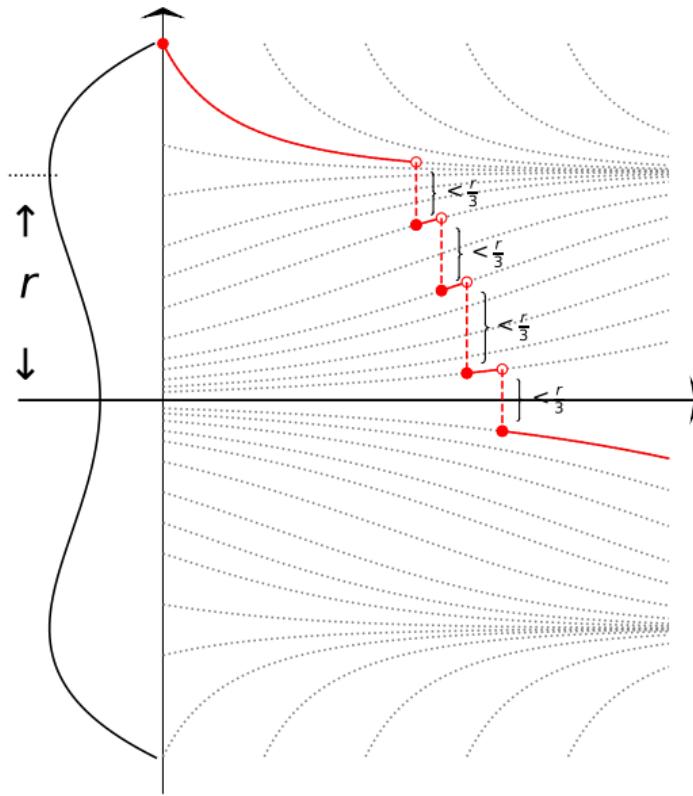
# How does $\mathcal{J}(A)$ change?

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



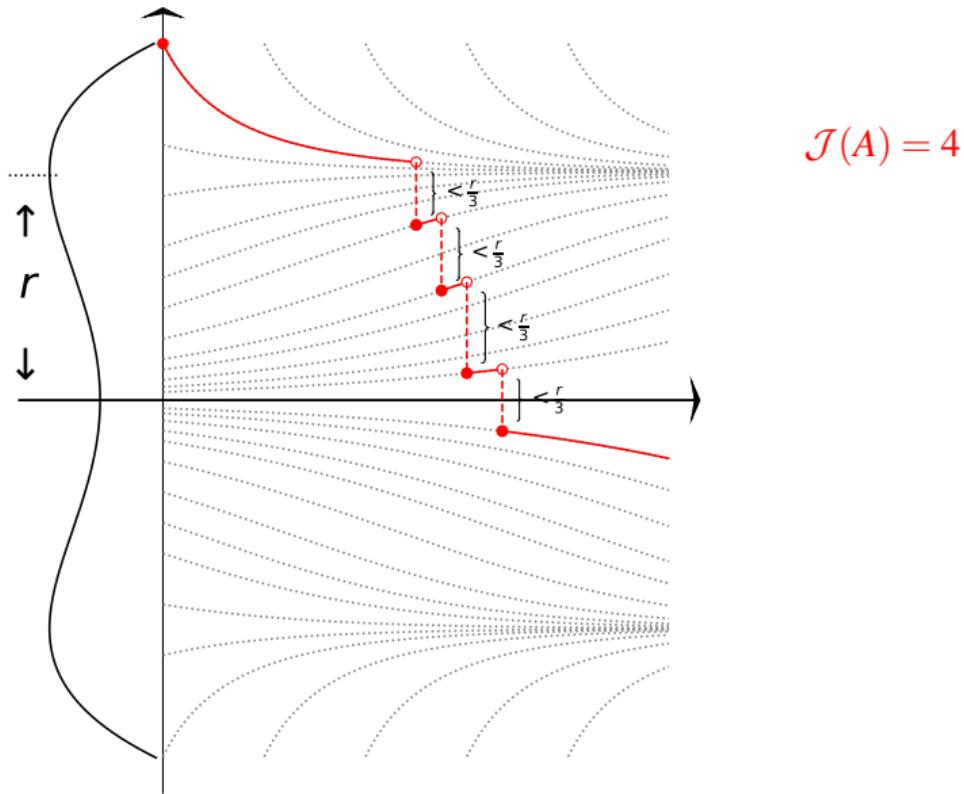
# How does $\mathcal{J}(A)$ change?

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



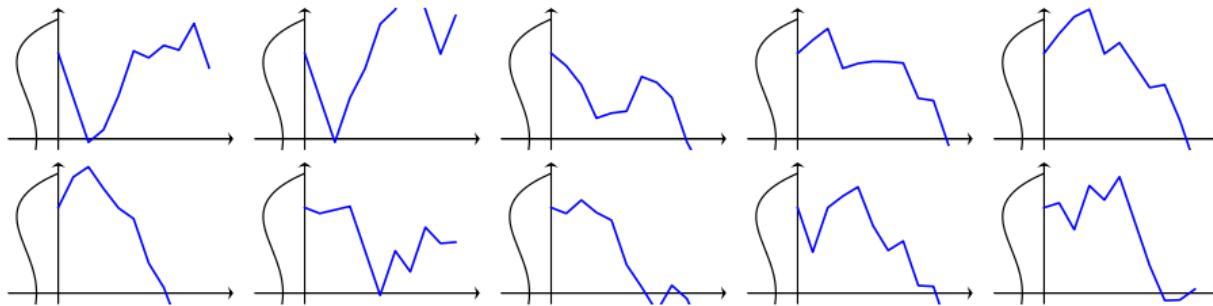
# How does $\mathcal{J}(A)$ change?

