Heavy-Tailed Large Deviations and How to Eliminate Sharp Minima from SGD

Chang-Han Rhee

Northwestern University

CSL SINE Seminar, UIUC, March 20, 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

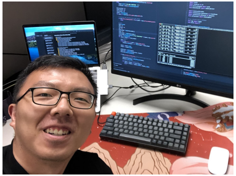
Jose Blanchet
Stanford

Bohan Chen
Munich Re

Sewoong Oh
U. of Washington

Zhe Su
Northwestern

Bert Zwart
CWI



Xingyu Wang
Northwestern

Generalization Mystery of Deep Learning

Modern “Artificial Intelligence”

“AI is in a ‘golden age’ and solving problems that were once in the realm of sci-fi.”

- Jeff Bezos

Empirical Success of Deep Neural Networks (DNNs)

"Deep Learning is eating the world."

- Jorge Nocedal

Empirical Success of Deep Neural Networks (DNNs)

- Despite massive over-parametrization, DNNs generalize well.

Empirical Success of Deep Neural Networks (DNNs)

- Despite massive over-parametrization, DNNs generalize well.

Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

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

Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

- Despite massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Stochastic Gradient Descent (SGD) is the method of choice.

Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

- Despite massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- Stochastic Gradient Descent (SGD) is the method of choice.

A Central Mystery of Deep Learning

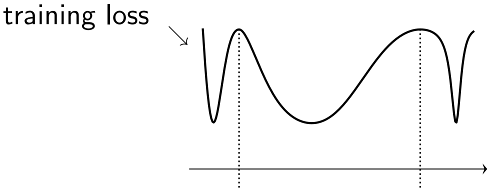
Heavy Tails may have something to do with the Mystery

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.

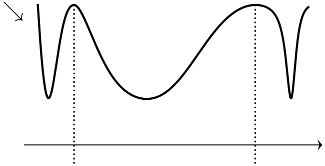


Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

tends to generalize well

training loss

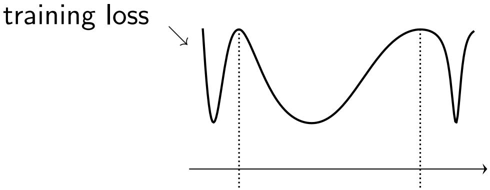


Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

tends to generalize well

- But how?



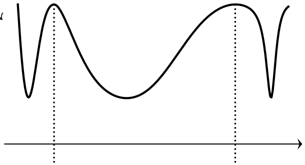
Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

  tends to generalize well

- But how?

training loss



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

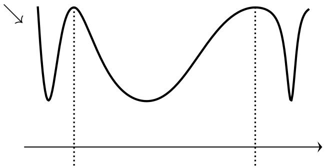
Heavy Tails may have something to do with the Mystery

- Popular explanation: SGD somehow finds flat local minima.

tends to generalize well

- But how? Previous theoretical attempts failed to explain how.

training loss



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 <u>flat local minima</u>.

  tends to generalize well

- But how? Previous theoretical attempts failed to explain how.

- Recent evidence/argument of 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.

- Recent evidence/argument of 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 <u>flat local minima</u>.

  tends to generalize well

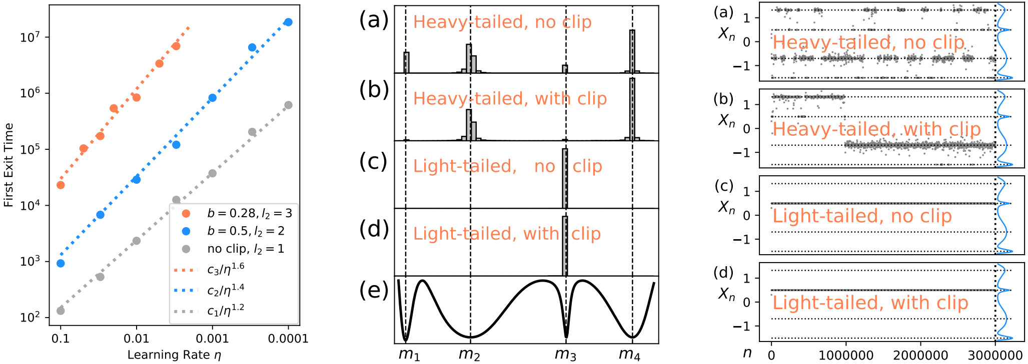
- But how? Previous theoretical attempts failed to explain how.

- Recent evidence/argument of heavy-tails in training DNNs with SGDs!

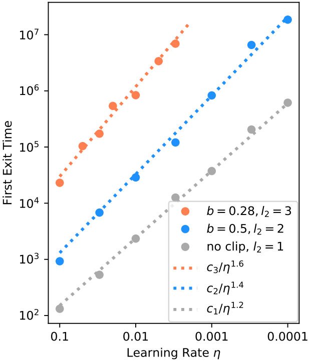
  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), etc

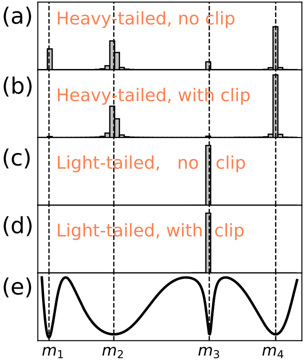
However, when SGD is heavy-tailed, one often truncates gradients.

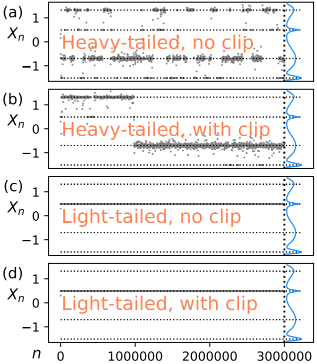
Entirely Different Global Dynamics Depending on Tail Behaviors



Entirely Different Global Dynamics Depending on Tail Behaviors

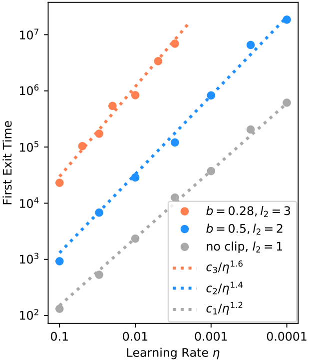


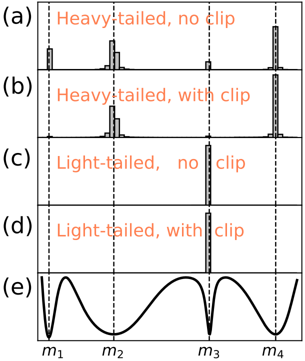


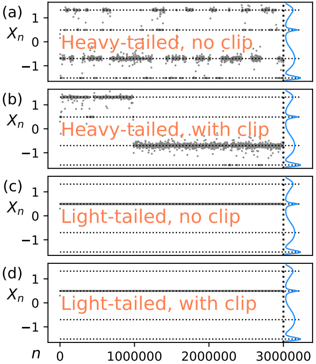


Explanation?

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

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)

Structural difference in the way systemwide rare events arise.

Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

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

Systemwide rare events

arise because

EVERYTHING goes wrong.

(Conspiracy Principle)

Heavy-Tailed Distributions

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

Systemwide rare events

arise because of

A FEW Catastrophes.

(Catastrophe Principle)

Structural difference in the way systemwide rare events arise.

Insurance Example: Capital Reserve

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

Insurance Example: Capital Reserve

Initial Capital

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

Insurance Example: Capital Reserve

Initial Capital

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

Premium

Insurance Example: Capital Reserve

Initial Capital

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

Premium

i.i.d. Claim Size

Insurance Example: Capital Reserve

Poisson Arrival

Initial Capital

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

Premium

i.i.d. Claim Size

Insurance Example: Capital Reserve

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

Insurance Example: Capital Reserve

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

Insurance Example: Capital Reserve

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

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

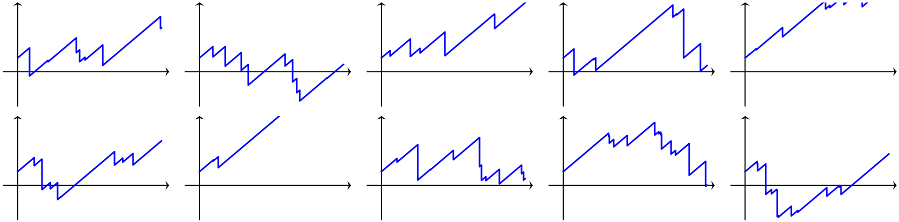
Typical Scenario

Typical Scenario

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

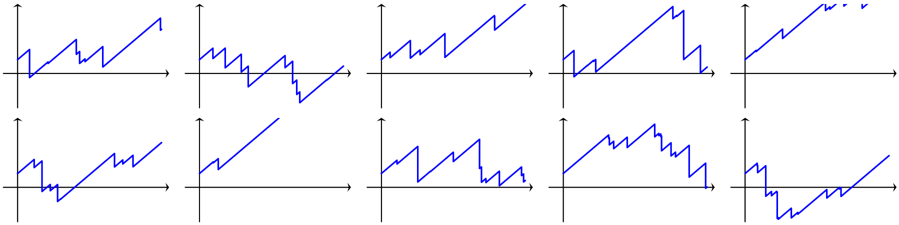
Typical Scenario

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



Typical Scenario

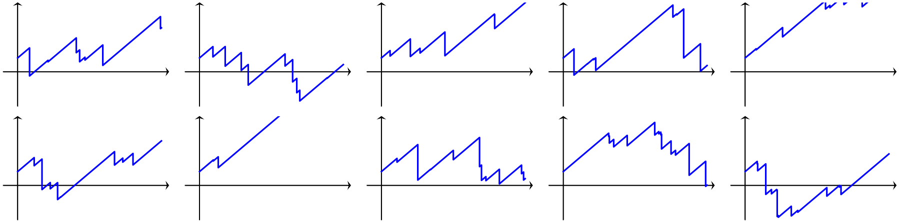
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**



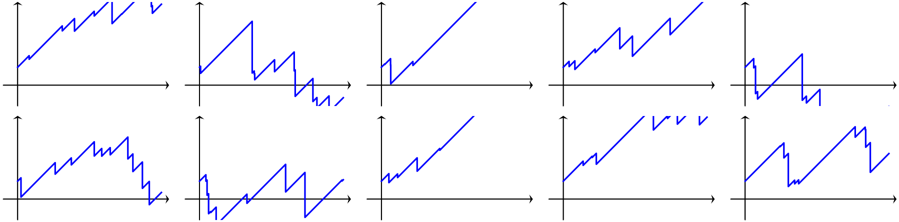
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

Typical Scenario

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

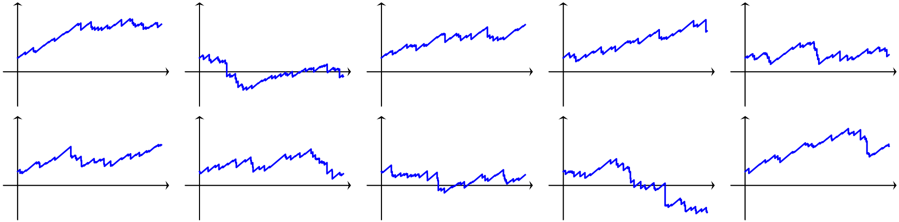


Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

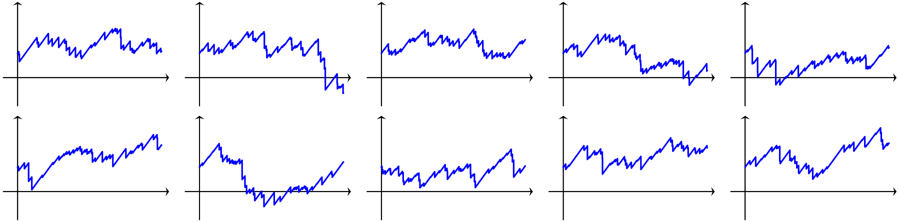


Typical Scenario

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

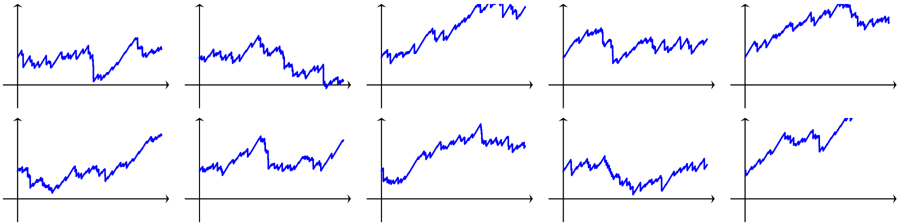


Sample paths of $\bar{Y}_n$: n=50 & claim sizes are **heavy-tailed**

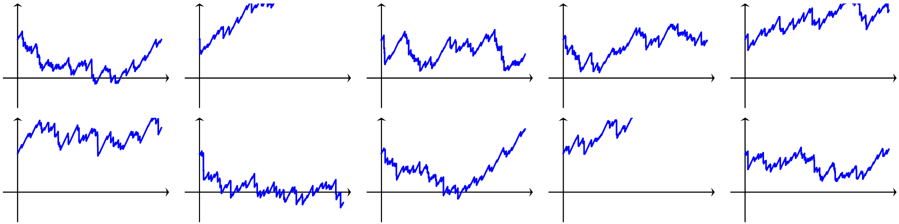


Typical Scenario

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

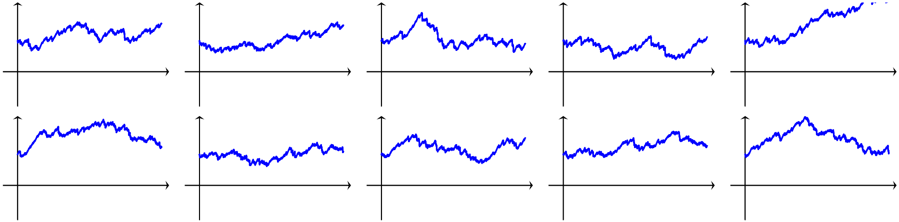


Sample paths of $\bar{Y}_n$: n=100 & claim sizes are **heavy-tailed**

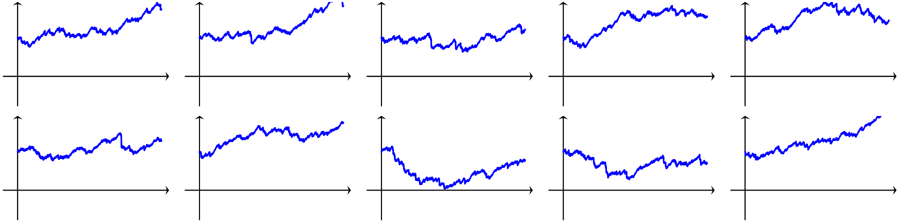


Typical Scenario

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

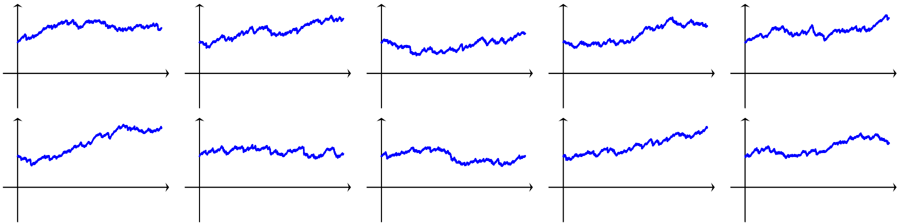


Sample paths of $\bar{Y}_n$: n=500 & claim sizes are **heavy-tailed**

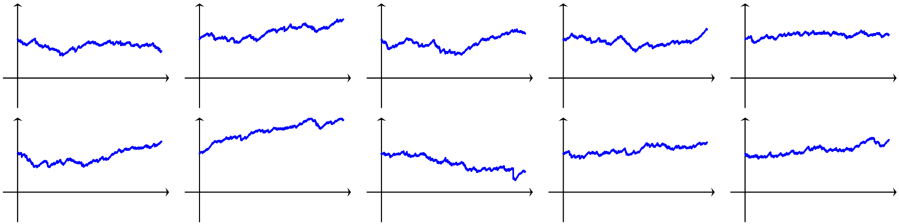


Typical Scenario

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

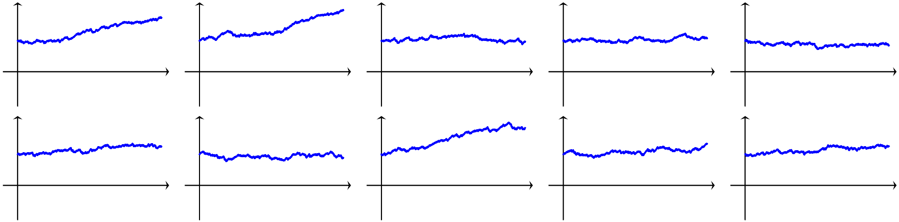


Sample paths of $\bar{Y}_n$: n=1000 & claim sizes are **heavy-tailed**

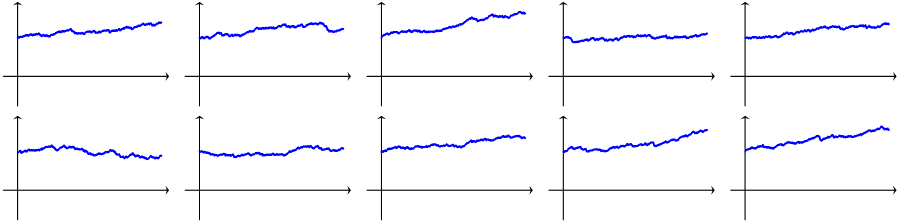


Typical Scenario

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

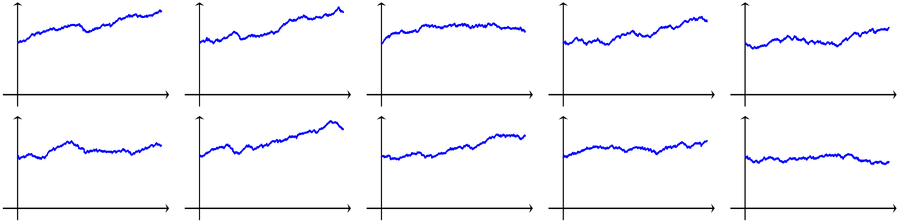


Sample paths of $\bar{Y}_n$: n=2500 & claim sizes are **heavy-tailed**

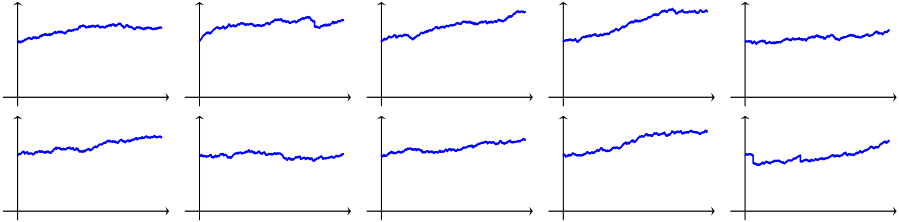


Typical Scenario

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



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

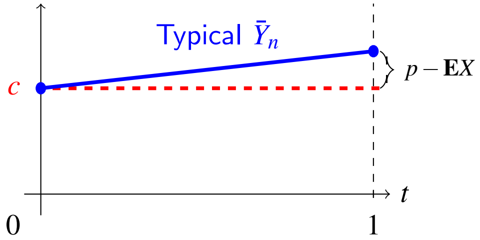


Typical Scenario

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

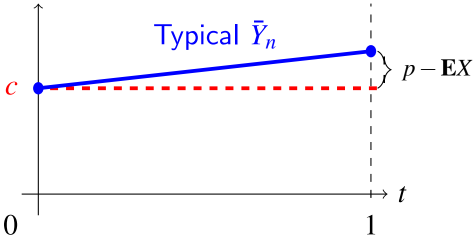
Typical Scenario

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



Typical Scenario

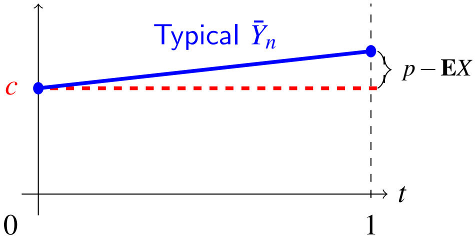
That is, $\quad \bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t \quad$ for large n.



Typically, your business will flourish

Typical Scenario

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

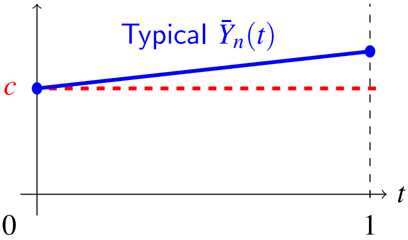


Typically, your business will flourish

regardless of the tail distributions.

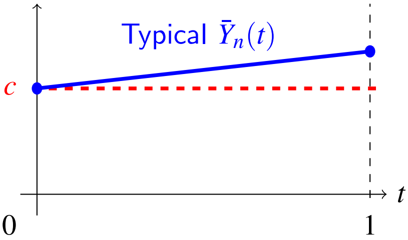
What about atypical cases?

A Rare Event: Bankruptcy



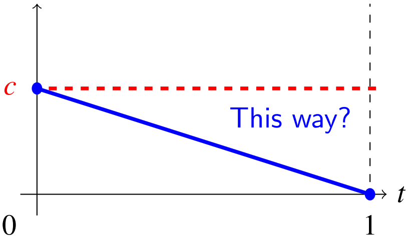
A Rare Event: Bankruptcy

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



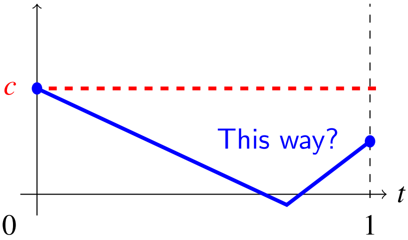
A Rare Event: Bankruptcy

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



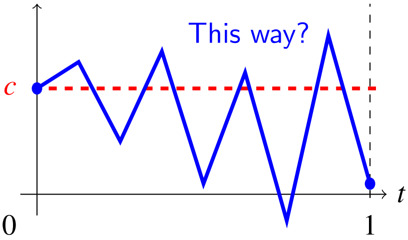
A Rare Event: Bankruptcy

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



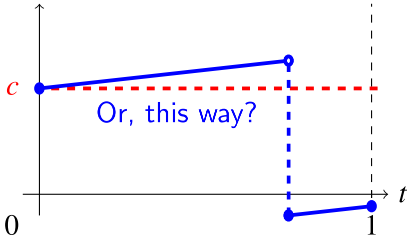
A Rare Event: Bankruptcy

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



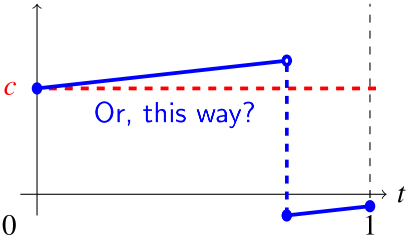
A Rare Event: Bankruptcy

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



A Rare Event: Bankruptcy

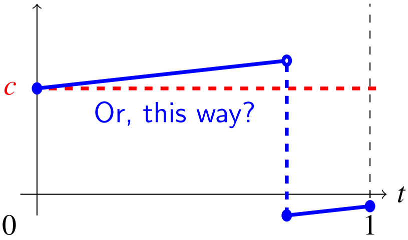
Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



Are we going to see clear patterns?

A Rare Event: Bankruptcy

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

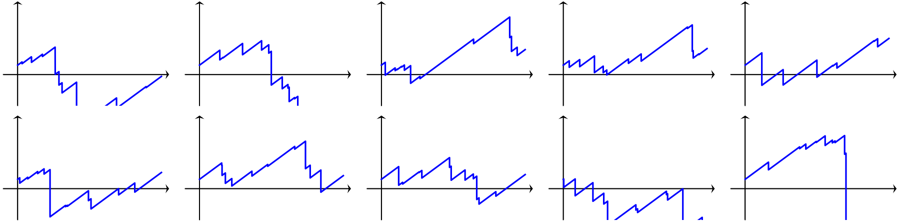


Are we going to see clear patterns?