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



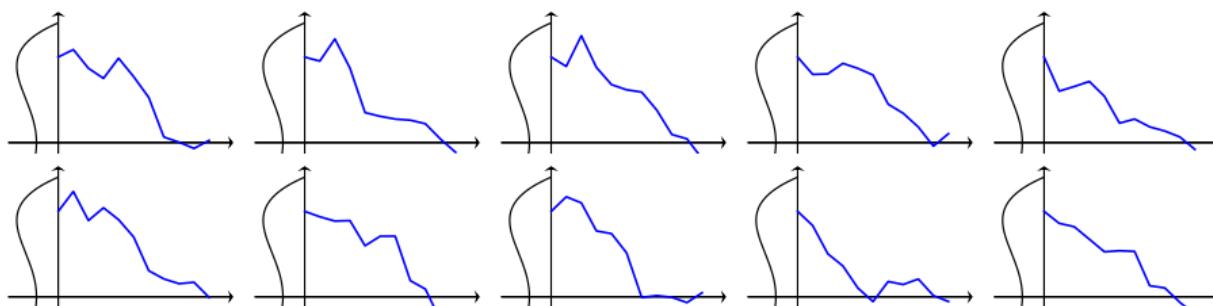
# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/10$

Trajectory of SGD  $X^\eta$  conditional on exit:

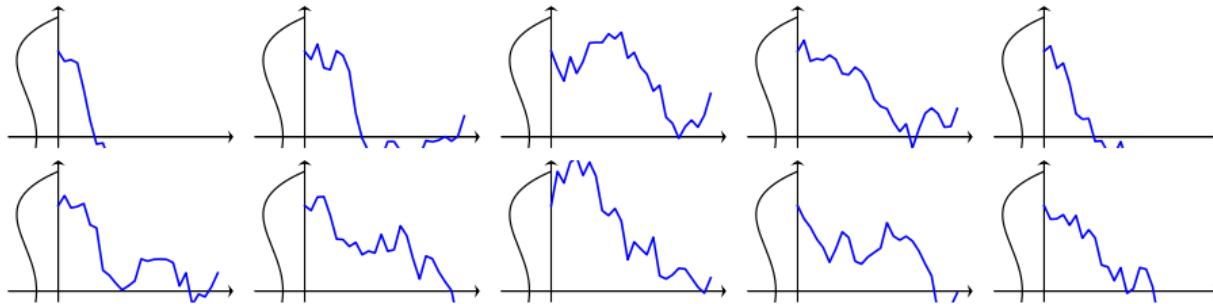


**heavy-tailed** noises with  $\eta = 1/10$

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

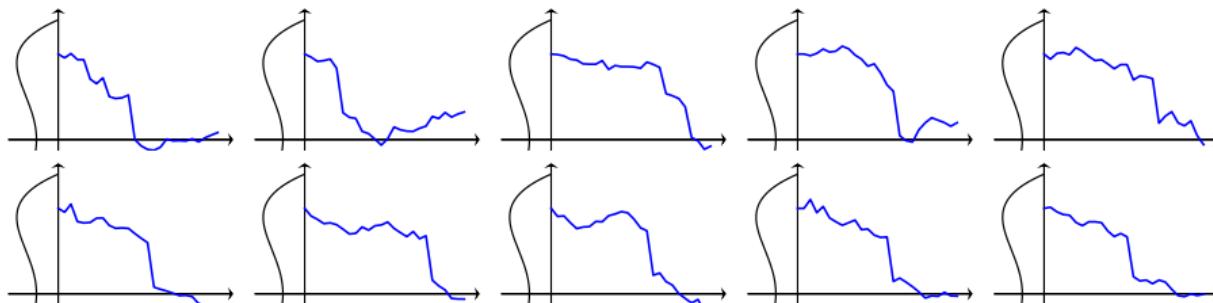
Trajectory of SGD  $X^\eta$  conditional on exit:

**light-tailed** noises with  $\eta = 1/25$



Trajectory of SGD  $X^\eta$  conditional on exit:

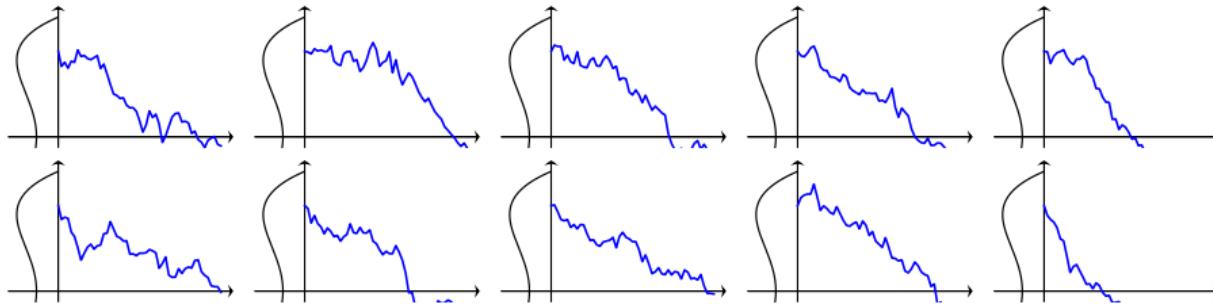
**heavy-tailed** noises with  $\eta = 1/25$



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

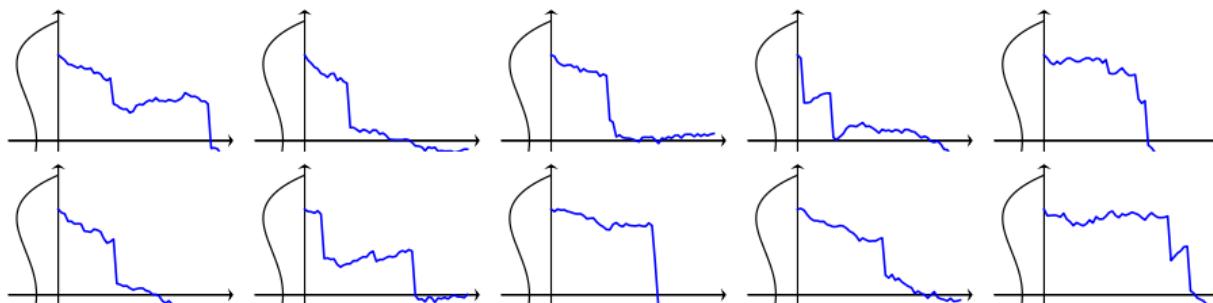
Trajectory of SGD  $X^\eta$  conditional on exit:

**light-tailed** noises with  $\eta = 1/50$



Trajectory of SGD  $X^\eta$  conditional on exit:

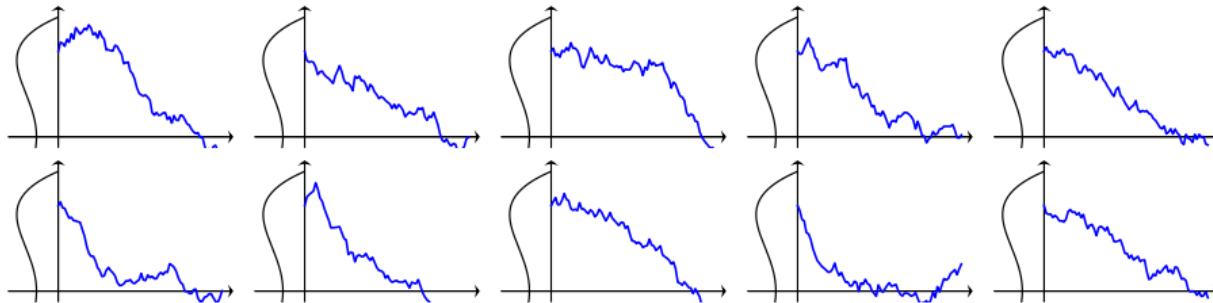
**heavy-tailed** noises with  $\eta = 1/10$



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

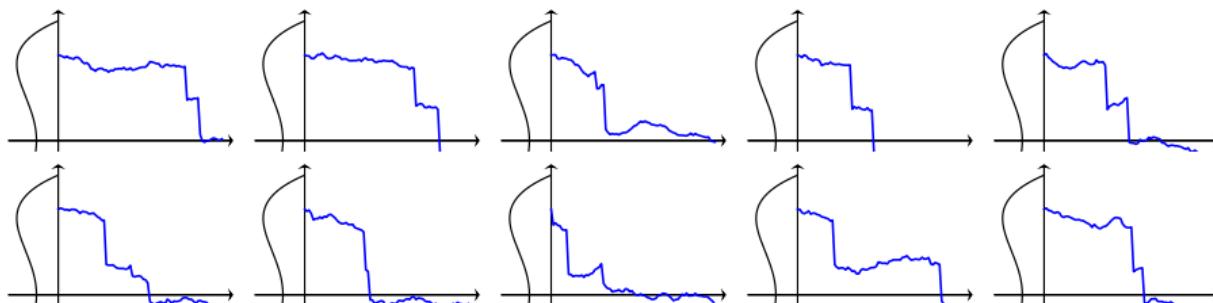
Trajectory of SGD  $X^\eta$  conditional on exit:

**light-tailed** noises with  $\eta = 1/75$



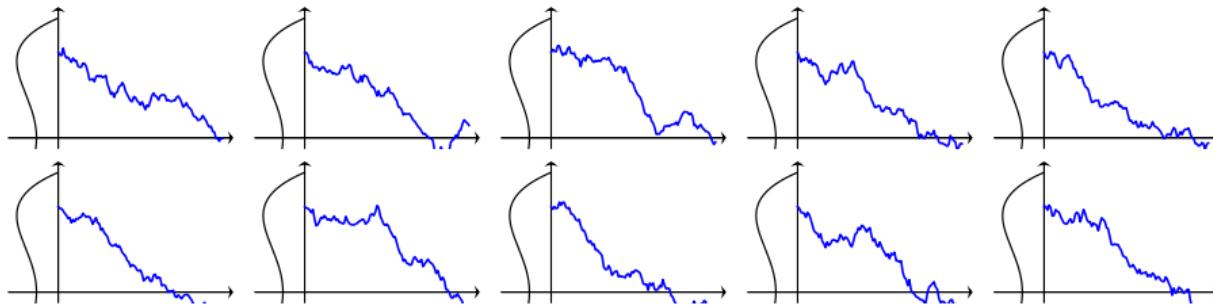
Trajectory of SGD  $X^\eta$  conditional on exit:

**heavy-tailed** noises with  $\eta = 1/75$



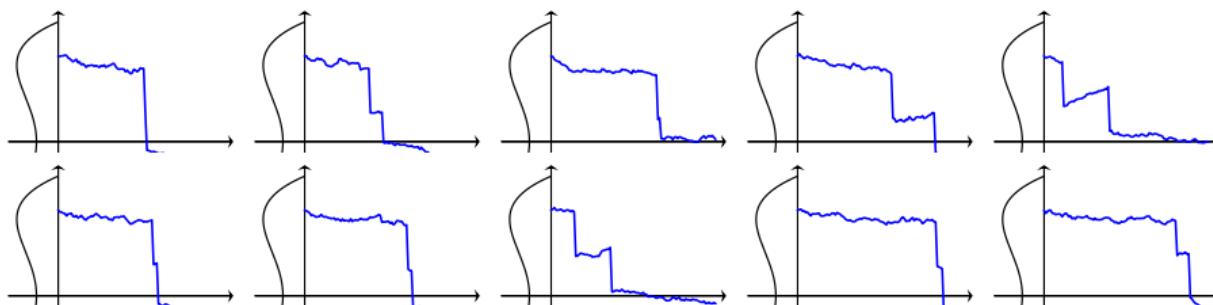
# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/100$