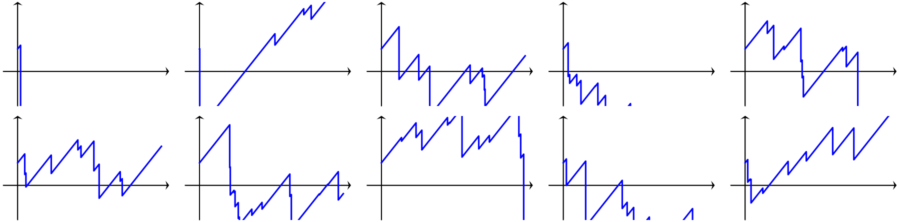
Do they depend on the tail distributions?

A Rare Event: Bankruptcy

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



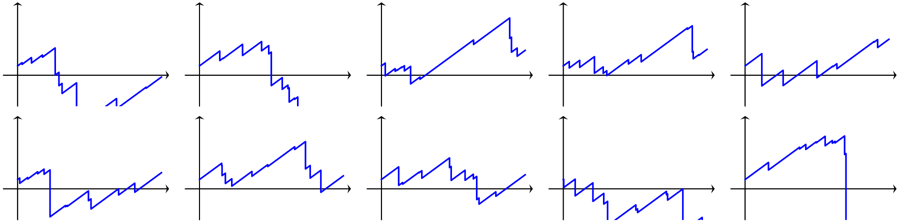
Sample paths of $\bar{Y}_{10}$ conditional on B for **heavy-tailed** claims:



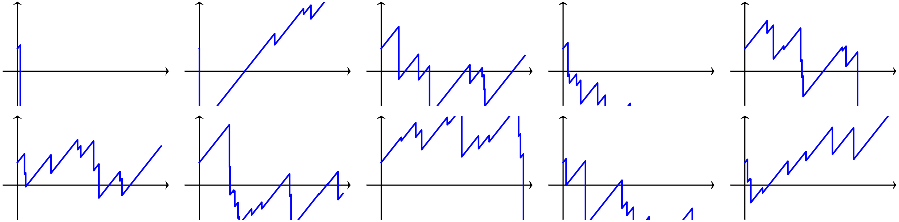
A Rare Event: Bankruptcy

Bankruptcy

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



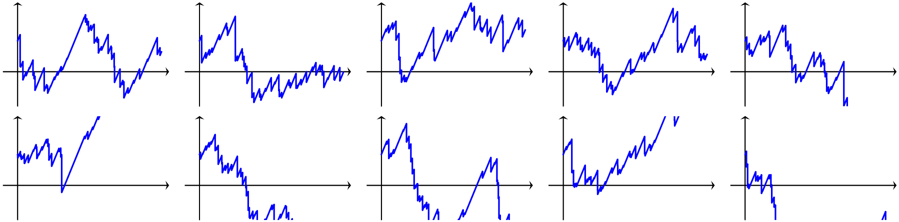
Sample paths of $\bar{Y}_{10}$ conditional on B for **heavy-tailed** claims:



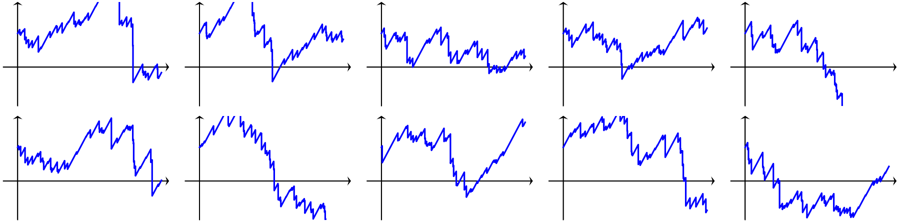
A Rare Event: Bankruptcy

Bankruptcy

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



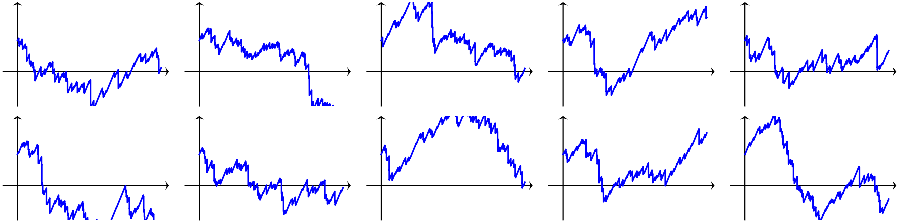
Sample paths of $\bar{Y}_{50}$ conditional on B for **heavy-tailed** claims:



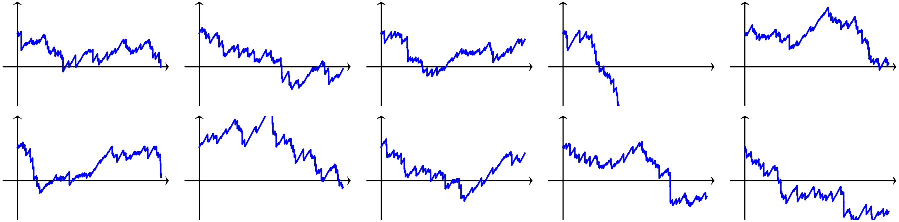
A Rare Event: Bankruptcy

Bankruptcy

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



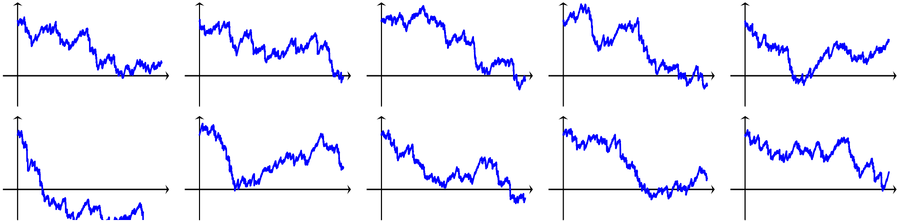
Sample paths of $\bar{Y}_{100}$ conditional on B for **heavy-tailed** claims:



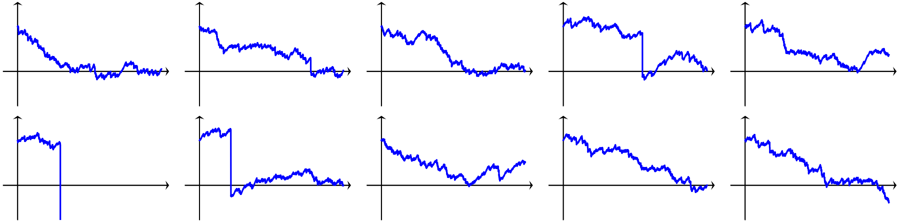
A Rare Event: Bankruptcy

Bankruptcy

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



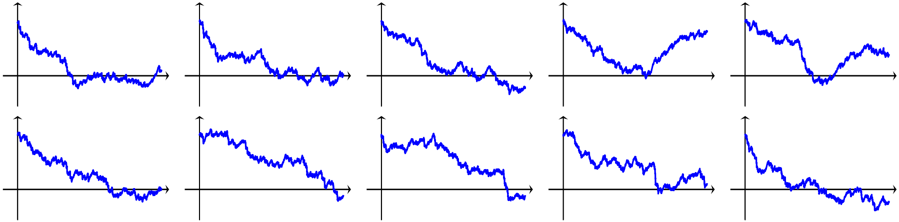
Sample paths of $\bar{Y}_{500}$ conditional on B for **heavy-tailed** claims:



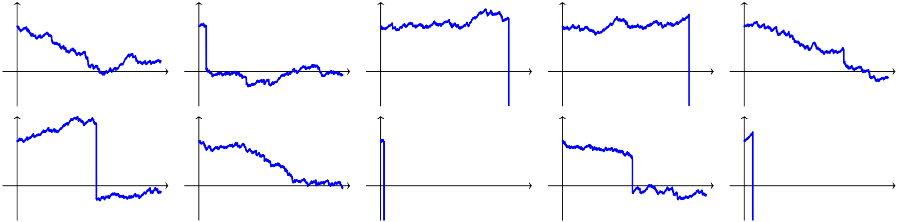
A Rare Event: Bankruptcy

Bankruptcy

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



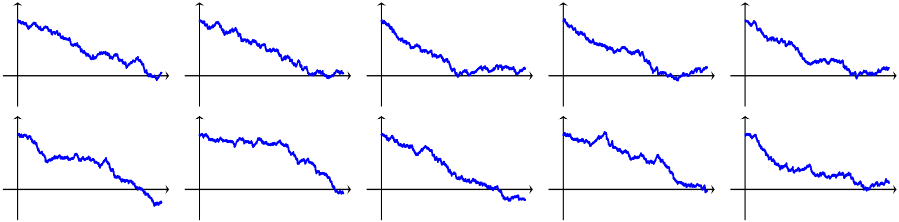
Sample paths of $\bar{Y}_{1000}$ conditional on B for **heavy-tailed** claims:



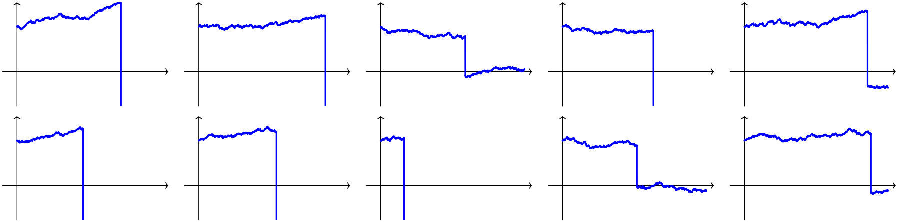
A Rare Event: Bankruptcy

Bankruptcy

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



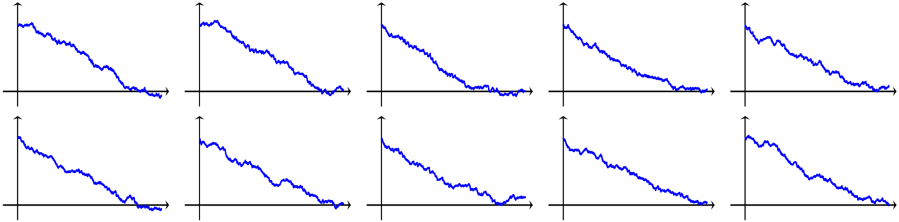
Sample paths of $\bar{Y}_{2500}$ conditional on B for **heavy-tailed** claims:



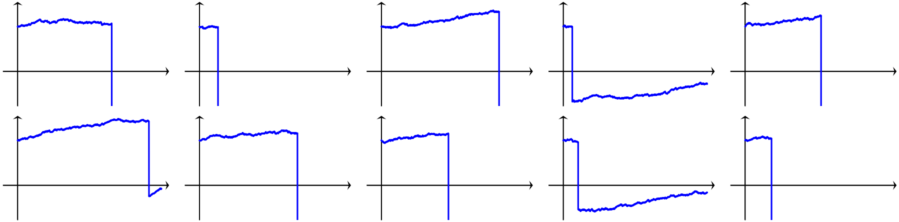
A Rare Event: Bankruptcy

Bankruptcy

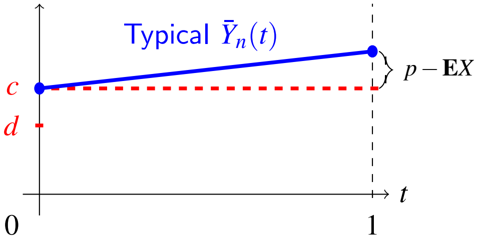
Sample paths of $\bar{Y}_{5000}$ conditional on B for **light-tailed** claims:



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

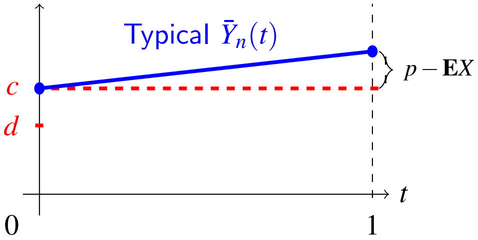


Bankruptcy Despite of Reinsurance



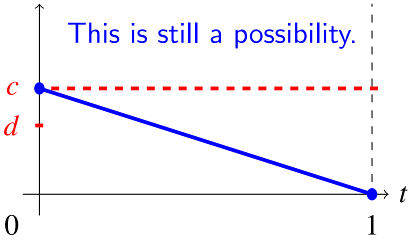
Bankruptcy Despite of Reinsurance

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



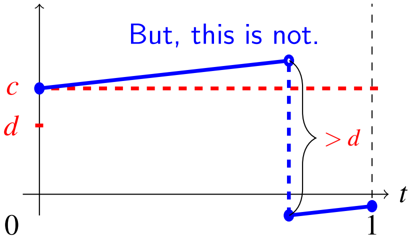
Bankruptcy Despite of Reinsurance

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



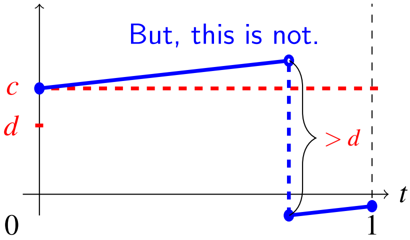
Bankruptcy Despite of Reinsurance

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



Bankruptcy Despite of Reinsurance

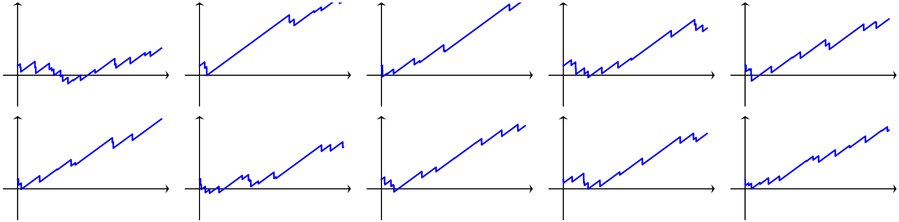
Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



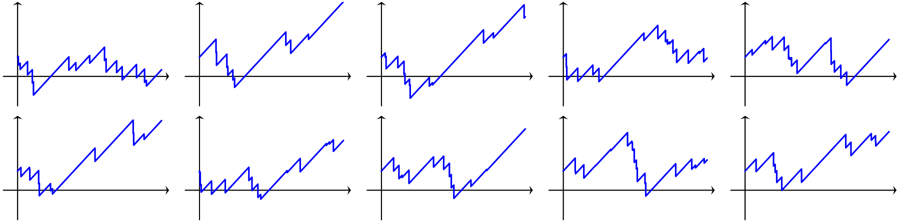
How does the pattern change in this case?

Bankruptcy Despite of Reinsurance

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

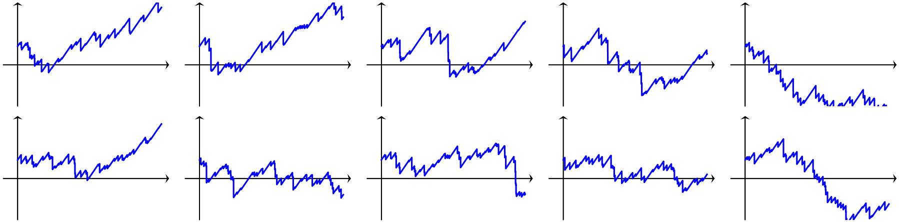


Sample paths of $\bar{Y}_{10}$ conditional on R for **heavy-tailed** claims:

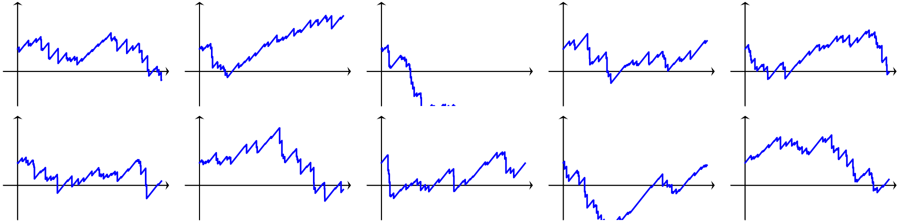


Bankruptcy Despite of Reinsurance

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

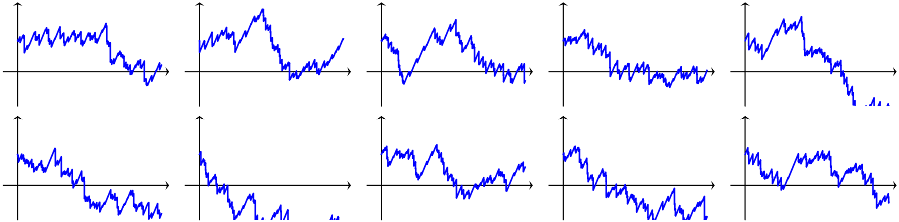


Sample paths of $\bar{Y}_{50}$ conditional on R for **heavy-tailed** claims:

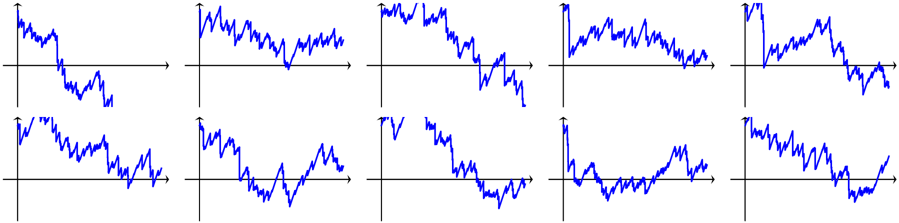


Bankruptcy Despite of Reinsurance

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

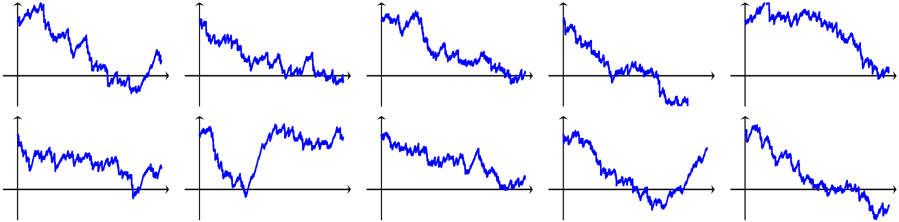


Sample paths of $\bar{Y}_{100}$ conditional on R for **heavy-tailed** claims:

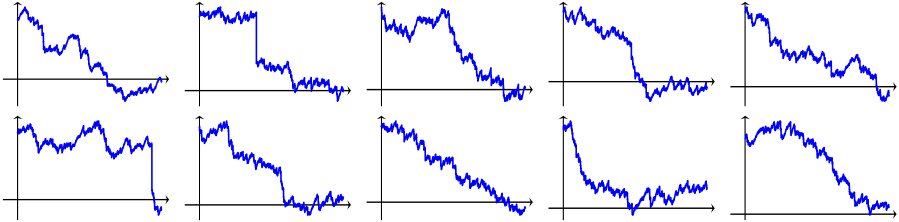


Bankruptcy Despite of Reinsurance

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

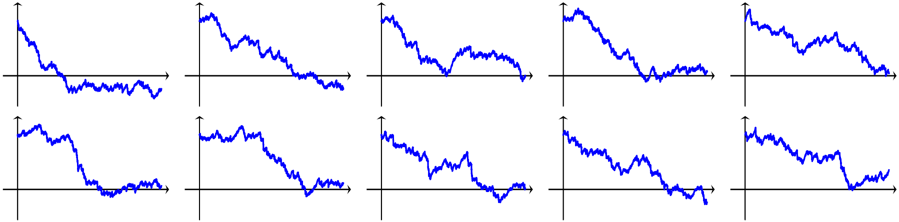


Sample paths of $\bar{Y}_{500}$ conditional on R for **heavy-tailed** claims:

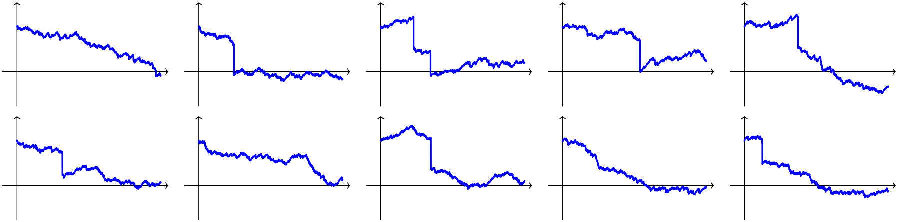


Bankruptcy Despite of Reinsurance

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

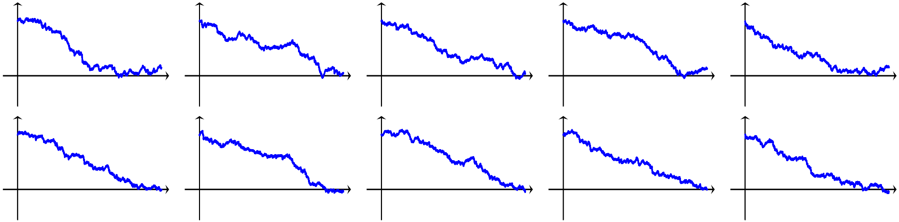


Sample paths of $\bar{Y}_{1000}$ conditional on R for **heavy-tailed** claims:

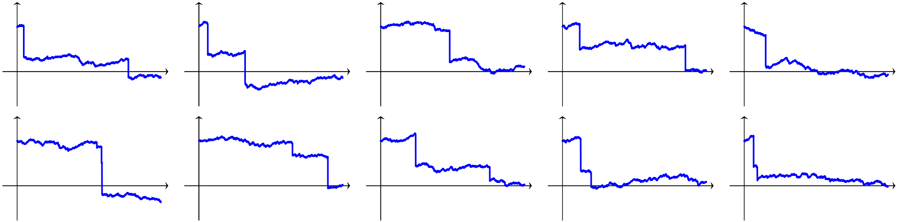


Bankruptcy Despite of Reinsurance

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

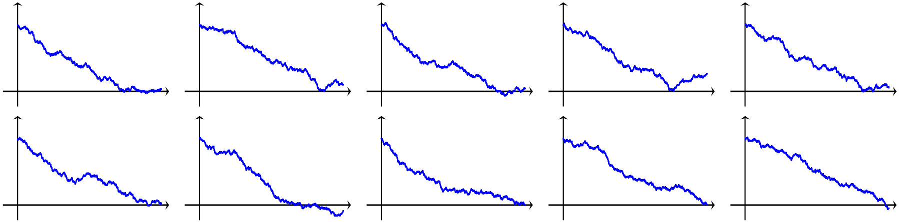


Sample paths of $\bar{Y}_{2500}$ conditional on R for **heavy-tailed** claims:

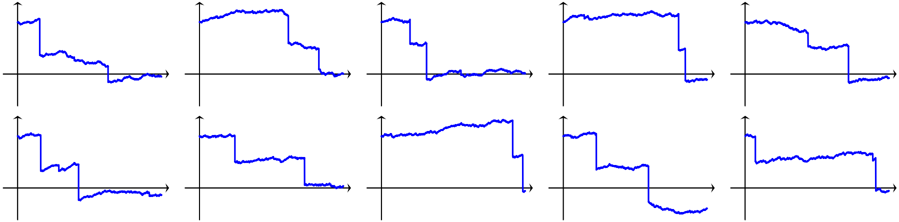


Bankruptcy Despite of Reinsurance

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

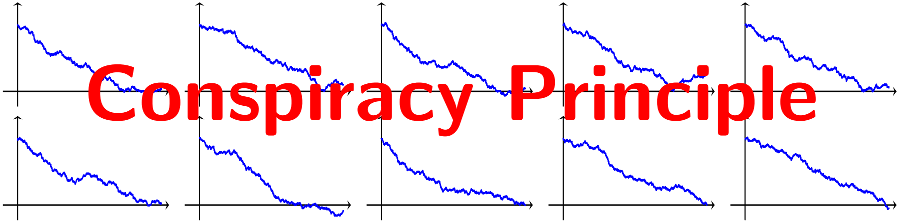


Sample paths of $\bar{Y}_{5000}$ conditional on R for **heavy-tailed** claims:

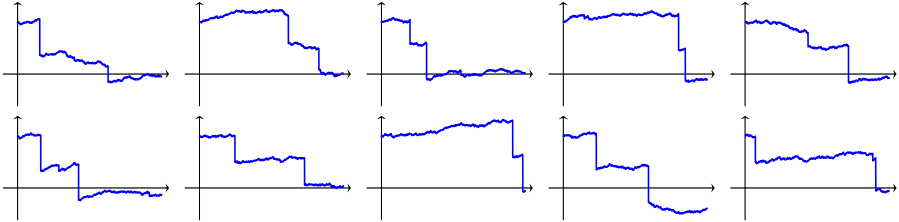


Bankruptcy Despite of Reinsurance

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

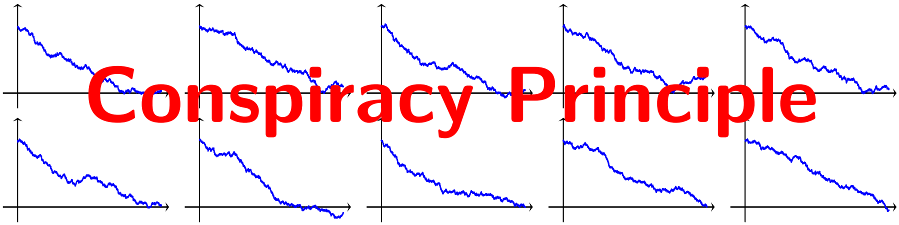


Sample paths of $\bar{Y}_{5000}$ conditional on R for **heavy-tailed** claims:

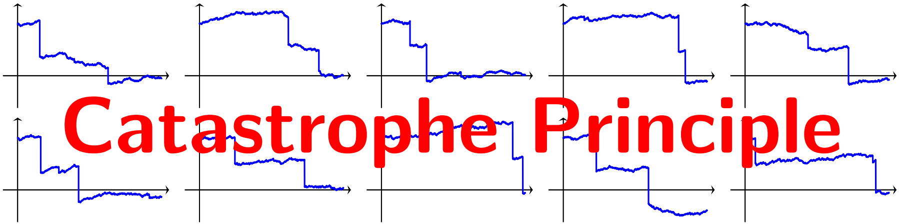


Bankruptcy Despite of Reinsurance

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



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



Heavy-Tailed Large Deviations:

Rigorous Characterization of Catastrophe Principle

Heavy-Tailed Large Deviations

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

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

Heavy-Tailed Large Deviations

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

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

Theorem (R., Blanchet, Zwart, 2019)

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

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

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

Heavy-Tailed Large Deviations

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

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

Theorem (R., Blanchet, Zwart, 2019)

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

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

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

Heavy-Tailed Large Deviations

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

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

Theorem (R., Blanchet, Zwart, 2019)

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

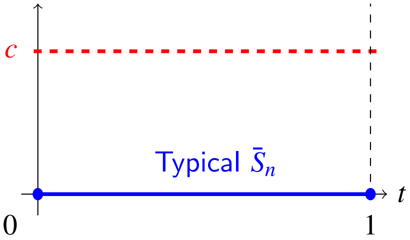
LD power index

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

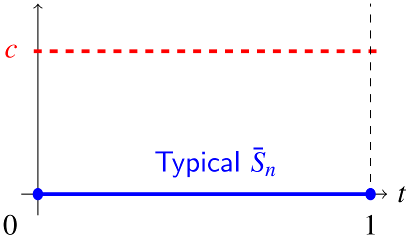
- $\mathcal{J}(A)$: *min #jumps for step functions to be inside* A

Conspiracy vs Catastrophe

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



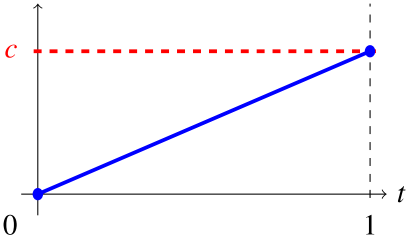
Light-Tailed Claim Size



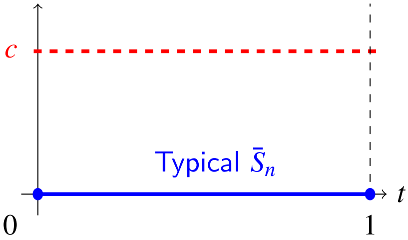
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



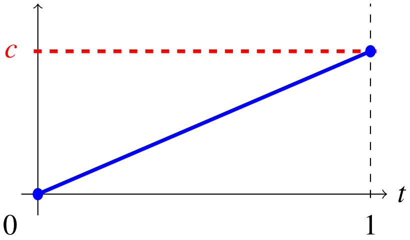
Light-Tailed Claim Size



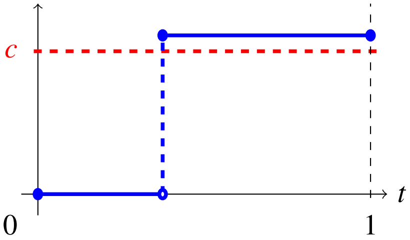
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



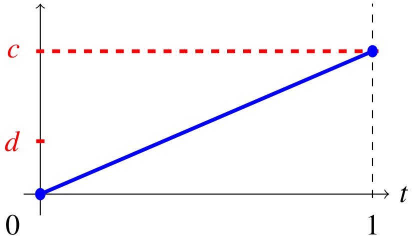
Light-Tailed Claim Size



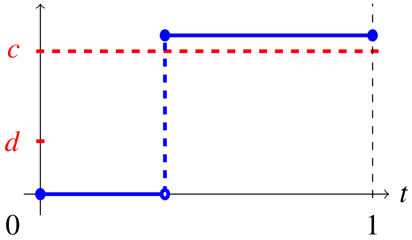
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



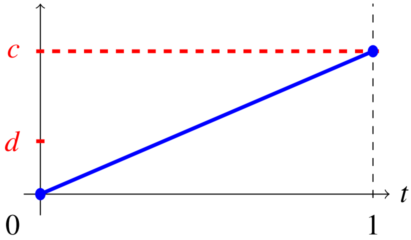
Light-Tailed Claim Size



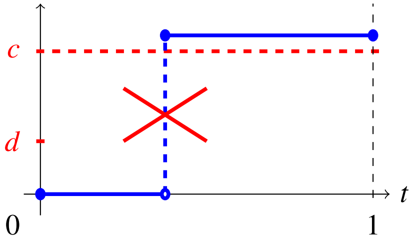
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



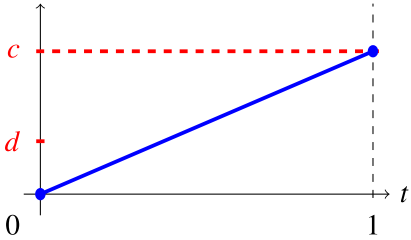
Light-Tailed Claim Size



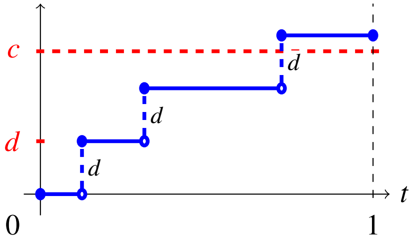
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



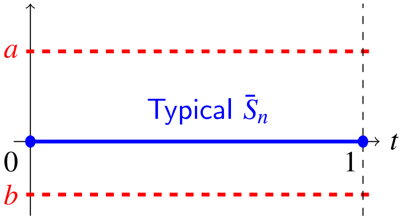
Light-Tailed Claim Size



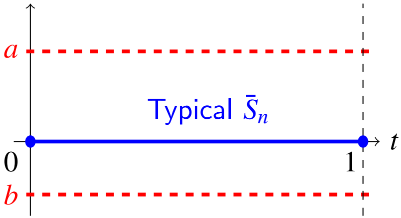
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



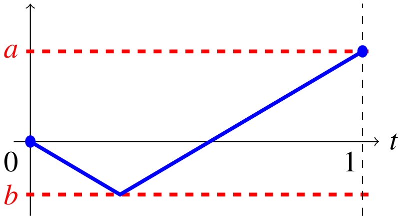
Light-Tailed Increments

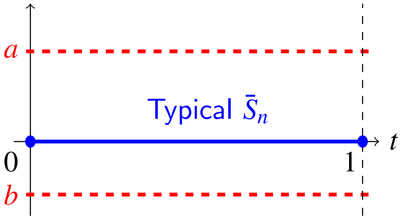


Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



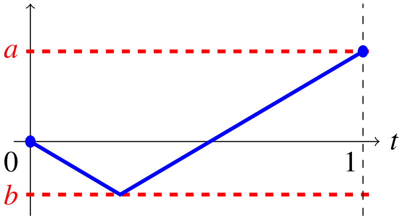


Light-Tailed Increments

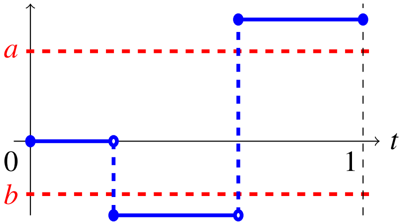
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



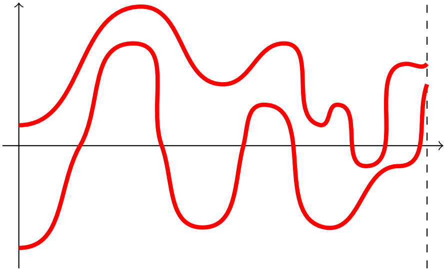
Light-Tailed Increments



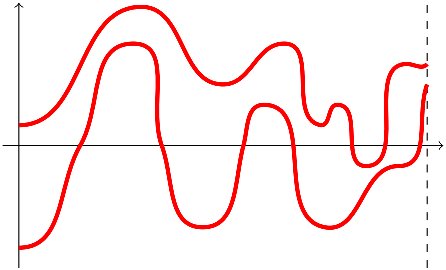
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



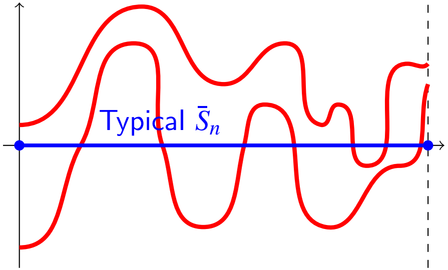
Light-Tailed Increments



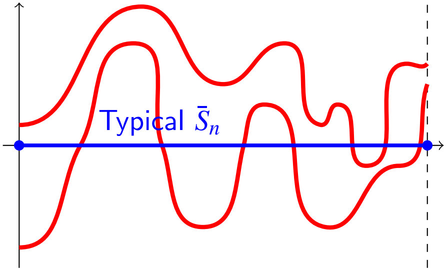
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



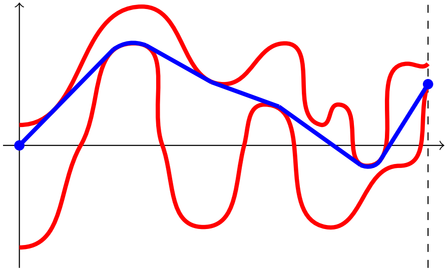
Light-Tailed Increments



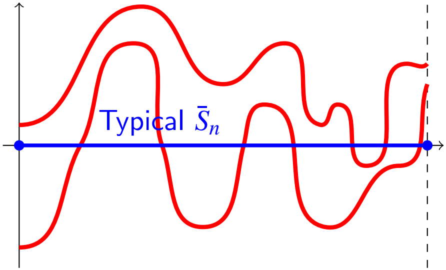
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



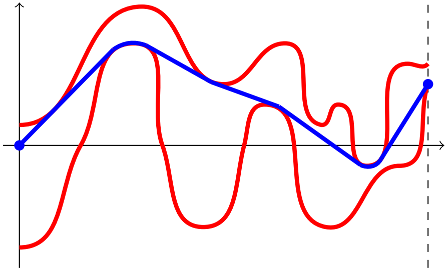
Light-Tailed Increments



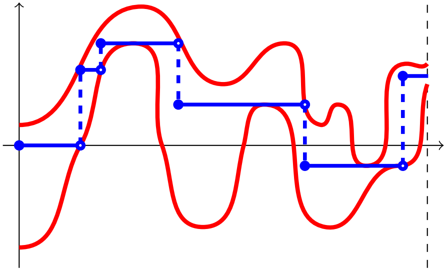
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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

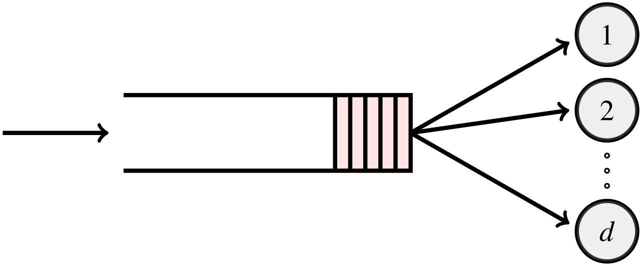


Light-Tailed Increments

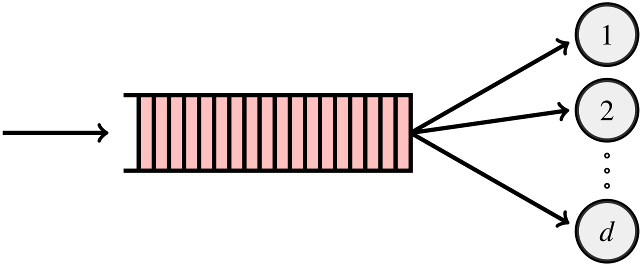


Heavy-Tailed Increments

Congestion of Multiple Server Queue:

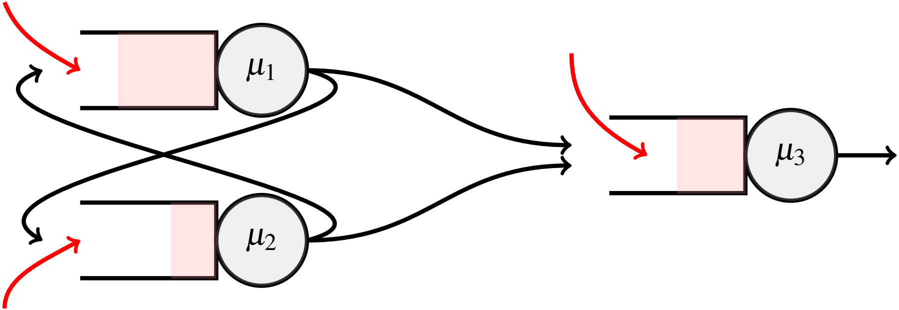


Congestion of Multiple Server Queue:



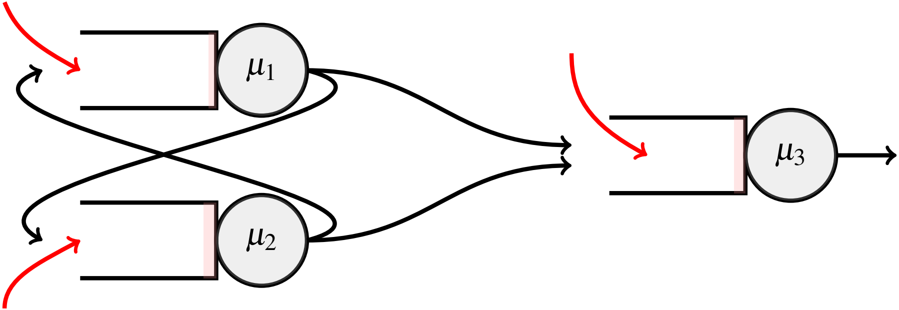
Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:

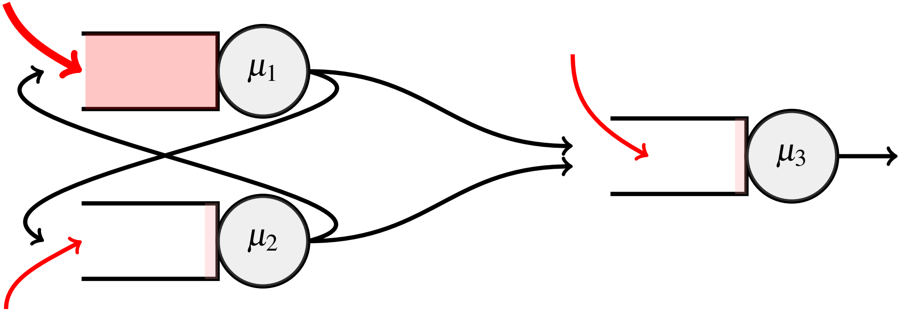


Conspiracy vs Catastrophe

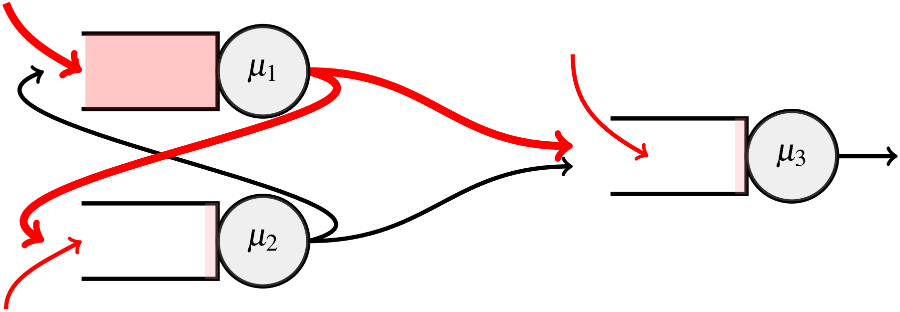
Congestion of Stochastic Fluid Networks:



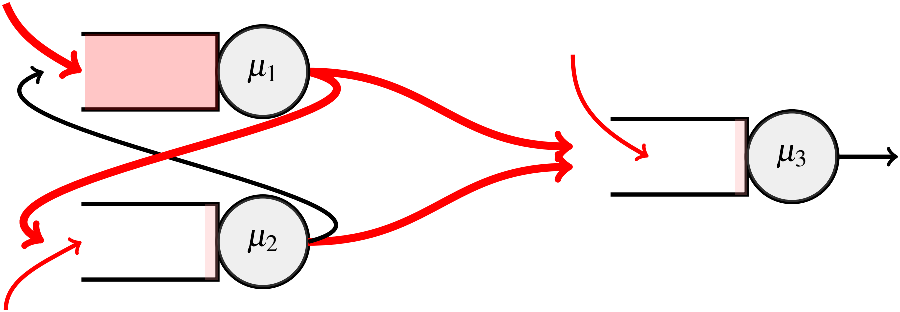
Congestion of Stochastic Fluid Networks:



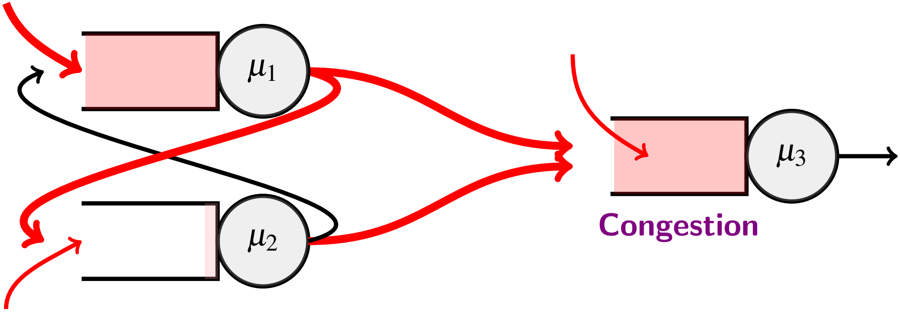
Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



The point is

Many heavy-tailed rare events can be written as

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

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

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

Catastrophe Principle Extends to General Heavy-Tailed Systems

Heavy-Tailed Large Deviations for

- Continuous-Time Processes
  R., Blanchet, Zwart (2019), Bazhba, Blanchet, R., Zwart (2020), 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 $\mathbf{P}(\text{Edge}) \overset{Suppose}{\approx} 10^{-6}$

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

Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

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

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

Challenge of Rare Event Simulation

Monte Carlo simulations as repetitive random experiments:

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

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

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

Importance Sampling

- Construct an alternative universe

Importance Sampling

- Construct an alternative universe

- Perform experiments there

Importance Sampling

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

Importance Sampling

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

- Perform experiments there

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

Importance Sampling

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

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

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

Importance Sampling

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

    as well as the likelihood ratio $\left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(1)}, \ldots, \left(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}\right)^{(m)}$
- Recover the true probability using the relationship between the two parallel universes

Importance Sampling

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Goal: Strongly Efficient IS Estimator

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

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

Goal: Strongly Efficient IS Estimator

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

IS estimator

Error ^{2}

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

Goal: Strongly Efficient IS Estimator

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

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

Goal: Strongly Efficient IS Estimator

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

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

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

Goal: Strongly Efficient IS Estimator

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

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

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

Considered Notoriously Hard for Heavy-Tailed Processes.