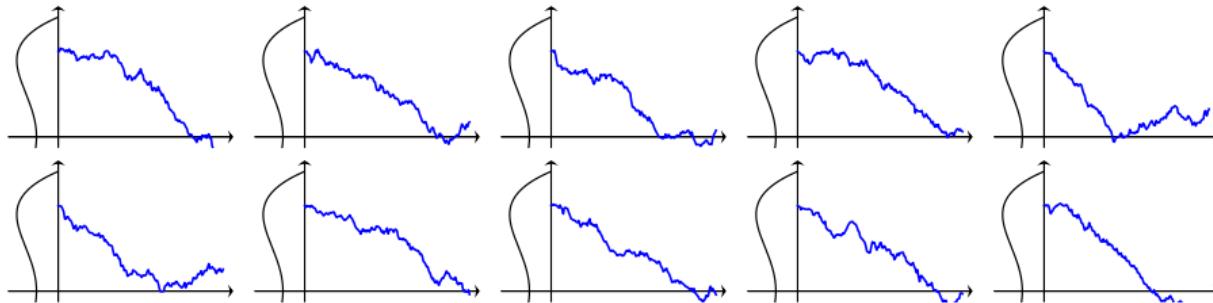
Trajectory of SGD  $X^\eta$  conditional on exit:



**heavy-tailed** noises with  $\eta = 1/100$

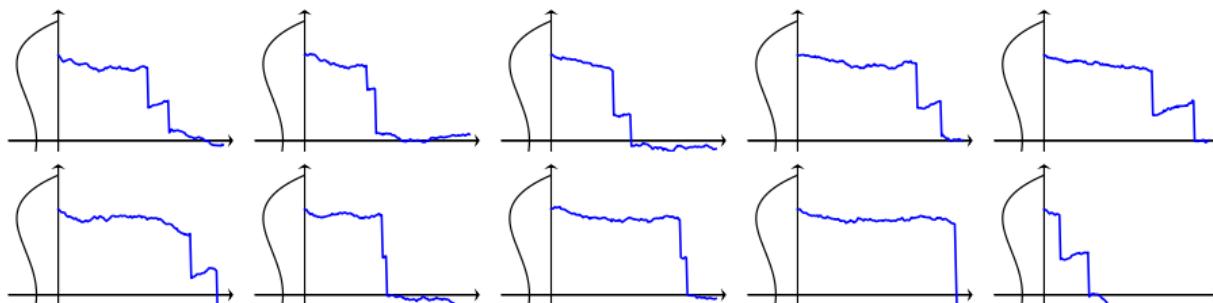
# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/150$

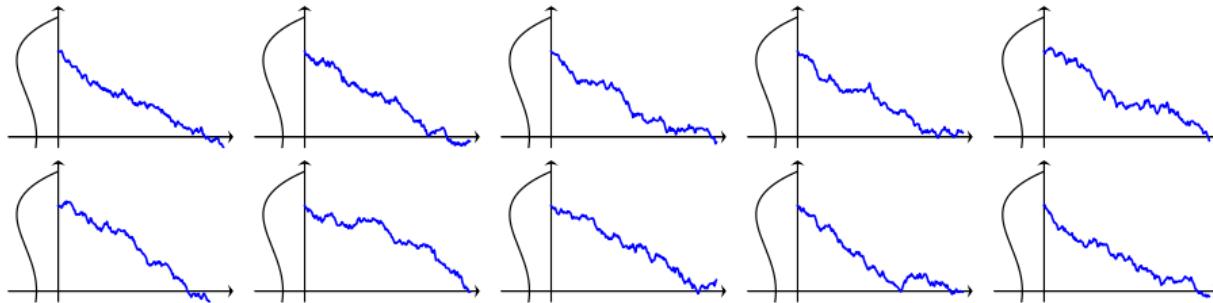
Trajectory of SGD  $X^\eta$  conditional on exit:



**heavy-tailed** noises with  $\eta = 1/150$

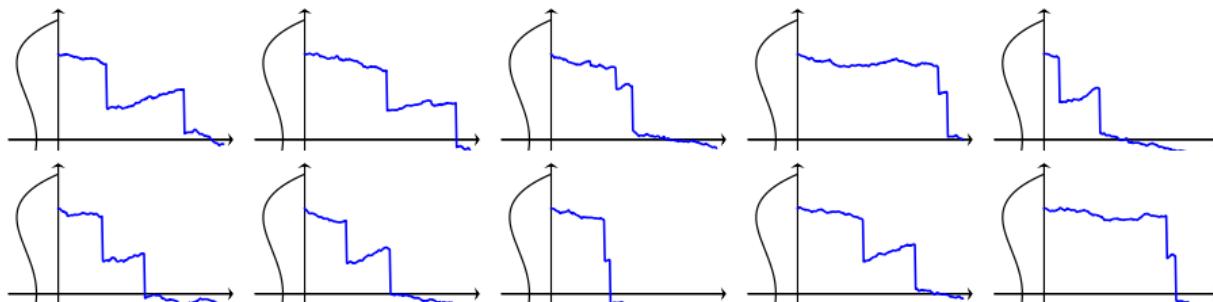
# SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

Trajectory of SGD  $X^\eta$  conditional on exit:



**light-tailed** noises with  $\eta = 1/200$

Trajectory of SGD  $X^\eta$  conditional on exit:

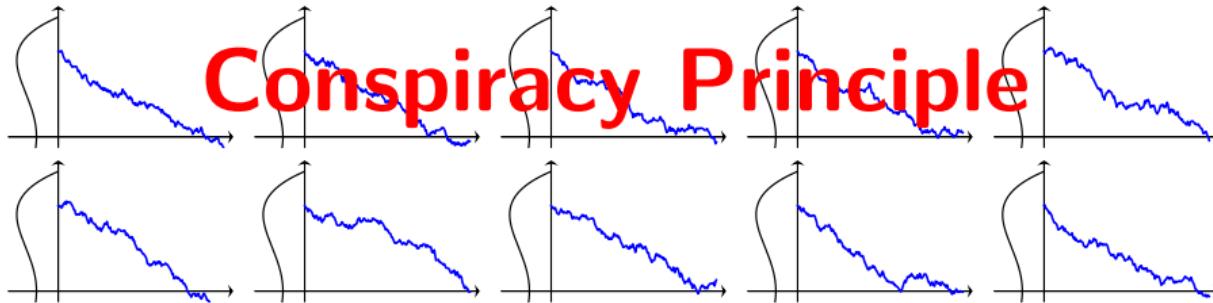


**heavy-tailed** noises with  $\eta = 1/200$

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

Trajectory of SGD  $X^\eta$  conditional on exit:

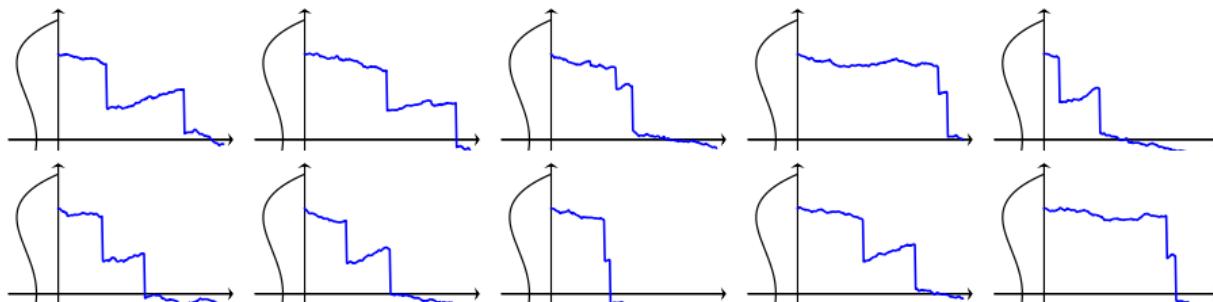
**light-tailed** noises with  $\eta = 1/200$



## Conspiracy Principle

Trajectory of SGD  $X^\eta$  conditional on exit:

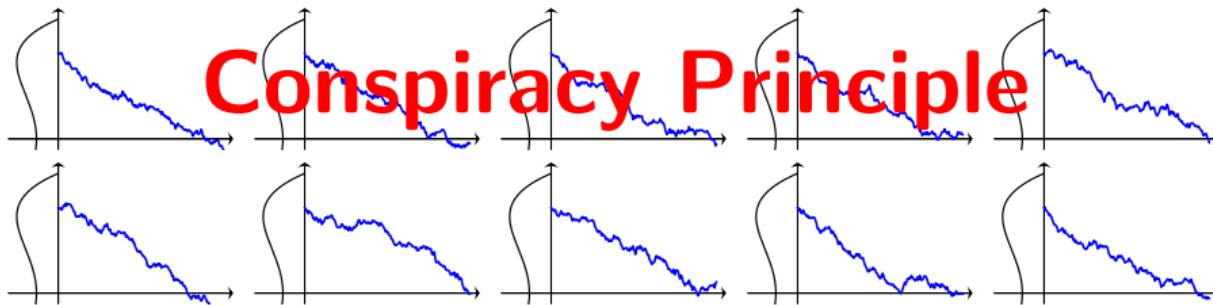
**heavy-tailed** noises with  $\eta = 1/200$



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

Trajectory of SGD  $X^\eta$  conditional on exit:

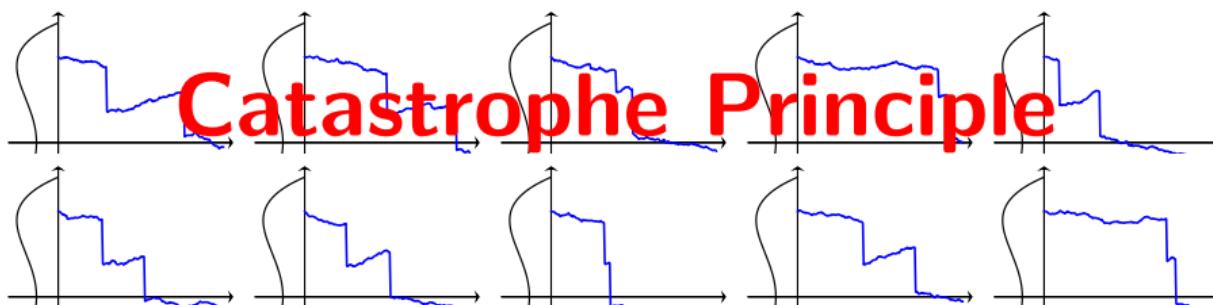
**light-tailed** noises with  $\eta = 1/200$



## Conspiracy Principle

Trajectory of SGD  $X^\eta$  conditional on exit:

**heavy-tailed** noises with  $\eta = 1/200$



## Catastrophe Principle

## **Metastability of SGD**

# Heavy-Tailed Large Deviations for SGD

Theorem (Wang, R., 2023+)

For “general”  $B \subseteq \mathbb{D}[0, T]$

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

- $\mathcal{J}(B)$ : 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”  $B \subseteq \mathbb{D}[0, T]$

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

- $\mathcal{J}(B)$ : 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”  $B \subseteq \mathbb{D}[0, T]$

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

- $\mathcal{J}(B)$ : 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”  $B \subseteq \mathbb{D}[0, T]$