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

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

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

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

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

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

Error^2 $\rightarrow$

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

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

$\leftarrow$ Target Quantity^2

First Universal Rare-Event Simulation Scheme for Heavy-Tails

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

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

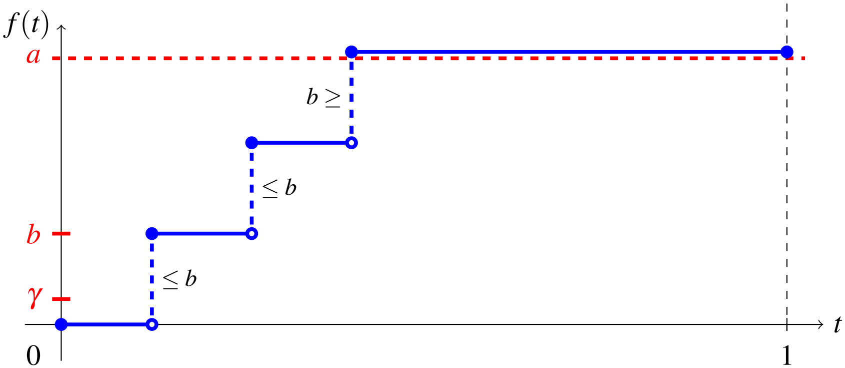
$$\text{Error}^2 \rightarrow \qquad \mathbf{E}^{\mathbf{Q}_n}\left(\mathbb{1}_{\{\bar{X}_n \in A\}}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}_n}\right)^2 \leq \frac{1}{w}\mathbf{P}(\bar{X}_n \in A \setminus B^\gamma) + \mathbf{P}(\bar{X}_n \in A)\cdot\mathbf{P}(\bar{X}_n \in B^\gamma)$$

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

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

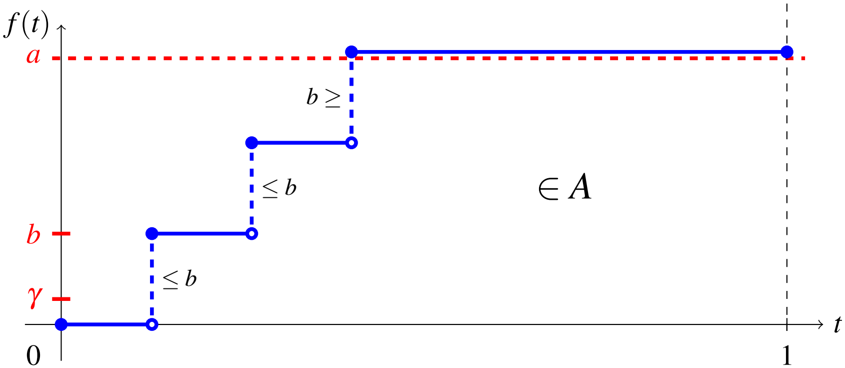
Chen, Blanchet, R., and Zwart (2019)

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



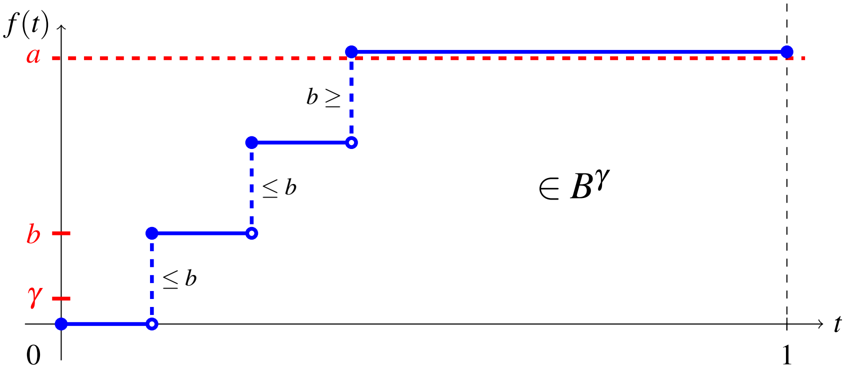
- $A \quad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^{\gamma} \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^{\gamma} = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



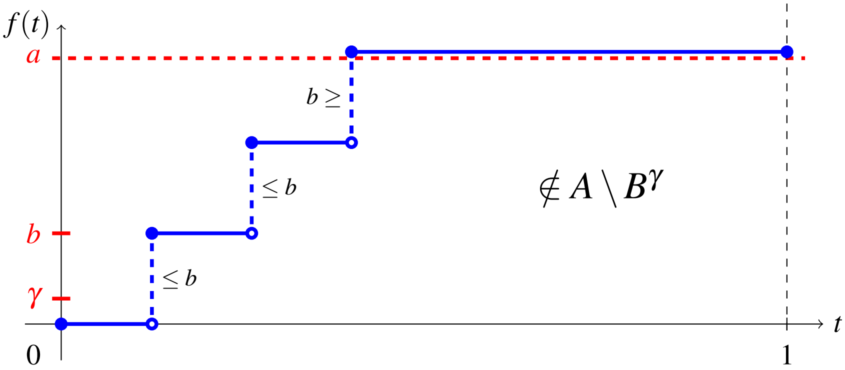
- $A \quad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



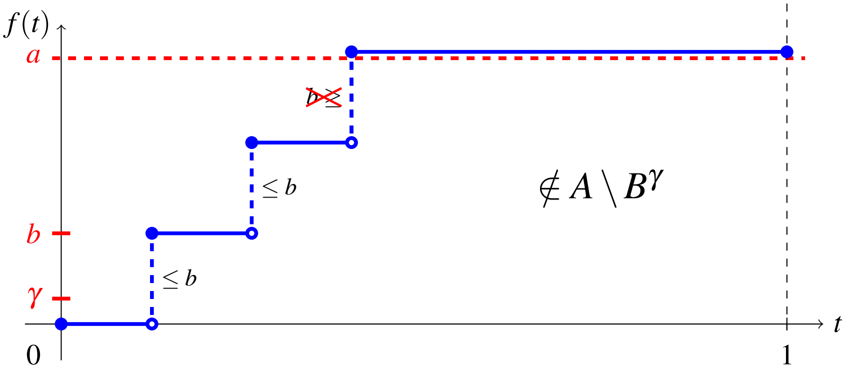
- $A \quad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



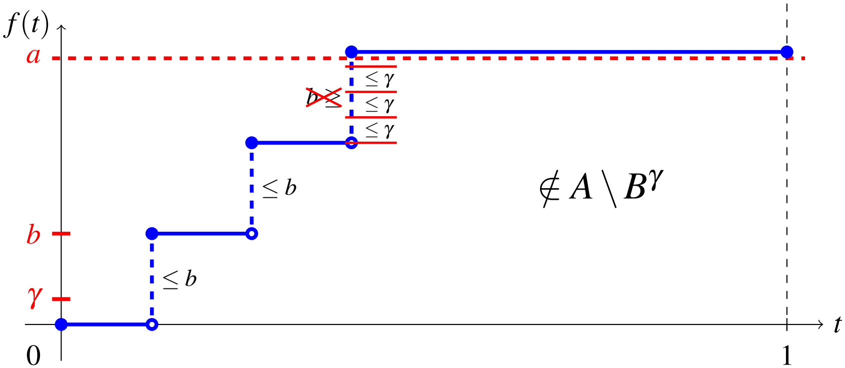
- $A \quad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



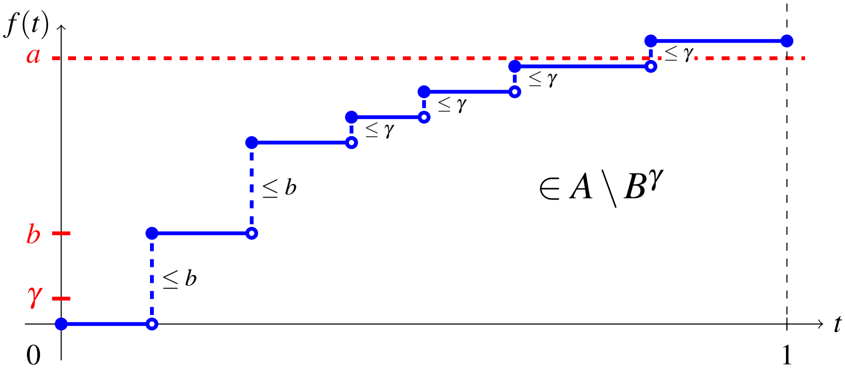
- $A \quad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \quad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



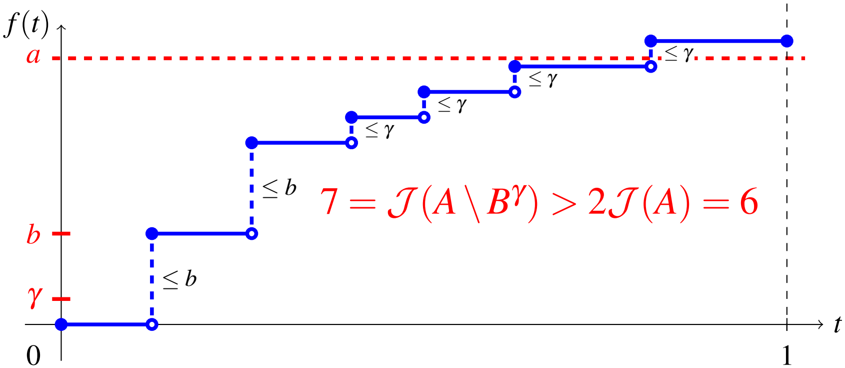
- $A \qquad = \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- $B^\gamma \qquad = \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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



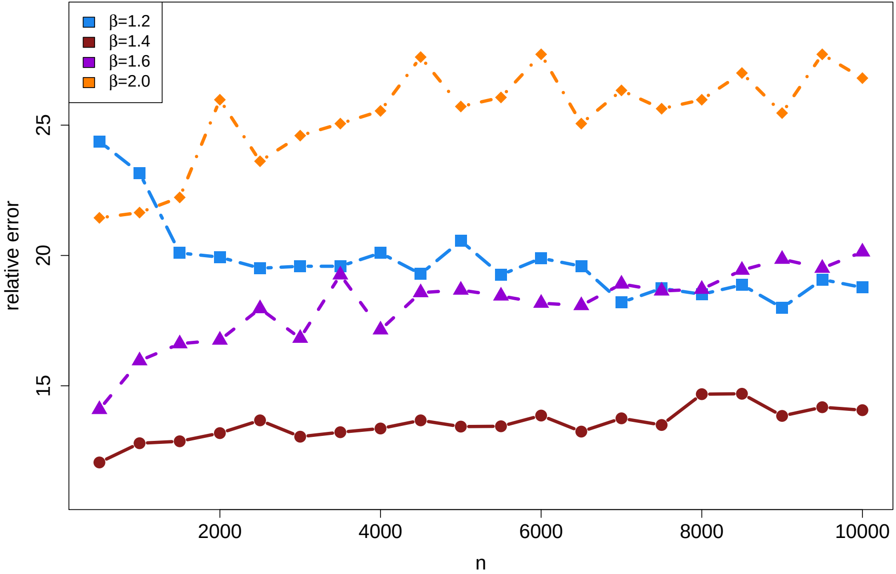
- A $= \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- B^γ $= \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

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

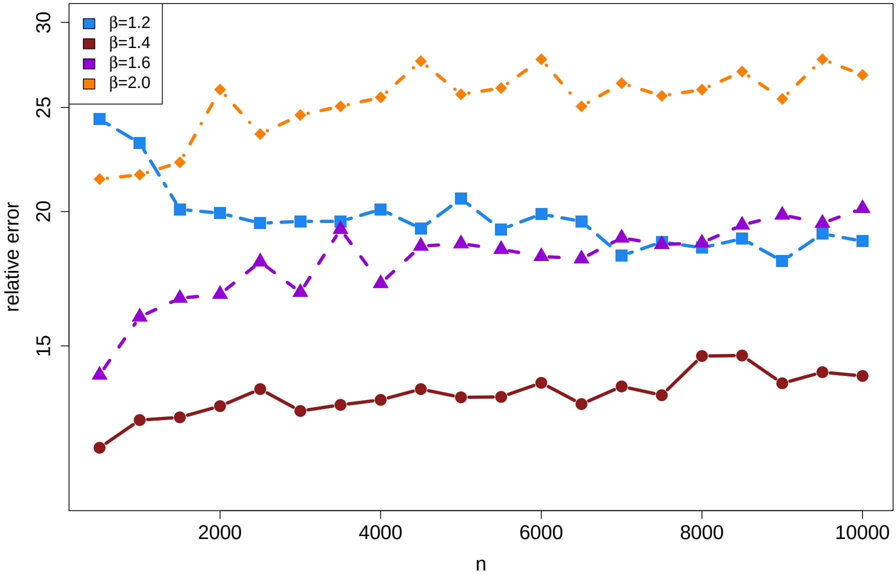


- A $= \{$paths that cross level a on $[0,1]$ & jump sizes $\leq b\}$
- B^γ $= \{$paths with at least 3 jumps of size $> \gamma\}$
- $A \setminus B^\gamma = \{$paths that cross level a on $[0,1]$
  & all jump sizes $\leq b$ & at most 2 jumps of size $> \gamma\}$

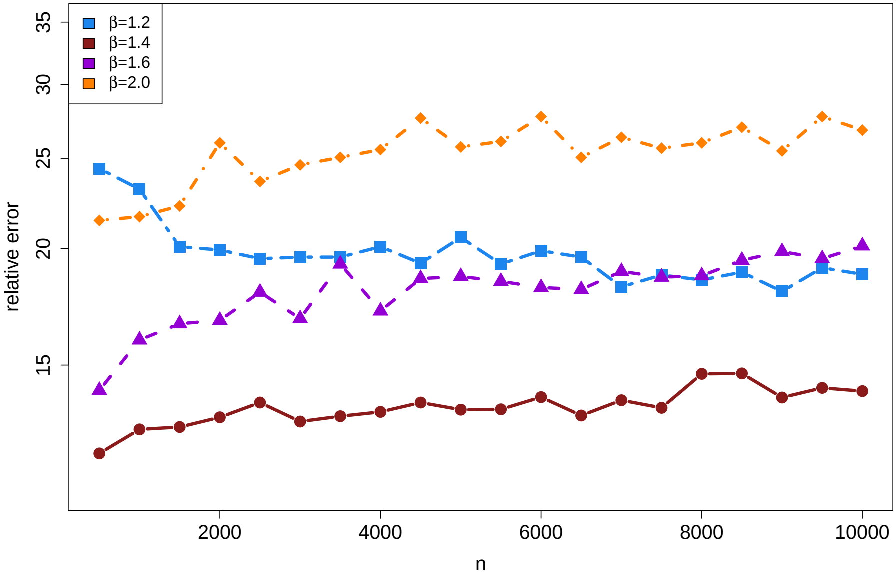
Numerical Experiments for Reinsurance Example



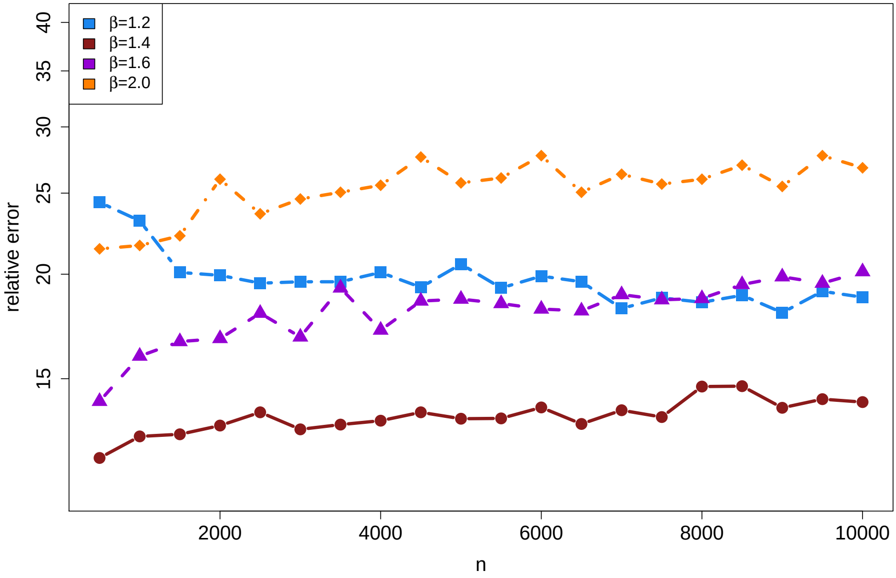
Numerical Results for Reinsurance Example



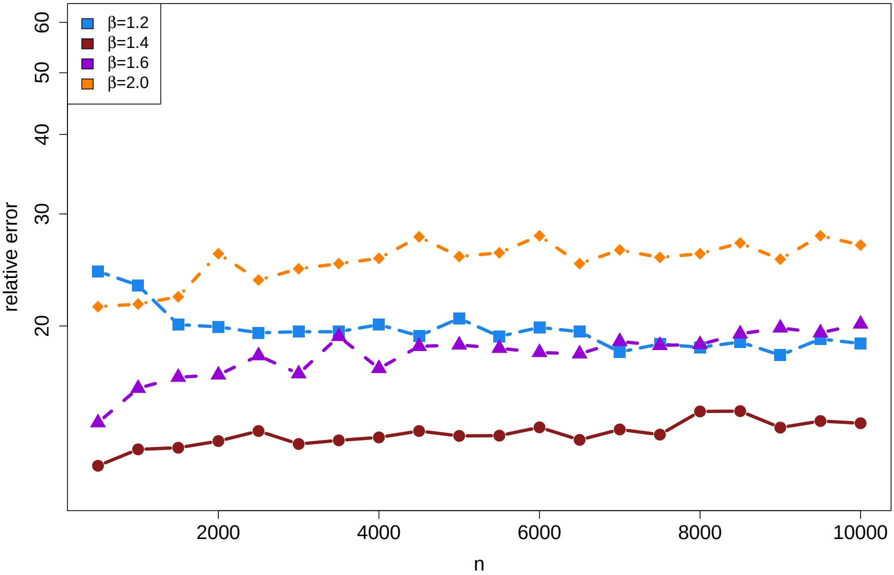
Numerical Results for Reinsurance Example



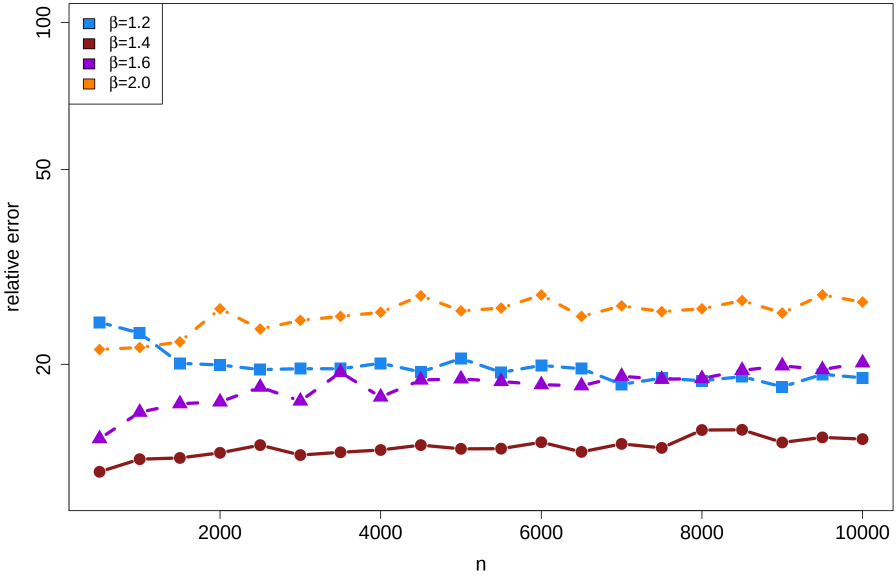
Numerical Results for Reinsurance Example



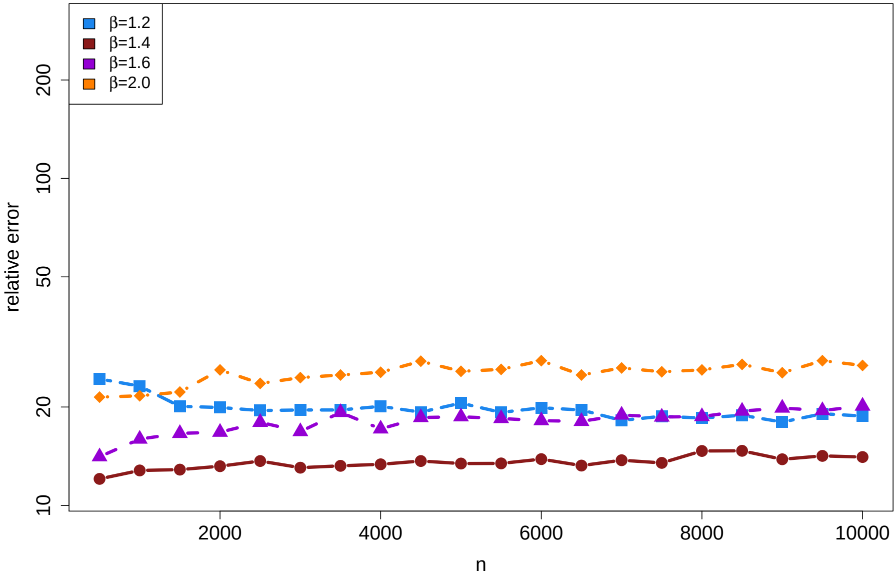
Numerical Results for Reinsurance Example



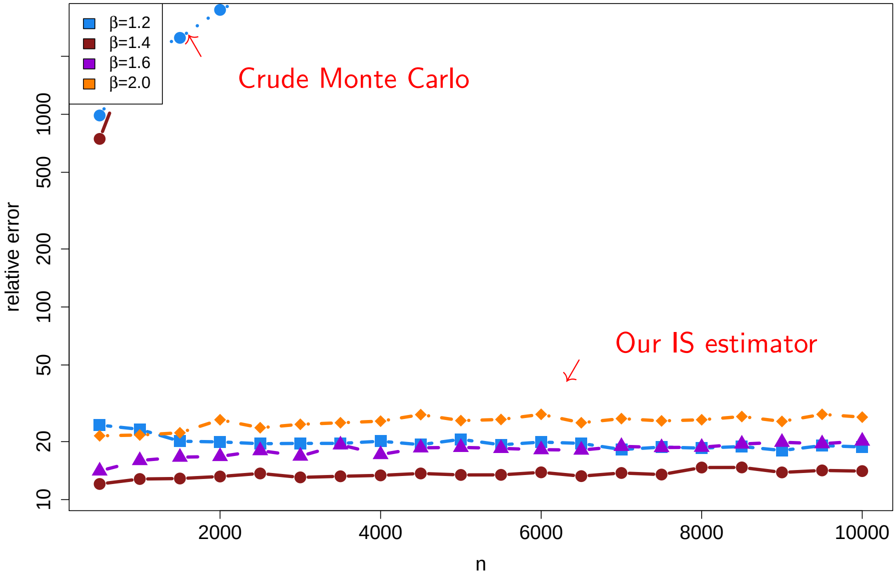
Numerical Results for Reinsurance Example



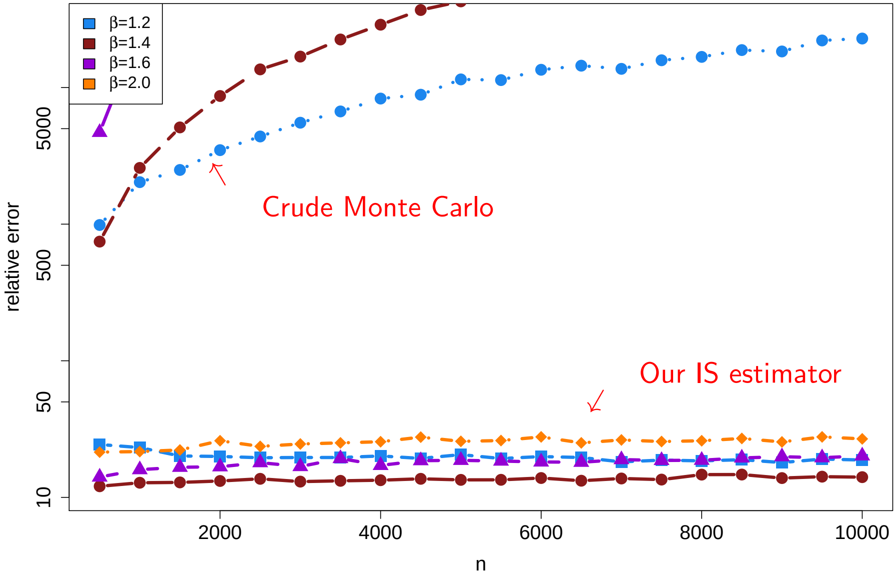
Numerical Results for Reinsurance Example



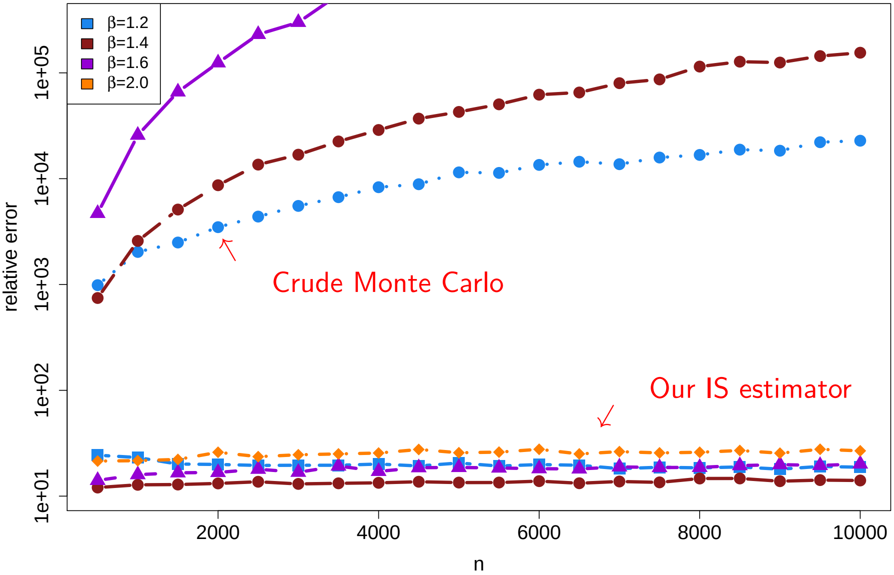
Numerical Results for Reinsurance Example



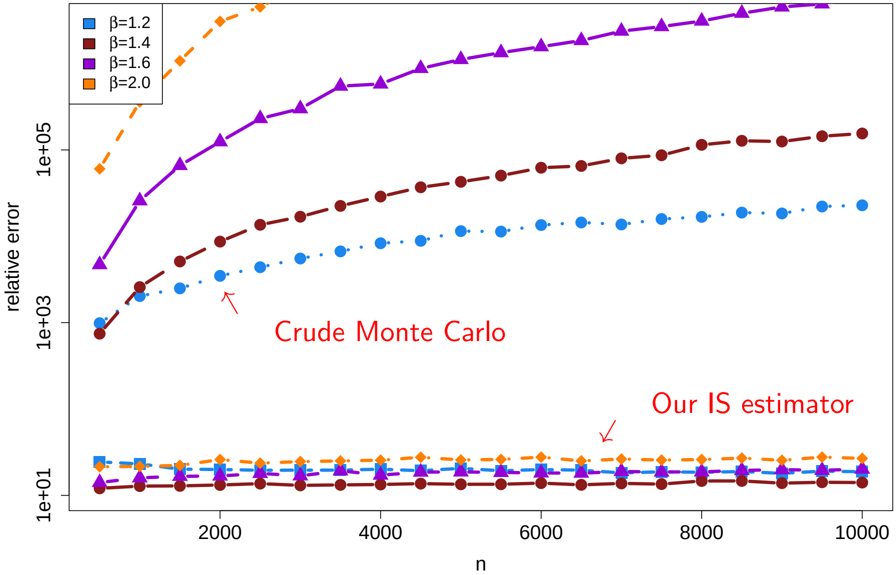
Numerical Results for Reinsurance Example



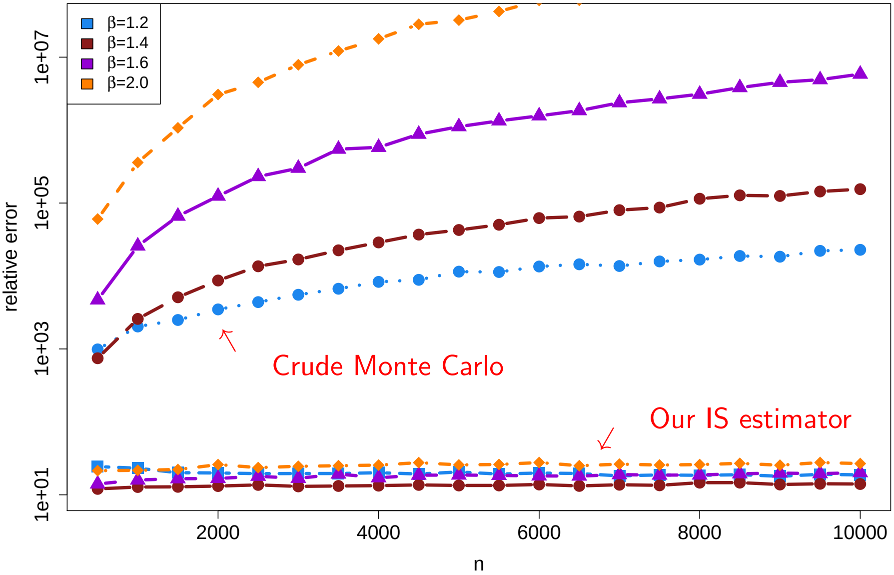
Numerical Results for Reinsurance Example



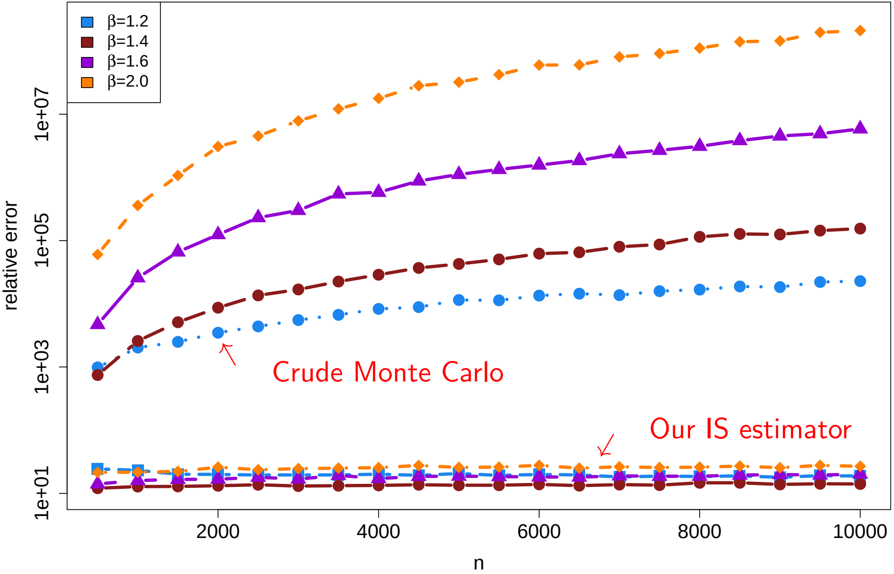
Numerical Results for Reinsurance Example



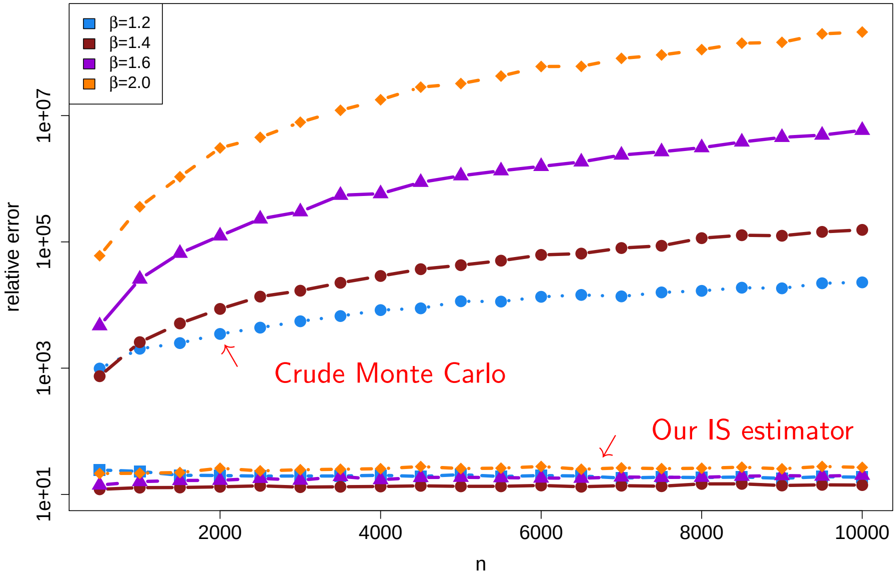
Numerical Results for Reinsurance Example



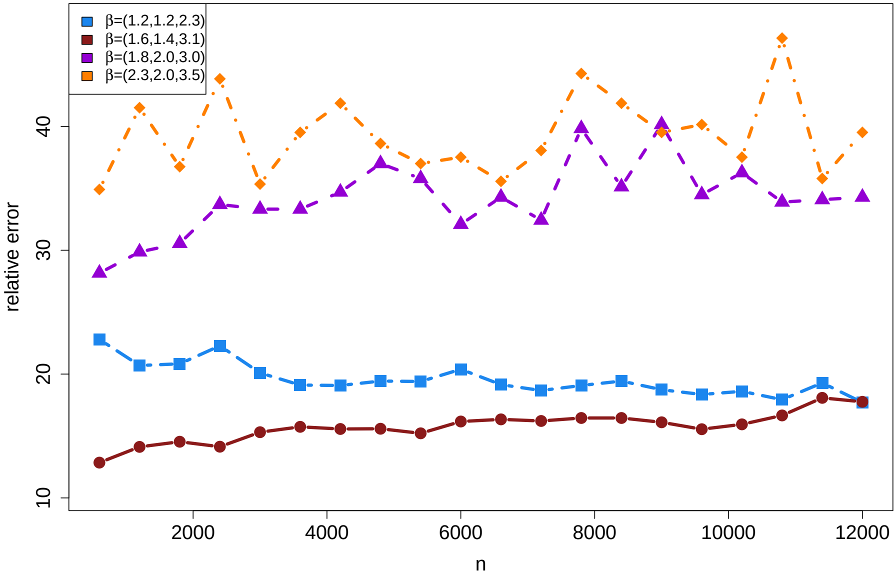
Numerical Results for Reinsurance Example



Numerical Results for Reinsurance Example



Numerical Results for Fluid Network Example



Large Deviations for Weibull Tails

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

Nina Gantert (1998)

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

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

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

Nina Gantert (1998)

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

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

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

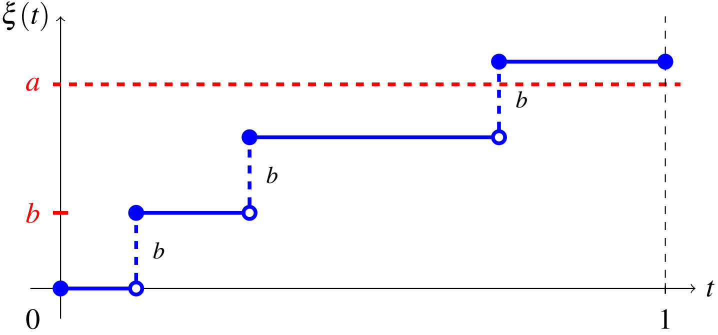
Nina Gantert (1998)

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

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

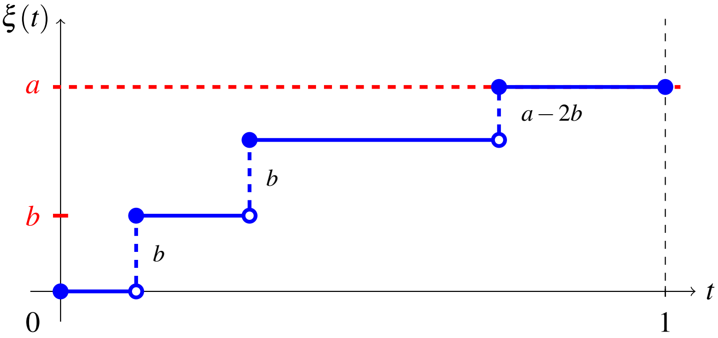
BUT

L_1 topology is weak



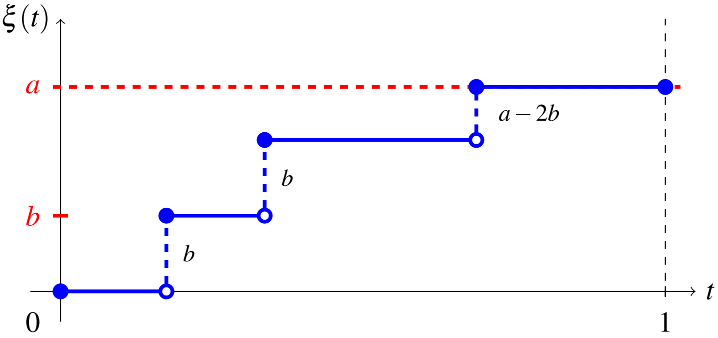
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + b^{\alpha}$

L_1 topology is weak



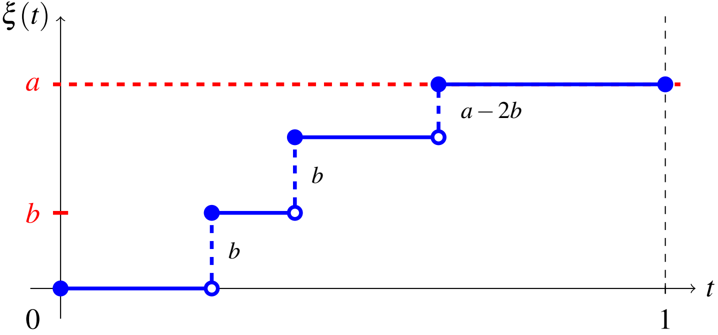
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



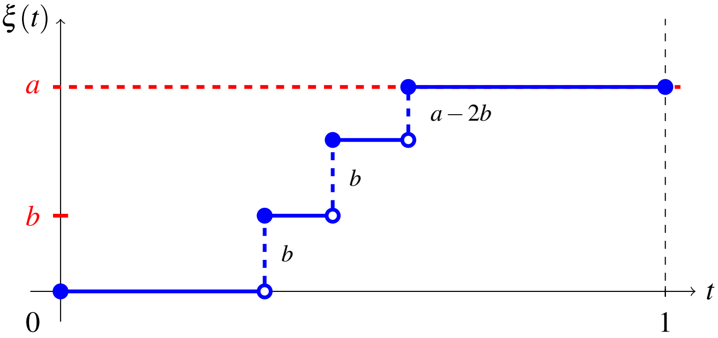
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



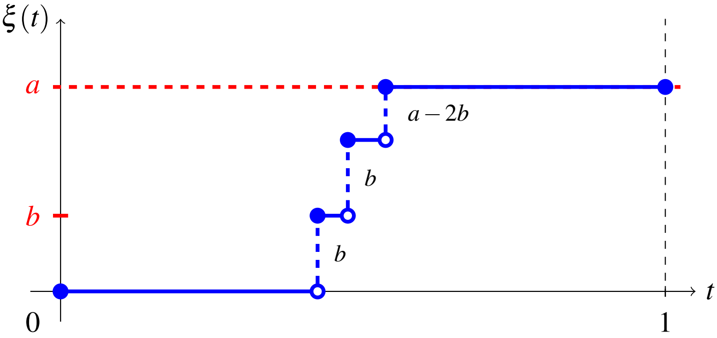
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



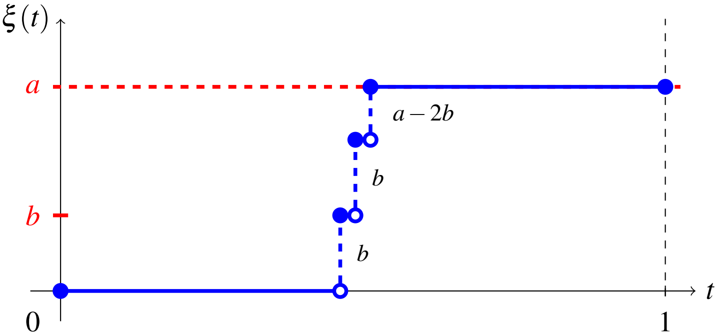
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^{-})\right)$, $\qquad I(A^{-}) = \inf_{\xi\in A^{-}} I(\xi)$
- $I(A^{-}) \stackrel{?}{=} b^{\alpha} + b^{\alpha}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

L_1 topology is weak



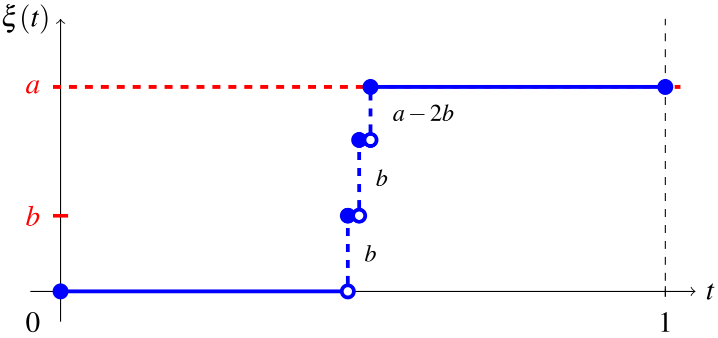
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \overset{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



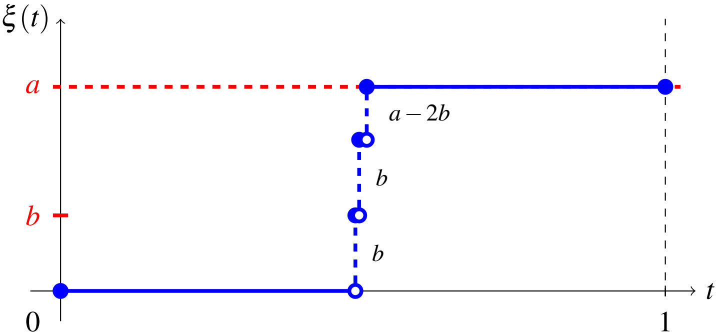
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \overset{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



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

L_1 topology is weak

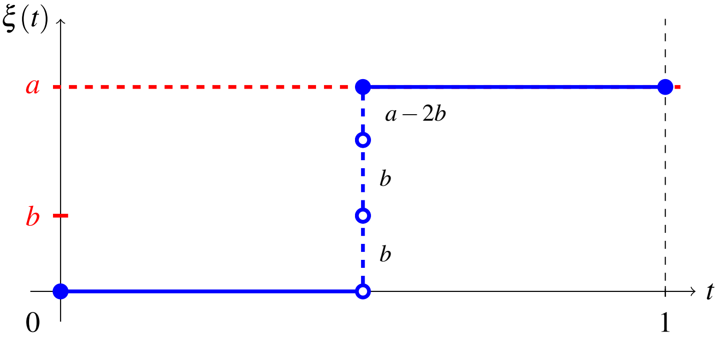


- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

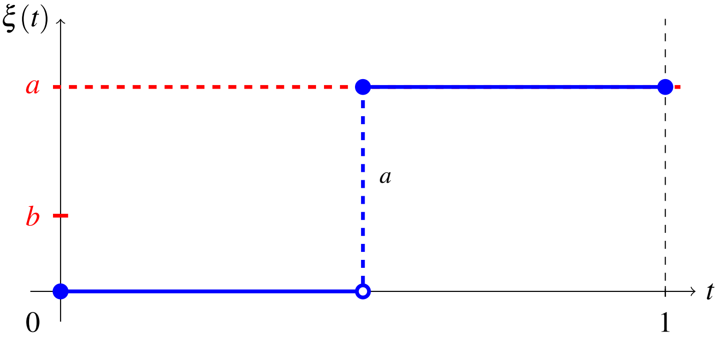
- $I(A^-) \overset{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



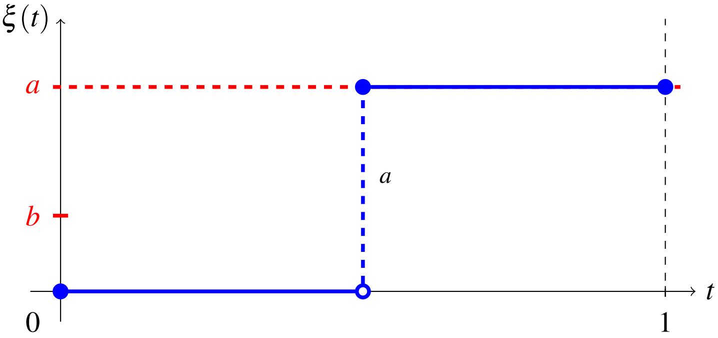
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^{\alpha} + b^{\alpha} + (a-2b)^{\alpha}$

L_1 topology is weak



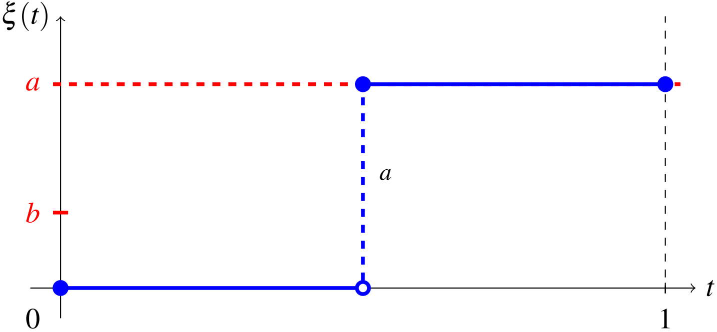
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^\alpha I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \stackrel{?}{=} b^\alpha + b^\alpha + (a-2b)^\alpha$

L_1 topology is weak



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

L_1 topology is weak

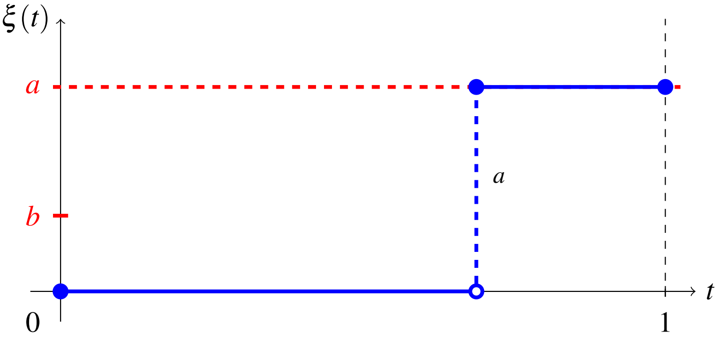


- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

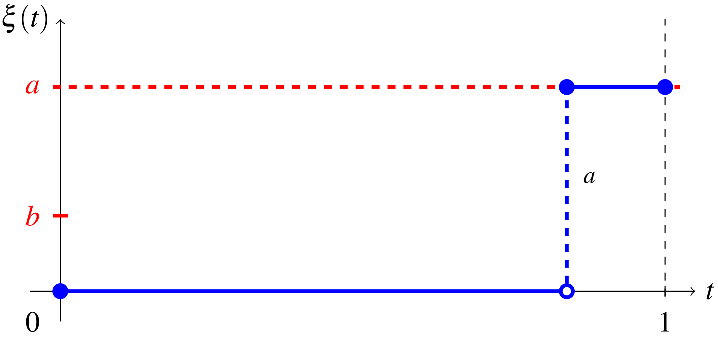
- $I(A^{-}) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^{-}$

L_1 topology is weak



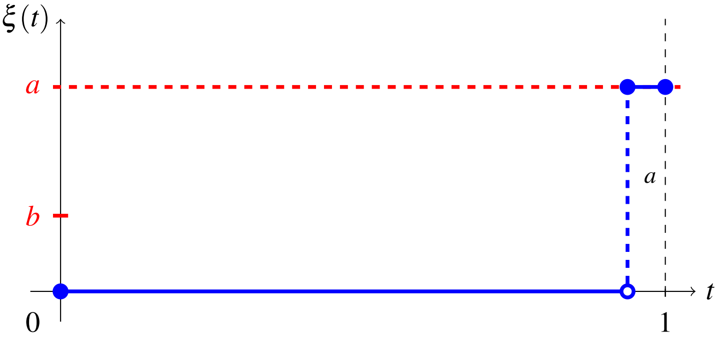
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^{-})\right)$, $\quad I(A^{-}) = \inf_{\xi\in A^{-}} I(\xi)$
- $I(A^{-}) \leq a^{\alpha} < b^{\alpha} + b^{\alpha}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

L_1 topology is weak



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

L_1 topology is weak

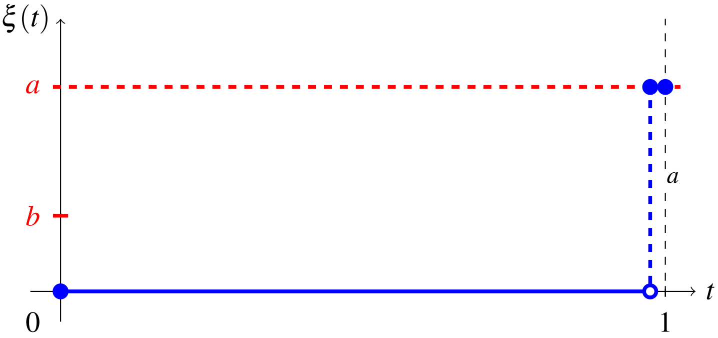


- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

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

L_1 topology is weak

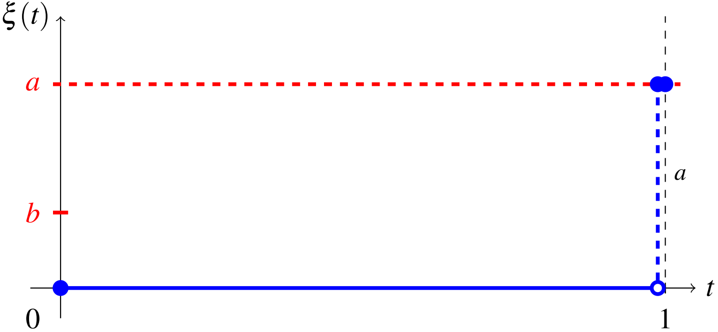


- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

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

L_1 topology is weak

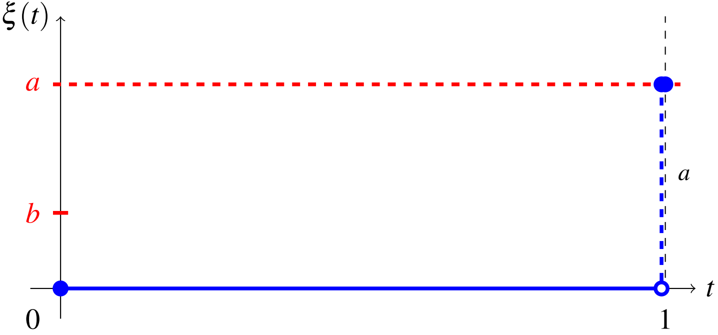


- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

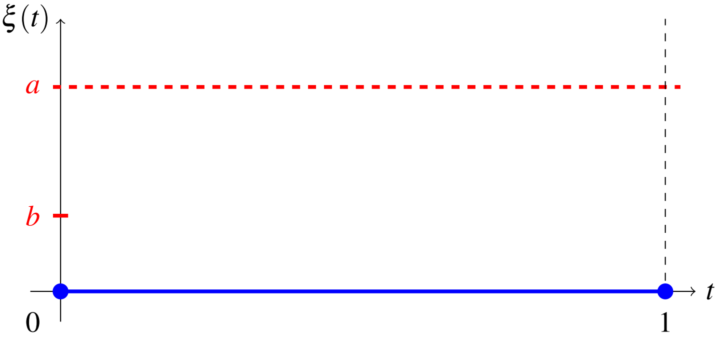
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad a\mathbb{1}_{[1/2,1]} \in A^-$