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

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

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

## M-Convergence

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

Definition (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  $\eta \rightarrow 0$  if

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

## M-Convergence

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

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

Definition (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  $\eta \rightarrow 0$  if

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

## M-Convergence

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

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

Definition (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  $\eta \rightarrow 0$  if

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



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

## M-Convergence

$\swarrow \epsilon\text{-fattening of } \mathbb{C}$

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

Definition (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  $\eta \rightarrow 0$  if

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

↑  
“bounded continuous functions supported on  $\mathbb{S} \setminus \mathbb{C}$ ”

Definition (Uniform 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  $\eta \rightarrow 0$  if

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

## 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.

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

# 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.

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

Then the following statements are equivalent:

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

# 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.

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

Then the following statements are equivalent:

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

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  $\delta_B : (0, \infty) \times (0, \infty) \rightarrow (0, \infty)$ ,  $\varepsilon_B > 0$ , and  $T_B > 0$  s.t.

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

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

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

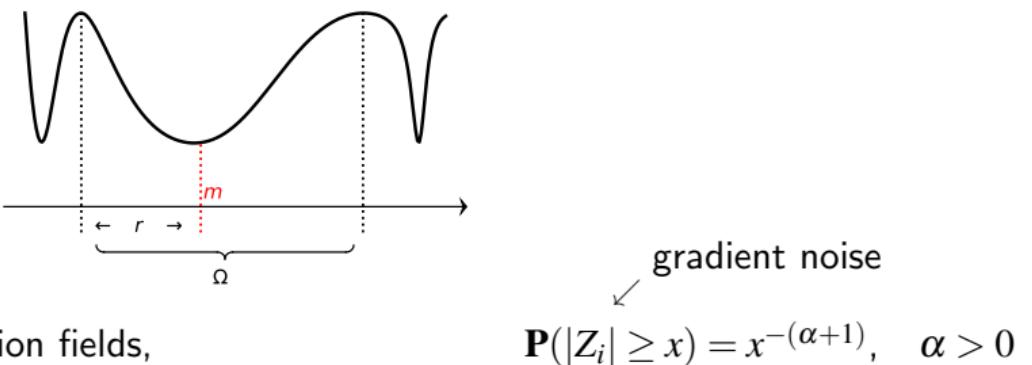
# Exit Time and Location under the Presence of Asymptotic Atom

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

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

# First Exit Time Analysis for SGD

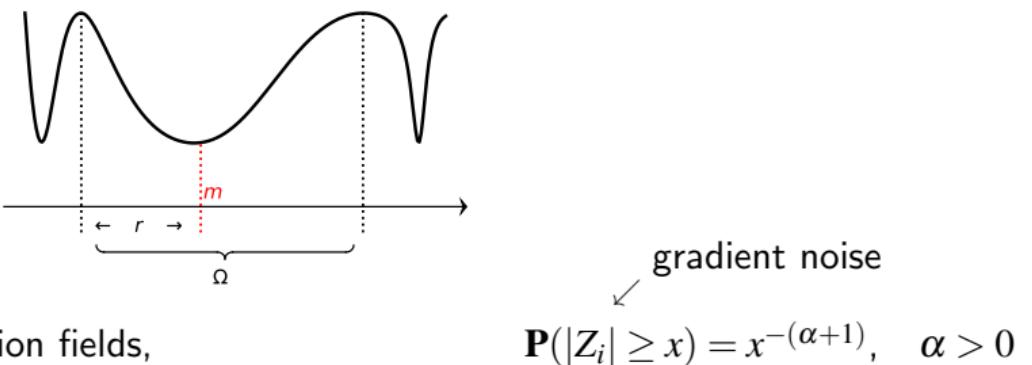


Theorem (Wang, Oh, R., 2022)

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

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

# First Exit Time Analysis for SGD



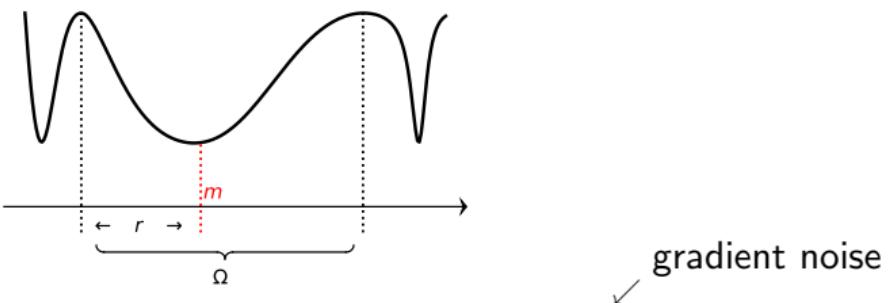
Theorem (Wang, Oh, R., 2022)

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



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

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

Theorem (Wang, Oh, R., 2022)

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

First Exit Time

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

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

# Eliminating Sharp Local Minima with Truncated Heavy-Tails

$l^*$ : “width” of the widest attraction fields,

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

gradient noise

Theorem (Wang, Oh, R., 2022)

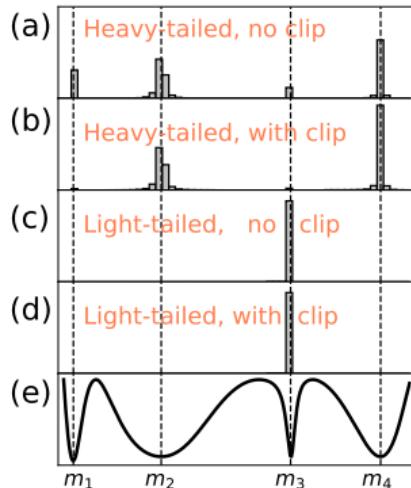
Under certain structural conditions, for any  $t > 0$  and  $\beta > 1 + \alpha \cdot l^*$ ,