L_1 topology is weak



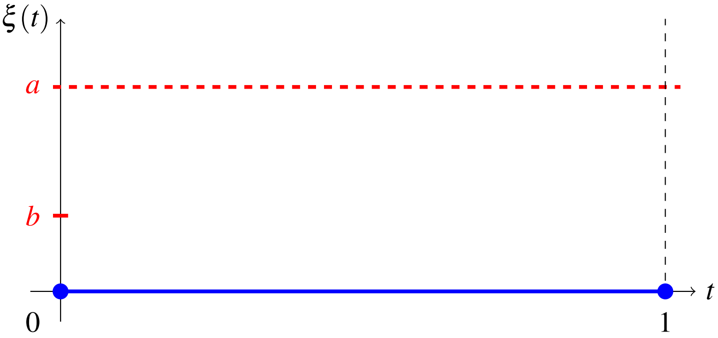
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right)$, $\qquad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha}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

L_1 topology is weak



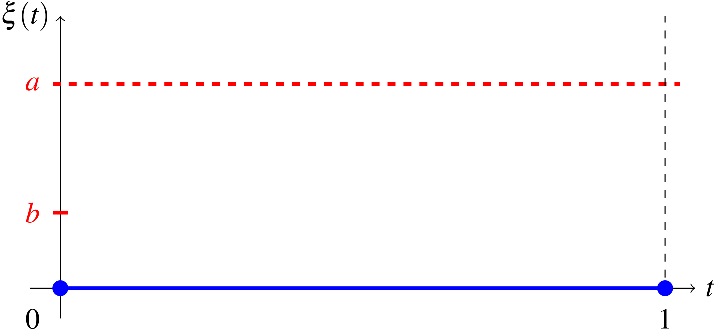
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a, \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \lesssim \exp\left(-L(n)n^{\alpha} I(A^-)\right),$ $\quad I(A^-) = \inf_{\xi\in A^-} I(\xi)$
- $I(A^-) \leq a^{\alpha} < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha}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

L_1 topology is weak



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

L_1 topology is weak



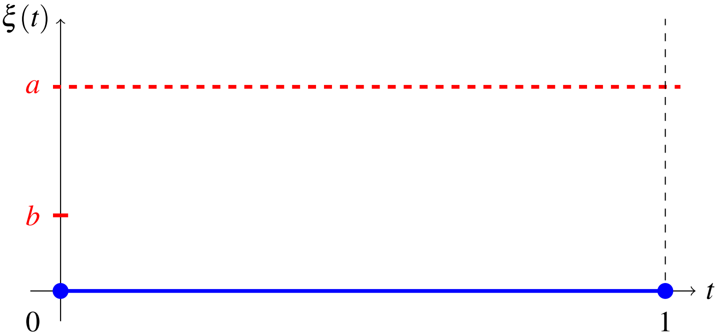
- $A = \{\xi \in \mathbb{D} : \sup_{t\in[0,1]} \xi(t) \geq a,\ \sup_{t\in[0,1]} |\xi(t) - \xi(t-)| \leq b\}$

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

- $I(A^-) = 0 < b^{\alpha} + b^{\alpha} + (b-2a)^{\alpha} \quad \Longleftarrow \quad 0 \in A^-$

No information!

L_1 topology is weak



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

Want a stronger topology, ideally J_1 topology!

LDP w.r.t. J_1 Topology is Impossible

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

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

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

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

“Extended” Large Deviation Principle

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

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

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

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

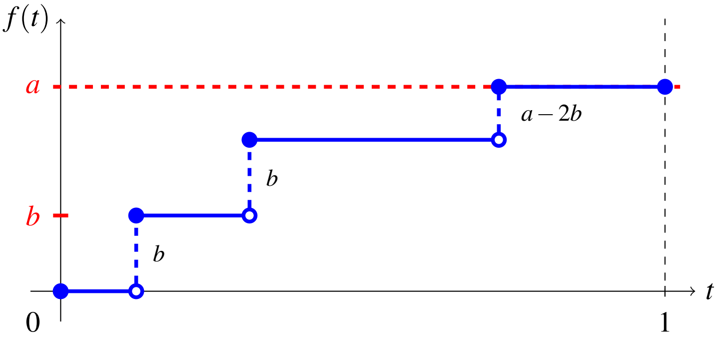
where

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

Corollary

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

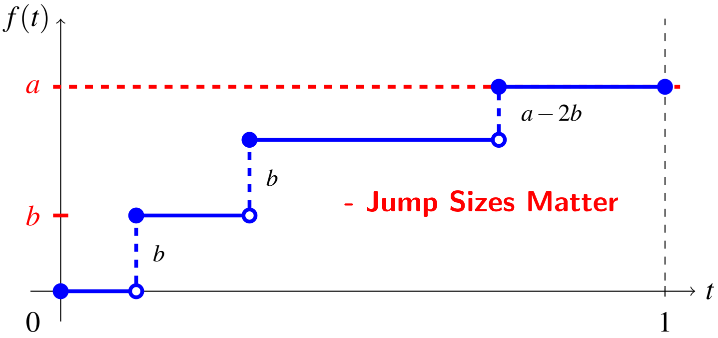
Back to our old example



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

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

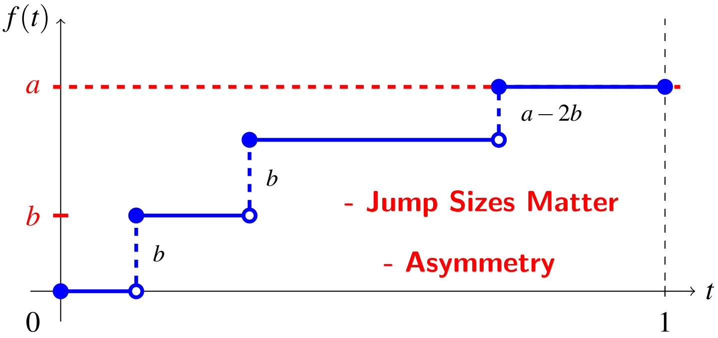
Back to our old example



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

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

Back to our old example

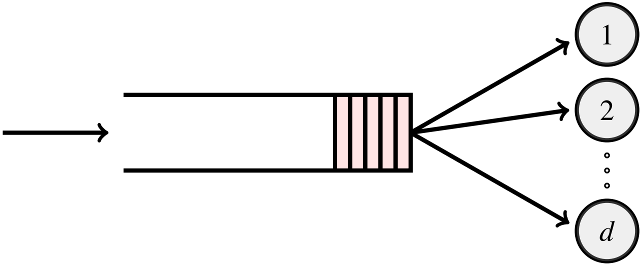


- $A = \{f \in \mathbb{D} : f$ crosses level a on $[0,1]$ & jump sizes $\leq b\}$
- $\mathbf{P}\left(\bar{S}_n \in A\right) \;\sim\; \exp\left(-n^{\alpha} I^*\right)$

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

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

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

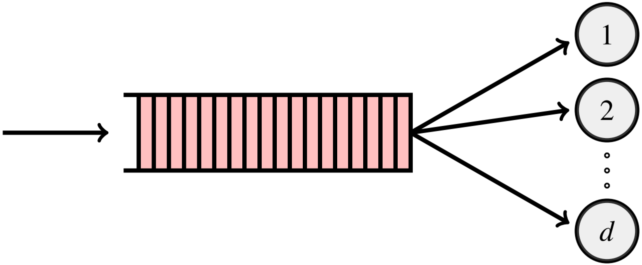


Tail Asymptotics? Most likely scenario?

So far, NOT EVEN a reasonable conjecture!

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?

So far, NOT EVEN a reasonable conjecture!

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

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

- $A(t) \;\; = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}, \qquad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

inverse of $A(\cdot)$

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times | iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\qquad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

inverse of $A(\cdot)$

- $M(t) = \inf\{s : A(s) > t\}$, $\qquad N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times — iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\qquad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

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

- $M(t) = \inf\{s : A(s) > t\}$, $\qquad N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$

Our Sample Path LDP Answers the Question!

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

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

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

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times — iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\qquad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

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

- $M(t) = \inf\{s : A(s) > t\}$, $\qquad N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$

scaled processes (fluid scale)

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

iid service times for server k

- $A(t) \;= A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\qquad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

inverse of $A(\cdot)$

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

- $M(t) \;= \inf\{s : A(s) > t\}$, $\qquad N^{(k)}(t) = \inf\{s : S^{(k)}(s) > t\}$

scaled processes (fluid scale)

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

- 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

iid inter-arrival times iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

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

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

scaled processes (fluid scale)

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

- LDP for A and $S^{(k)}$ $\overset{\text{"contraction principle"}}{\Longrightarrow}$ 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

iid inter-arrival times

iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

inverse of $A(\cdot)$

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

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

scaled processes (fluid scale)

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

"contraction principle"

- LDP for A and $S^{(k)}$ $\Longrightarrow$ LDP for $\bar{M}_n$ and $\bar{N}_n^{(k)}$ (in M_1' topology)

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

Our Sample Path LDP Answers the Question!

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

iid inter-arrival times

iid service times for server k

- $A(t) = A_1 + A_2 + \cdots + A_{\lfloor t \rfloor}$, $\quad S^{(k)}(t) = S_1^{(k)} + S_2^{(k)} + \cdots + S_{\lfloor t \rfloor}^{(k)}$

inverse of $A(\cdot)$

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

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

scaled processes (fluid scale)

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

- LDP for A and $S^{(k)}$ $\overset{\text{“contraction principle”}}{\Longrightarrow}$ LDP for $\bar{M}_n$ and $\bar{N}_n^{(k)}$ (in M_1' topology)

$$\mathbf{P}(Q(\gamma n) > n) \approx \mathbf{P}\Bigg(\underbrace{\sup_{0 \le s \le \gamma} \Big(\bar{M}_n(s) - \sum_{k=1}^{d} \bar{N}_n^{(k)}(s) \Big)}_{\text{continuous functional of } \bar{M}_n \text{ and } \bar{N}_n^{(k)}\text{'s in } M_1' \text{ topology}} > 1 \Bigg)$$

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

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

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

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

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

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

where

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

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

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

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

where

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

- Special case of L^α-norm minimization problem.
- We have explicit solution.

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

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

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

where

- Nontrivial Tradeoff between Size and Number of Jumps!

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

- Special case of L^α-norm minimization problem.
- We have explicit solution.

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

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

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

where

- Nontrivial Tradeoff between Size and Number of Jumps!

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

- Asymmetry in Jump Sizes!

- Special case of L^α-norm minimization problem.
- We have explicit solution.

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

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

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

where

- **Nontrivial Tradeoff between Size and Number of Jumps!**
- **Asymmetry in Jump Sizes!**

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

- Special case of L^α-norm minimization problem.
- We have explicit solution.

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

More Specifically,

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

- If $\lfloor \lambda \rfloor \leq \frac{\lambda - \alpha d}{1-\alpha}$: smallest possible number of big jobs to block enough servers
  $\Rightarrow$ **same as the power law case**
- If $\lfloor \lambda \rfloor > \frac{\lambda - \alpha d}{1-\alpha}$: larger number of moderately big jobs might be more likely
  $\Rightarrow$ **qualitatively different from the power law case**

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

Back to Stochastic Gradient Descent

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Then

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Then

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

where

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

Stochastic Gradient Descent

Minimizing loss function f:

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

Then

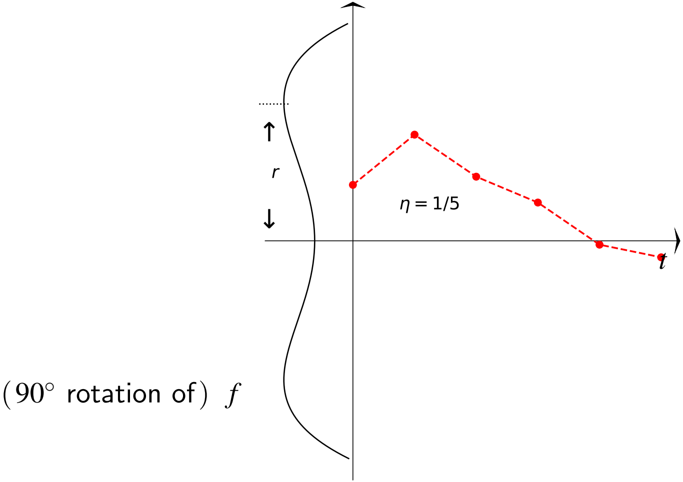
$$W^\eta(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

Gradient Flow

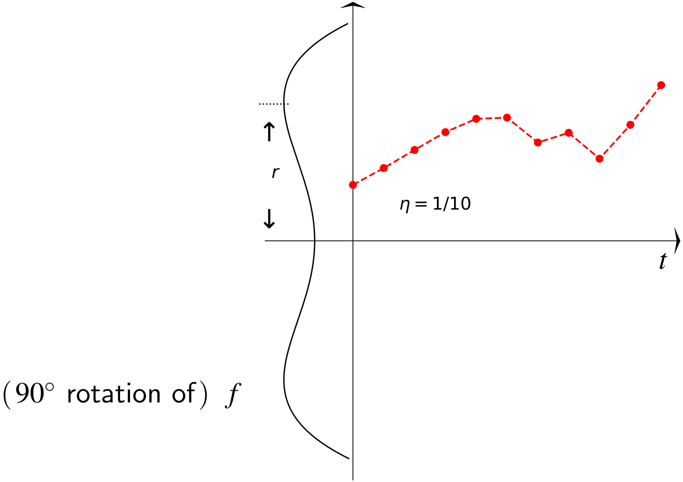
where

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

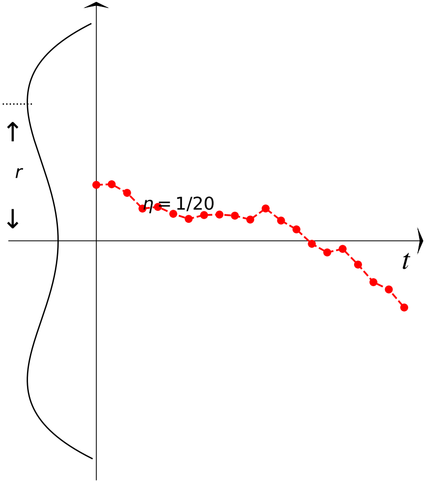
Gradient Flow: Law of Large Numbers



Gradient Flow: Law of Large Numbers

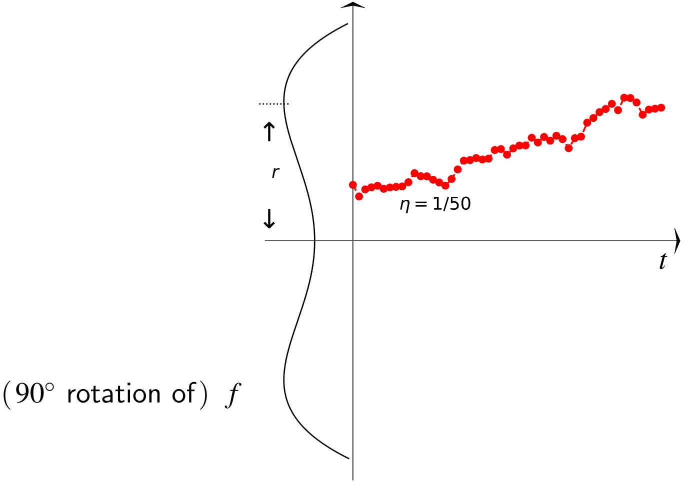


Gradient Flow: Law of Large Numbers

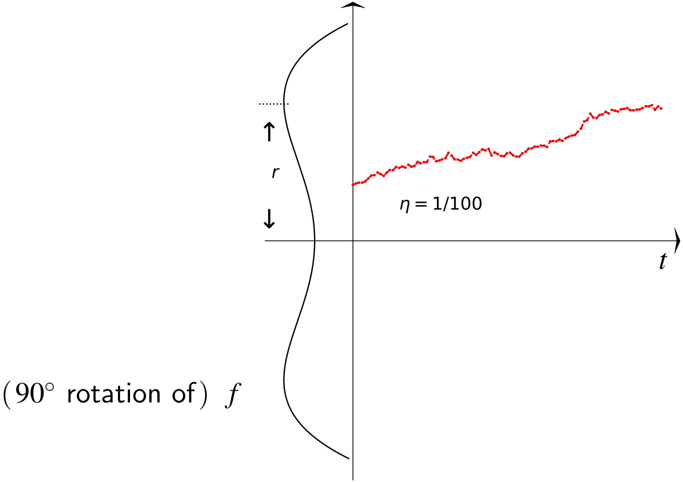


$(90^\circ$ rotation of$)$ f

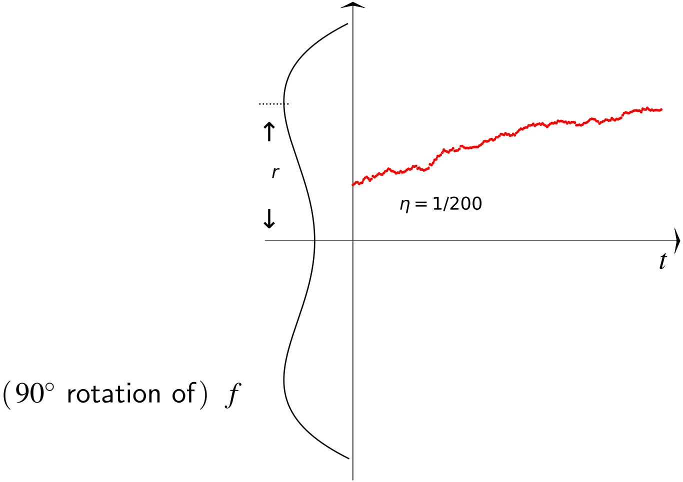
Gradient Flow: Law of Large Numbers



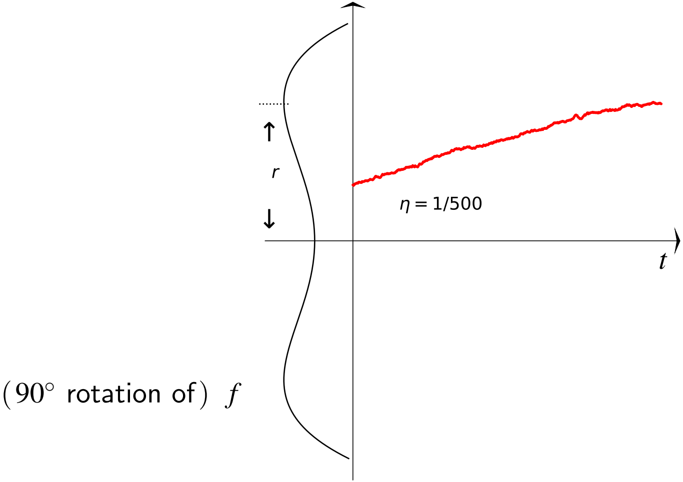
Gradient Flow: Law of Large Numbers



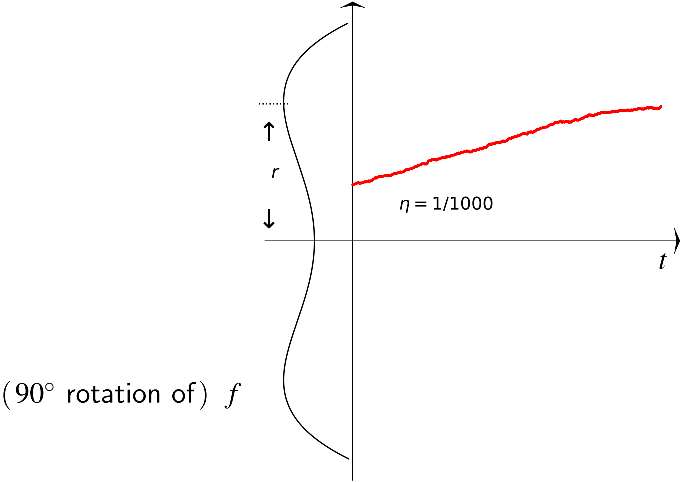
Gradient Flow: Law of Large Numbers



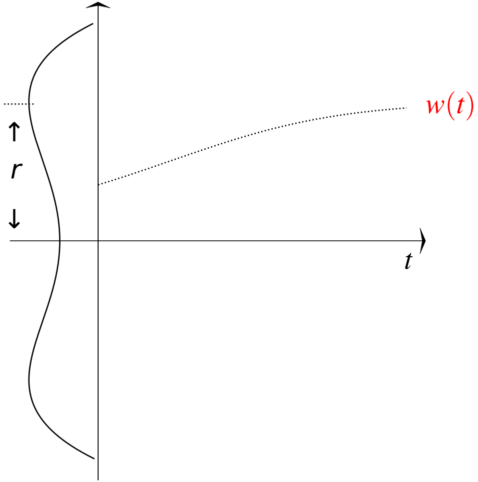
Gradient Flow: Law of Large Numbers



Gradient Flow: Law of Large Numbers

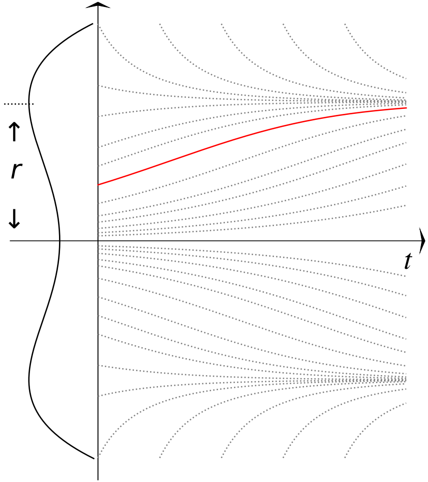


Gradient Flow: Law of Large Numbers



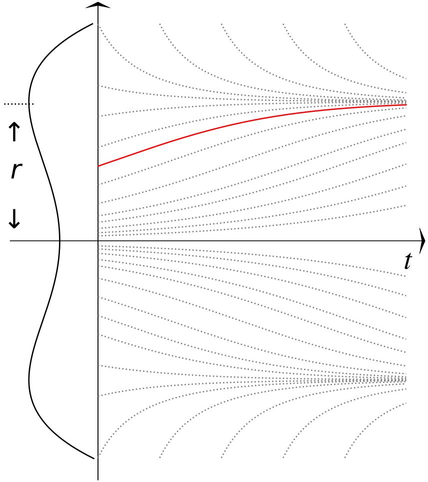
$(90^\circ$ rotation of$)$ f

Gradient Flow: Law of Large Numbers



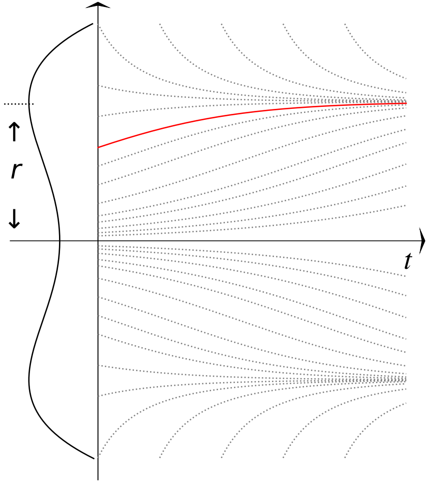
$(90^\circ$ rotation of$)$ f

Gradient Flow: Law of Large Numbers



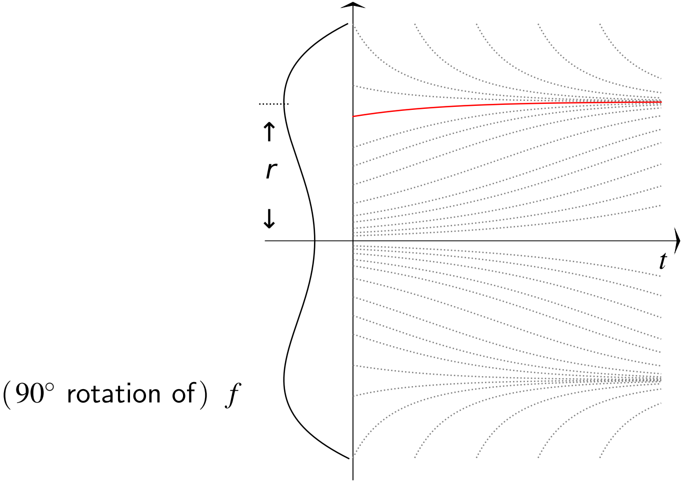
$(90^\circ$ rotation of$)$ f

Gradient Flow: Law of Large Numbers

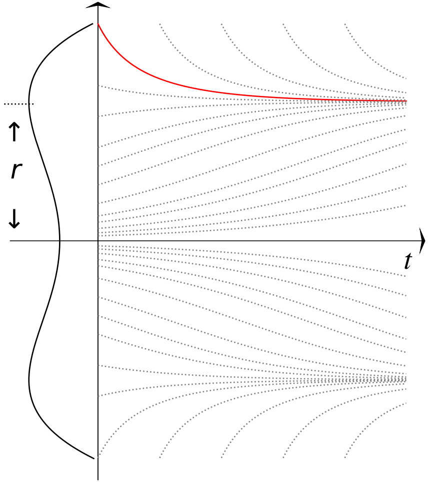


(90° rotation of) f

Gradient Flow: Law of Large Numbers

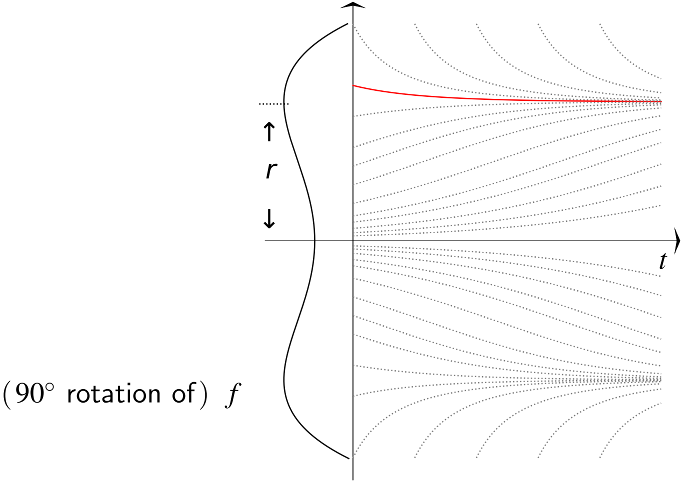


Gradient Flow: Law of Large Numbers

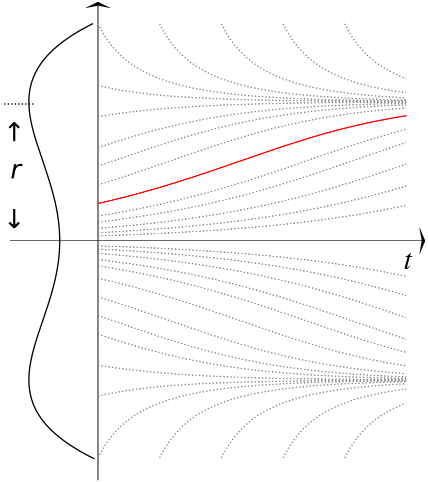


$(90^\circ \text{ rotation of})\ f$

Gradient Flow: Law of Large Numbers

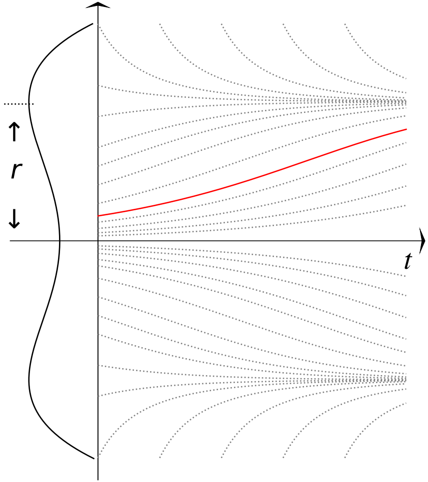


Gradient Flow: Law of Large Numbers



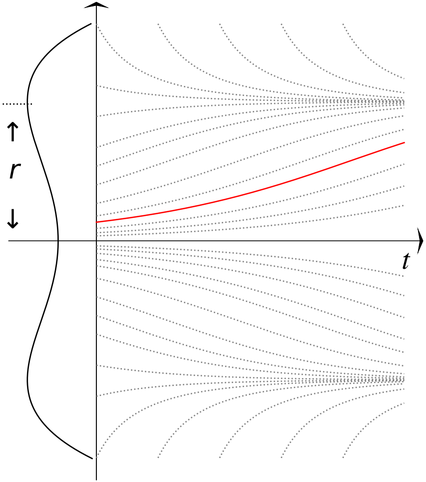
$(90^\circ \text{ rotation of})\ f$

Gradient Flow: Law of Large Numbers



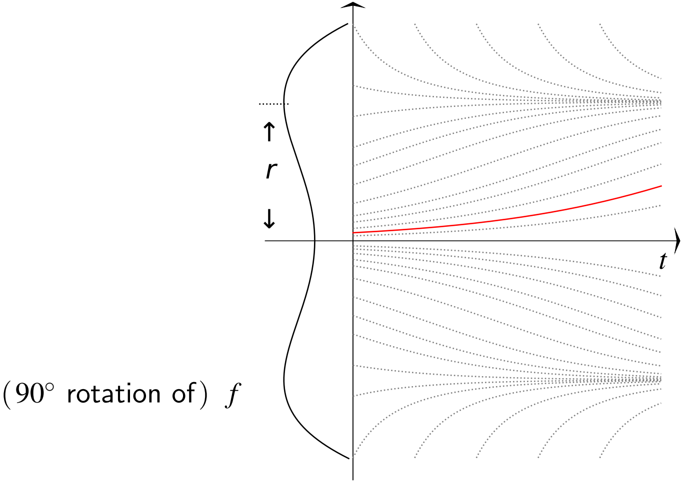
(90° rotation of) f

Gradient Flow: Law of Large Numbers

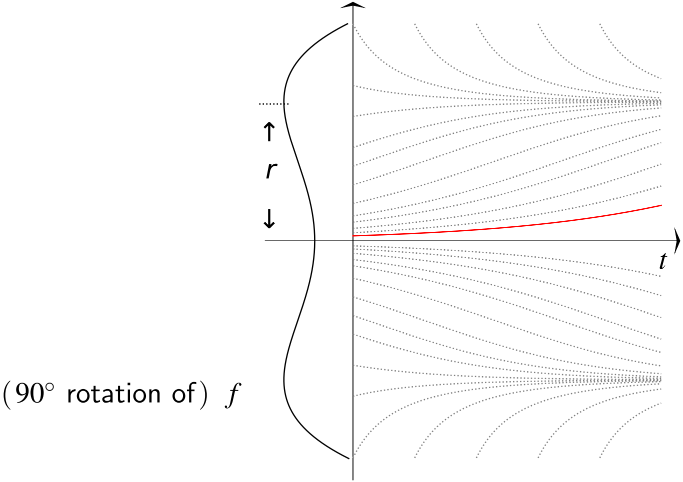


$(90^\circ$ rotation of$)$ f

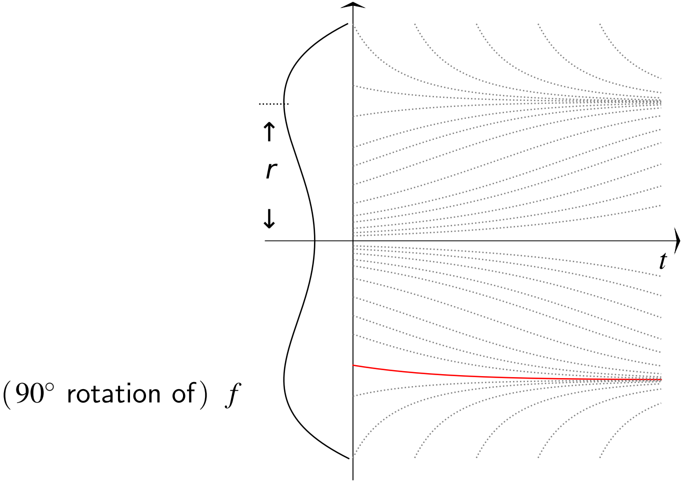
Gradient Flow: Law of Large Numbers



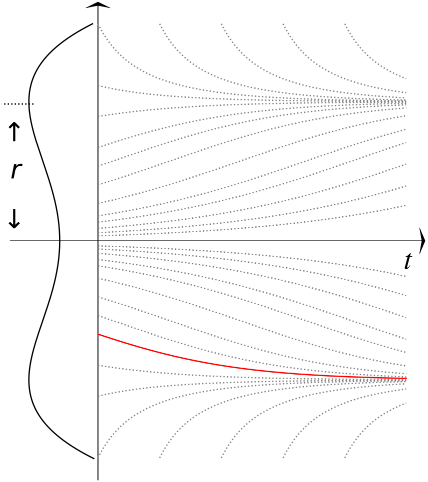
Gradient Flow: Law of Large Numbers



Gradient Flow: Law of Large Numbers

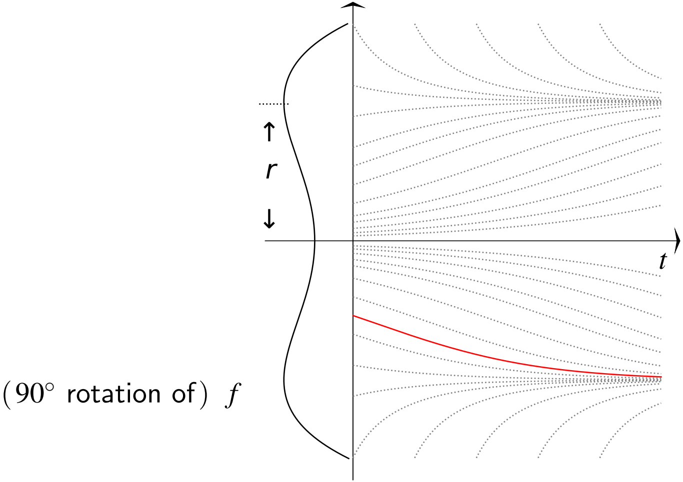


Gradient Flow: Law of Large Numbers

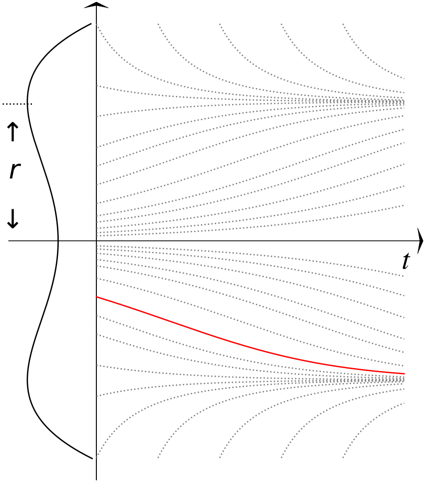


(90° rotation of) f

Gradient Flow: Law of Large Numbers

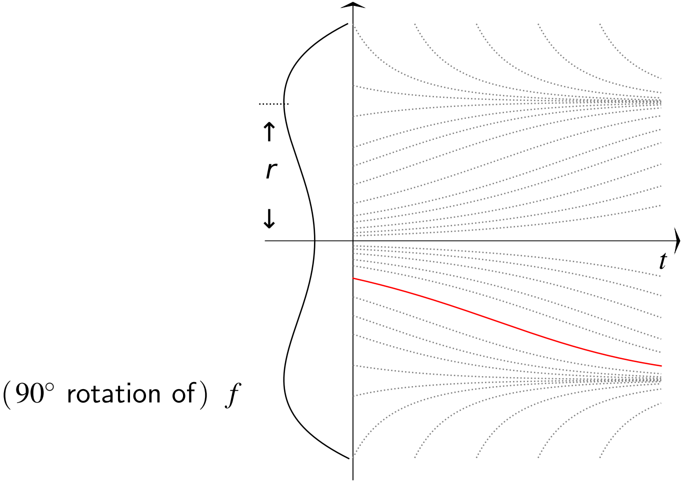


Gradient Flow: Law of Large Numbers



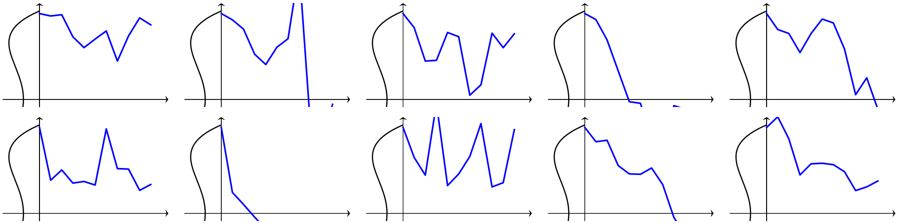
$(90^\circ$ rotation of$)$ f

Gradient Flow: Law of Large Numbers

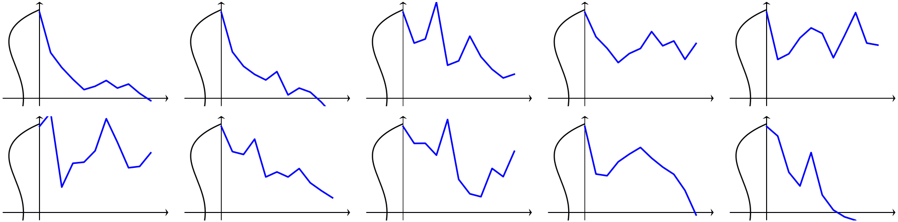


Typical Scenario

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

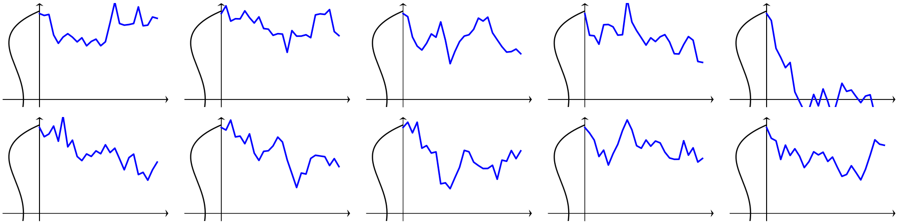


Trajectory of SGD W^η: $\eta = 1/10$ & noises are **heavy-tailed**

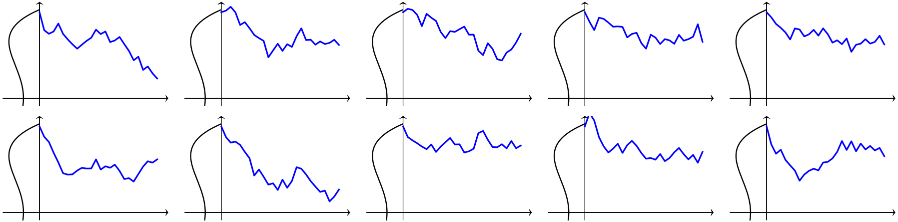


Typical Scenario

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

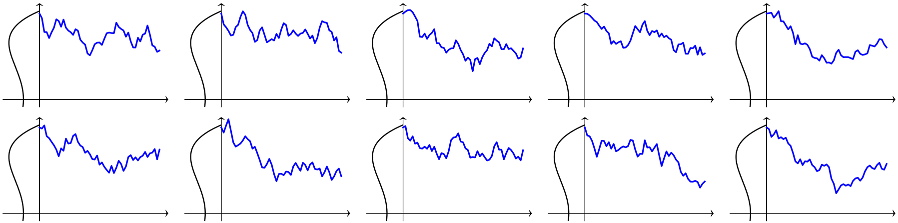


Trajectory of SGD W^η: $\eta = 1/25$ & noises are **heavy-tailed**

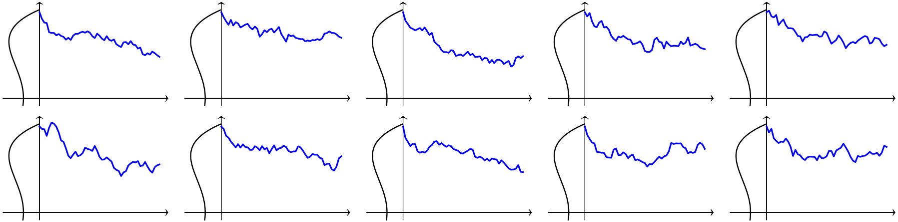


Typical Scenario

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

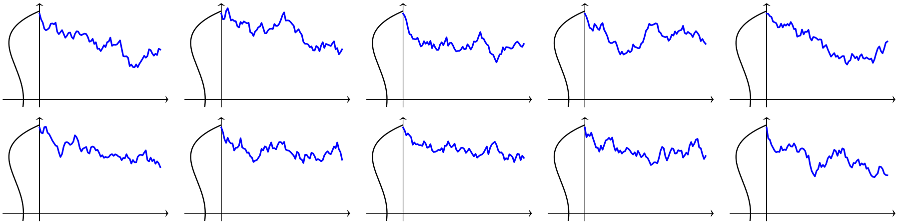


Trajectory of SGD W^η: $\eta = 1/50$ & noises are **heavy-tailed**

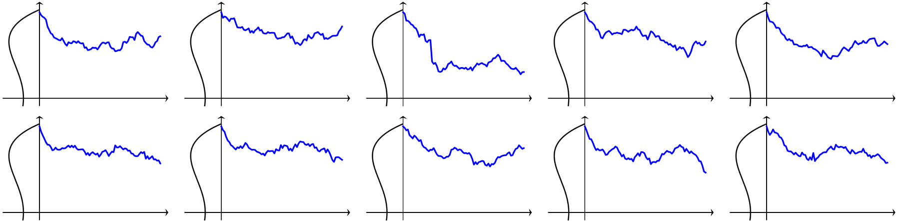


Typical Scenario

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

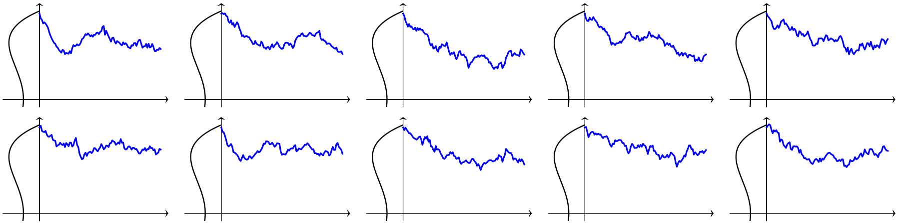


Trajectory of SGD W^{η}: $\eta = 1/75$ & noises are **heavy-tailed**

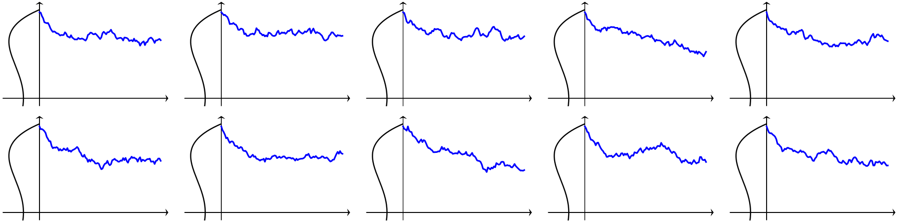


Typical Scenario

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

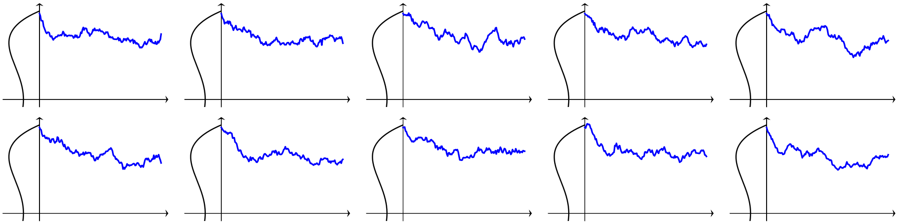


Trajectory of SGD W^{η}: $\eta = 1/100$ & noises are **heavy-tailed**

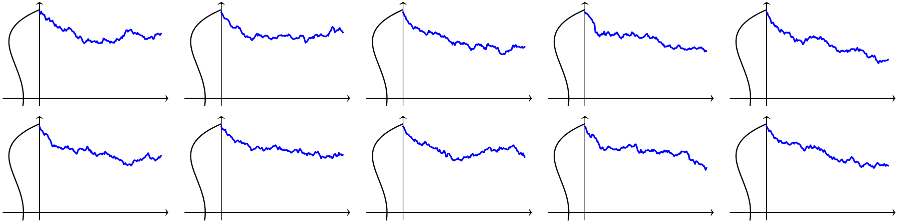


Typical Scenario

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

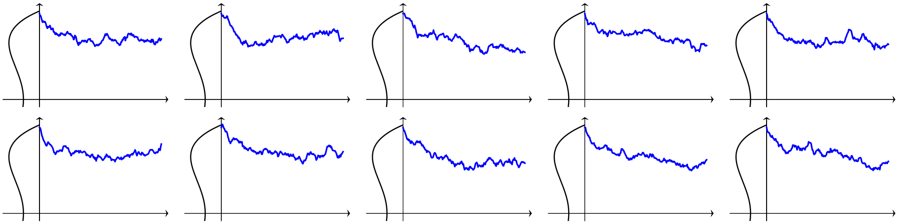


Trajectory of SGD W^{η}: $\eta = 1/150$ & noises are **heavy-tailed**

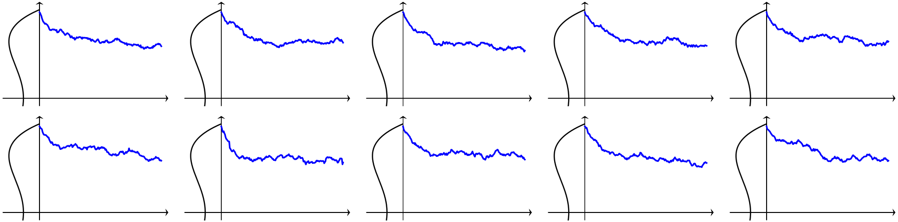


Typical Scenario

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



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



Heavy-Tailed Large Deviations for SGD

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

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

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

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

Heavy-Tailed Large Deviations for SGD

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

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

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

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

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

Heavy-Tailed Large Deviations for SGD

gradient noise $\swarrow$

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

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

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

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

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

Heavy-Tailed Large Deviations for SGD

gradient noise

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

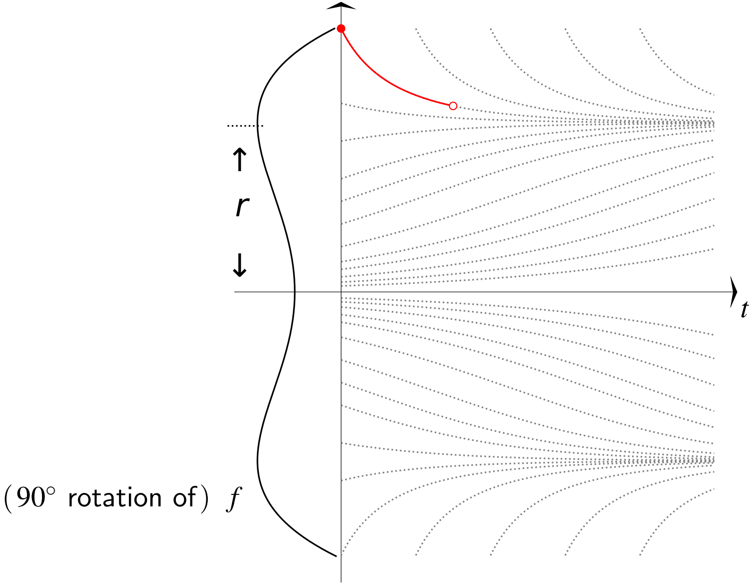
Theorem (Wang, Su, R., 2022+)