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

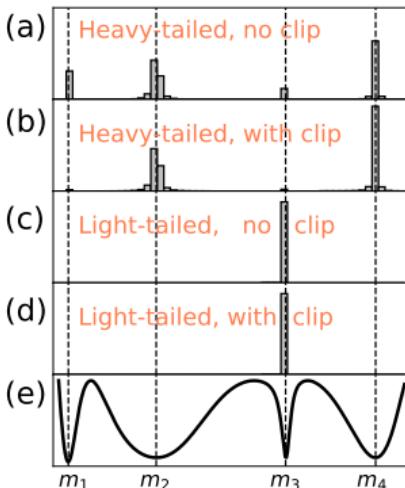
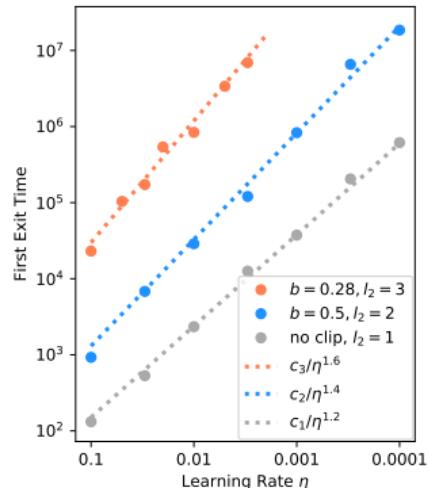
In fact,  $W_{\lfloor t/\eta^{1+\alpha \cdot l^*} \rfloor}^\eta$  converges to a Markov jump processes

whose state space consists of wide local minima only.

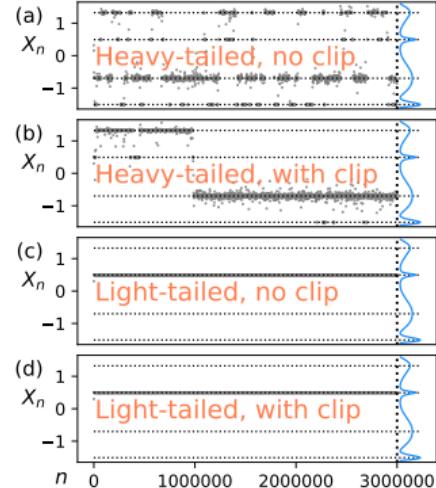
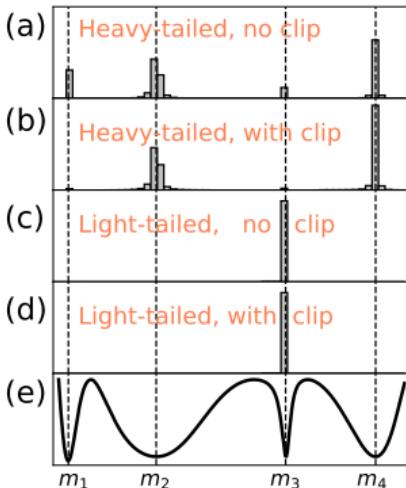
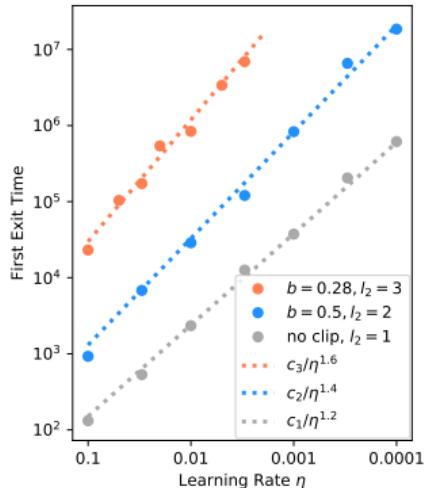
# Eliminating Sharp Local Minima with Truncated Heavy-Tails



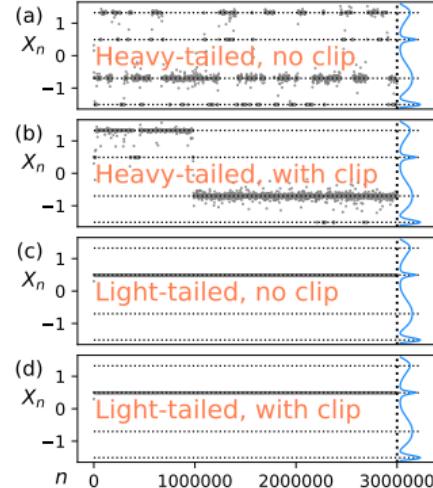
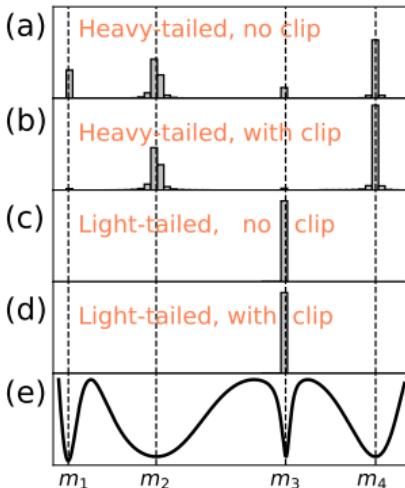
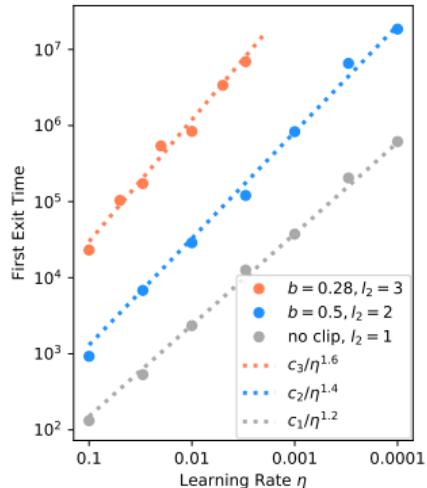
# Eliminating Sharp Local Minima with Truncated Heavy-Tails



# Eliminating Sharp Local Minima with Truncated Heavy-Tails

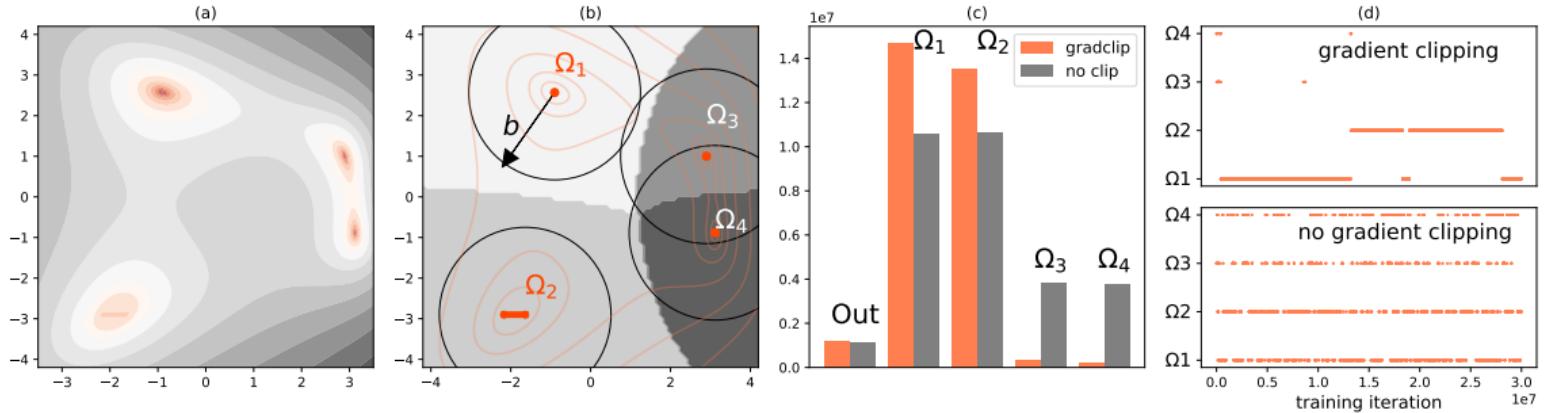


# Eliminating Sharp Local Minima with Truncated Heavy-Tails

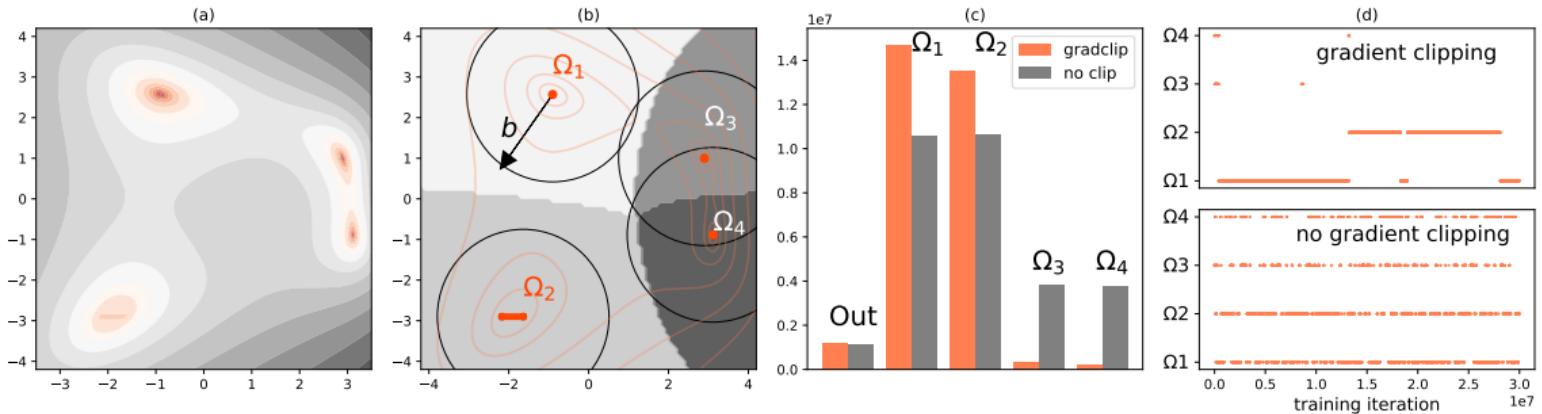


Consistent with what theory predicts!

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



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



Again, consistent with what theory predicts!

## **New Training Strategy: Tail-INflation-Truncation**

## Tail-Inflation-Truncation Scheme

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

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

# Tail-Inflation-Truncation Scheme

Stochastic Gradient

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

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

# Tail-Inflation-Truncation Scheme

Stochastic Gradient

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

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

## Tail-Inflation-Truncation Scheme

Stochastic Gradient

Pareto RV

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

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

## Tail-Inflation-Truncation Scheme

Heavy-Tailed Stochastic Gradient

Pareto RV

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

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

# 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%	<b>70.06%</b>
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	<b>88.42%</b>	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	<b>75.87%</b>
Expected Sharpness	LB	SB	SB+Clip	SB+Noise	Our 1	Our 2
FashionMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	<b>0.002</b>
SVHN, VGG11	0.694	0.037	0.041	0.012	<b>0.002</b>	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	<b>0.024</b>	0.037

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

We will know soon:

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

# Summary

## Summary

- Catastrophe Principle for various heavy-tailed stochastic processes

## Summary

- Catastrophe Principle for various heavy-tailed stochastic processes
- Solutions to open problems in queueing theory and rare-event simulation

## Summary

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

## Summary

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

## Summary

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