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

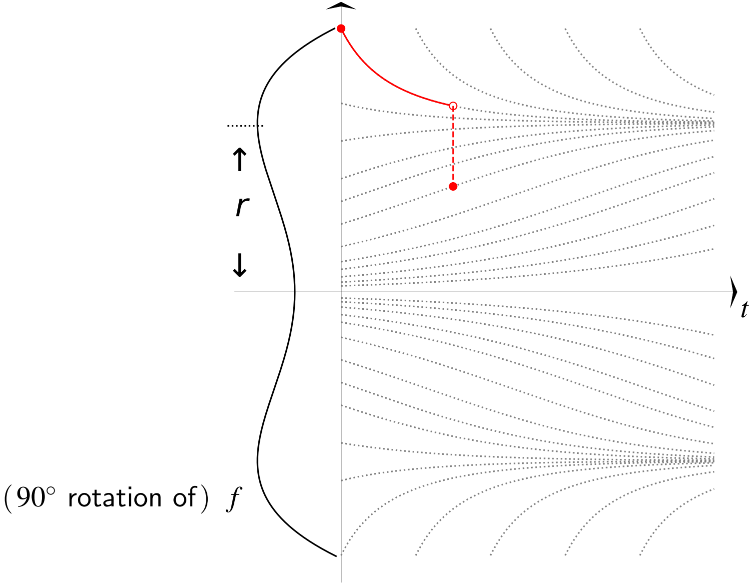
$$\mathbf{P}(W^\eta \in A) \sim n^{-\alpha \mathcal{J}(A)}$$

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

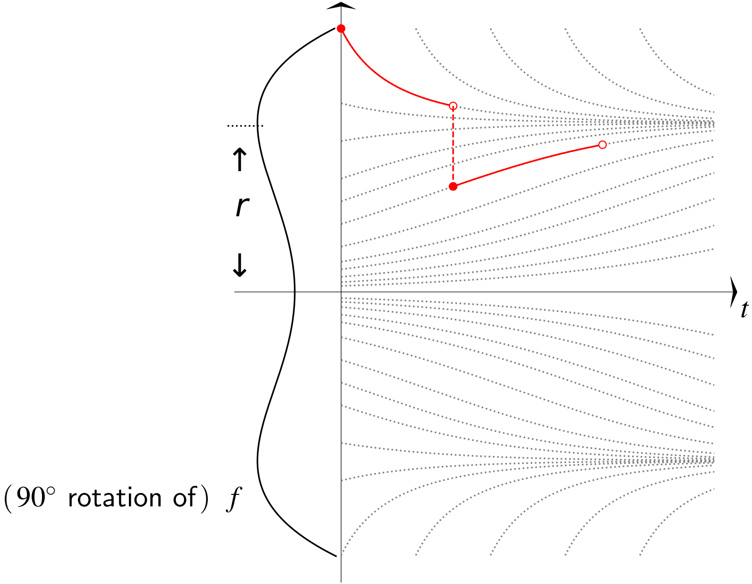
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



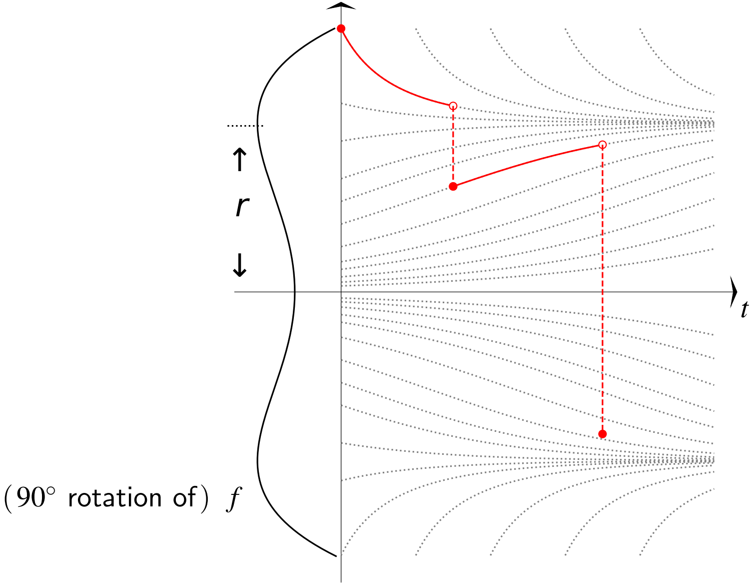
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



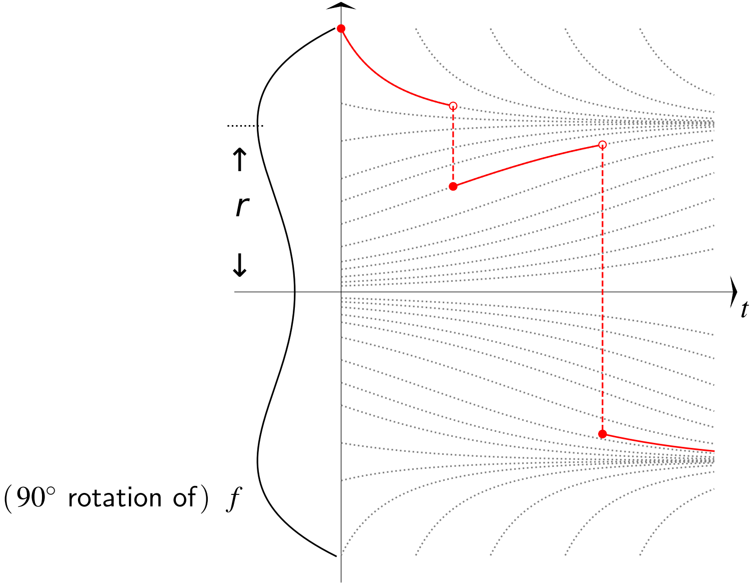
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



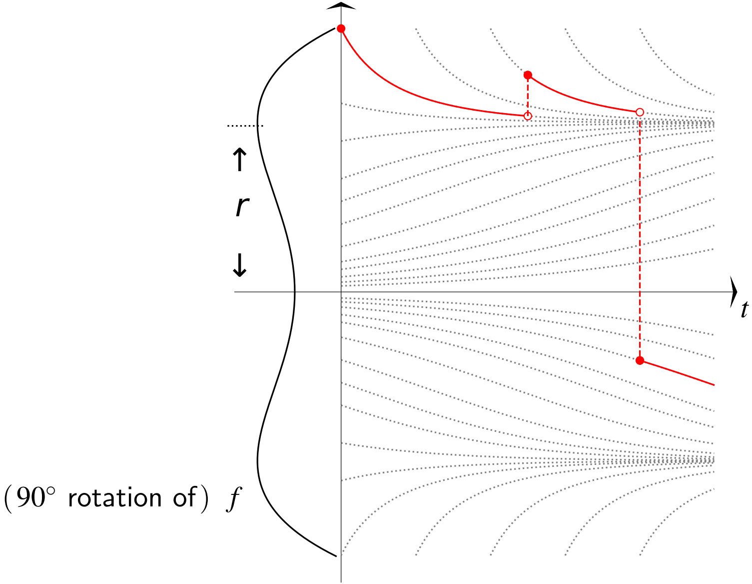
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



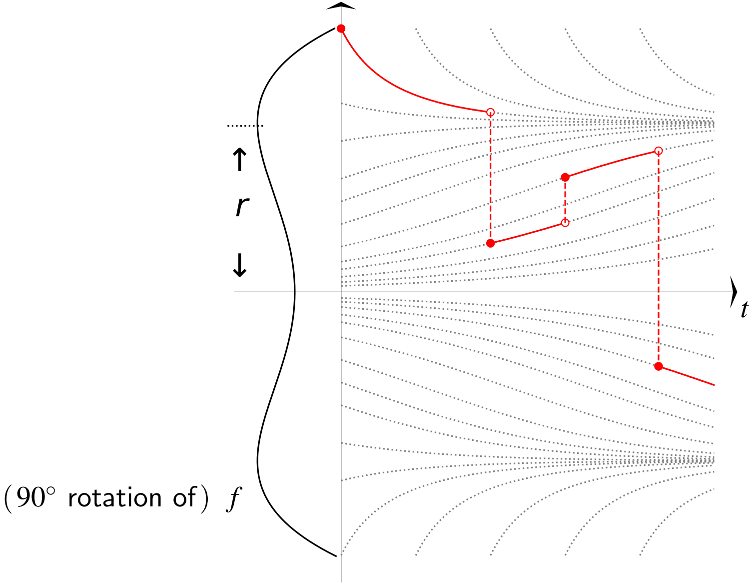
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



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+)

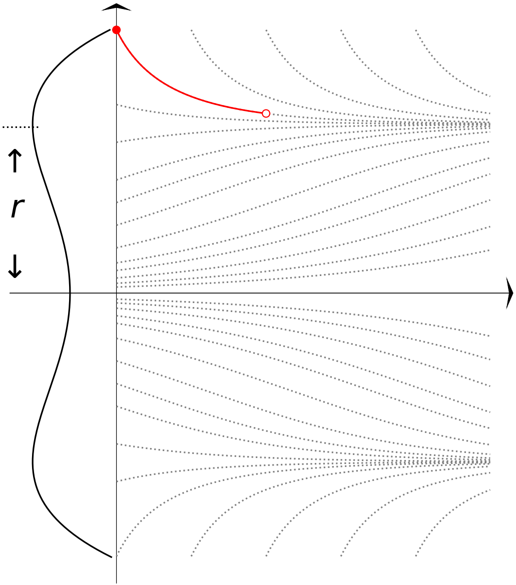
For "general" $A \subseteq \mathbb{D}$

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

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

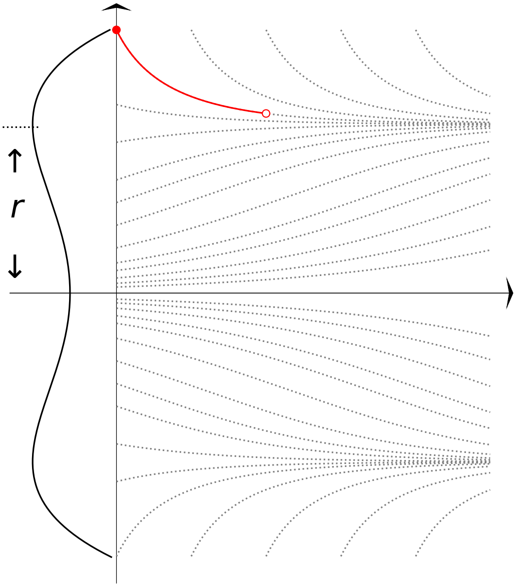
Catastrophe Principle: Most Likely Escape Route

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



Catastrophe Principle: Most Likely Escape Route

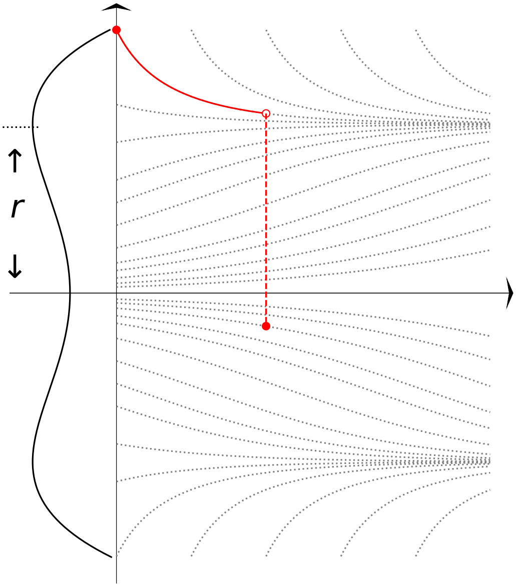
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

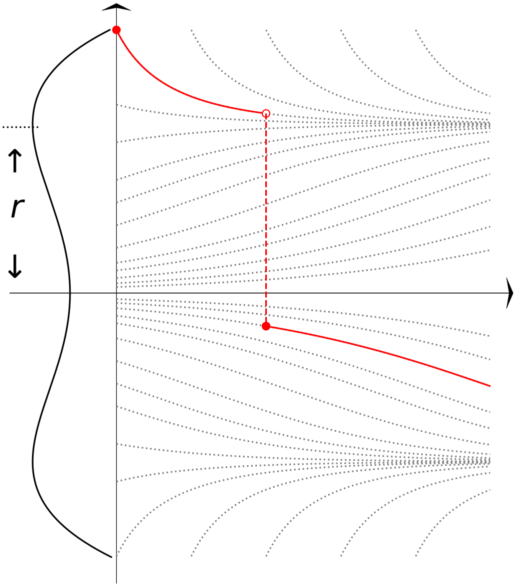
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

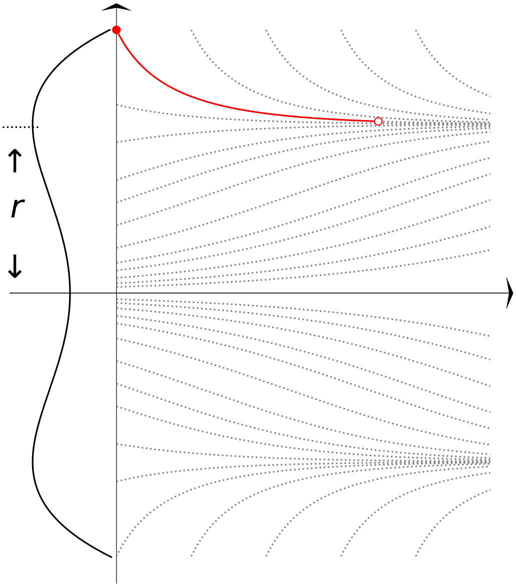
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

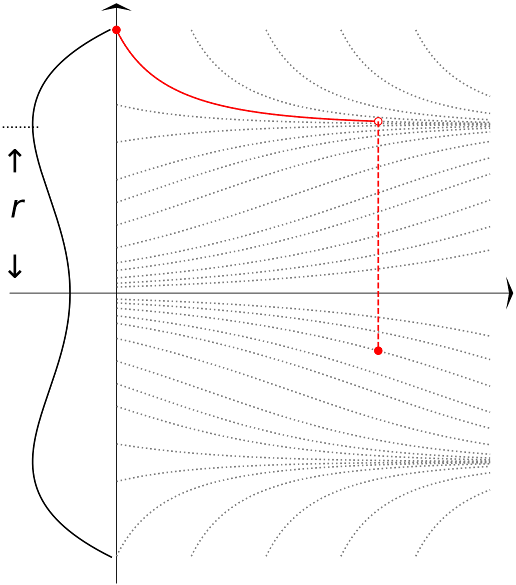
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

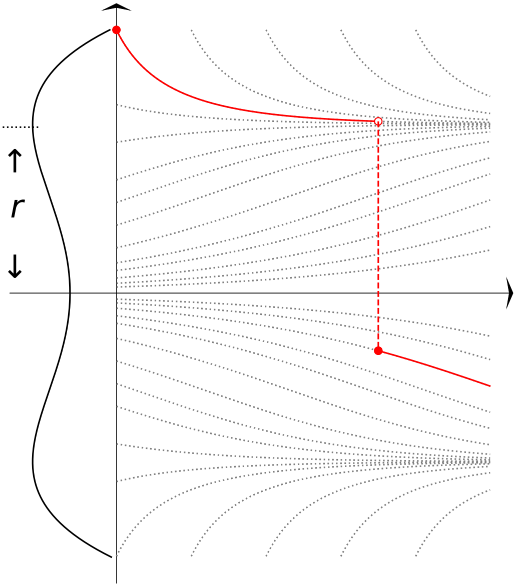
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

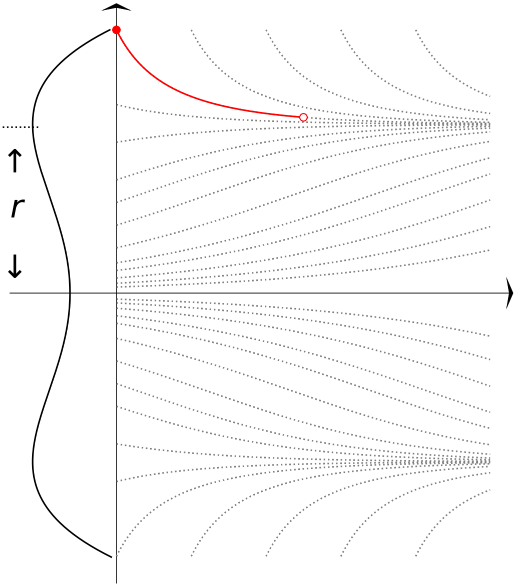
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

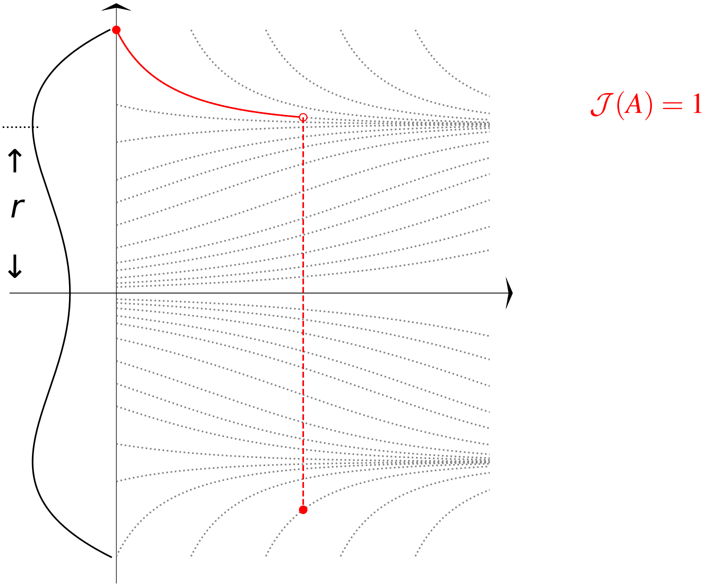
Most Likely Paths for $\{\, W^\eta$ escapes the local minimum $\}$



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

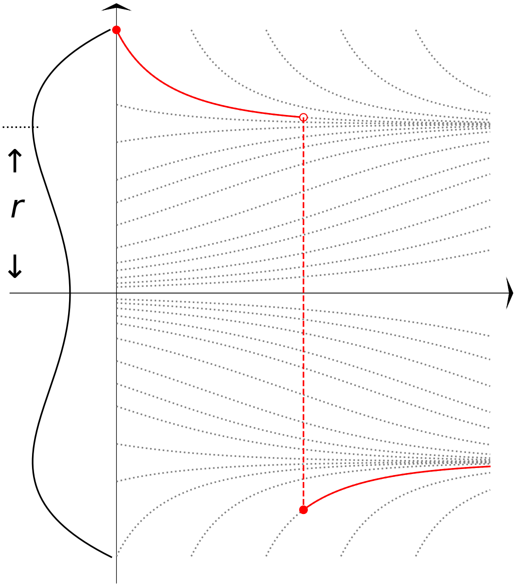
Catastrophe Principle: Most Likely Escape Route

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



Catastrophe Principle: Most Likely Escape Route

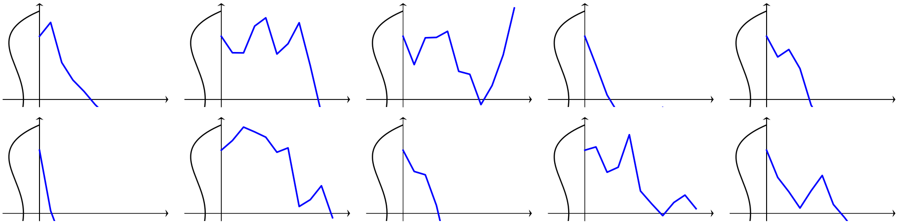
Most Likely Paths for $\{\, W^{\eta}$ escapes the local minimum $\}$



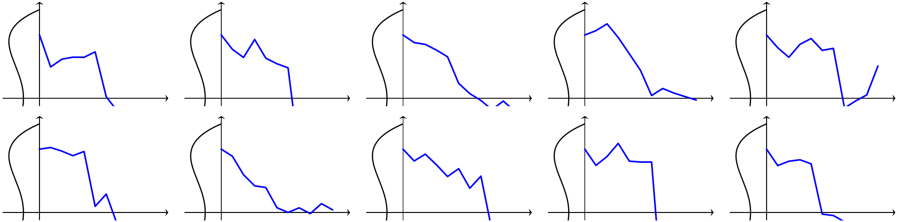
$\mathcal{J}(A) = 1$

Catastrophe Principle Dictates SGD's Escape Route

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

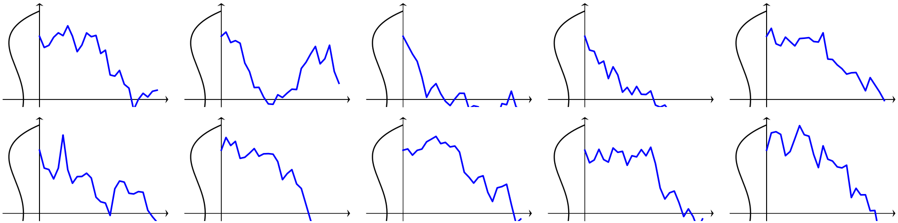


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

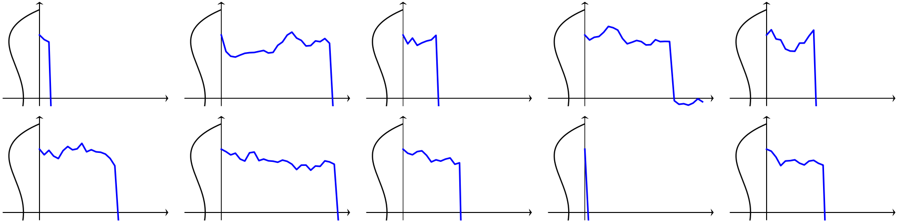


Catastrophe Principle Dictates SGD's Escape Route

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

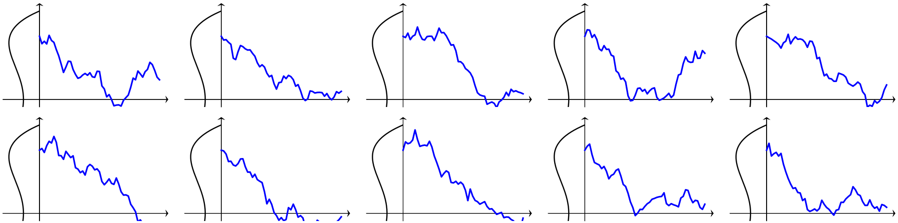


Trajectory of SGD X^{η} conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

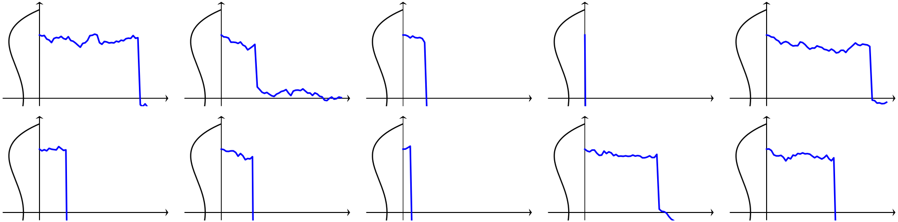


Catastrophe Principle Dictates SGD's Escape Route

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

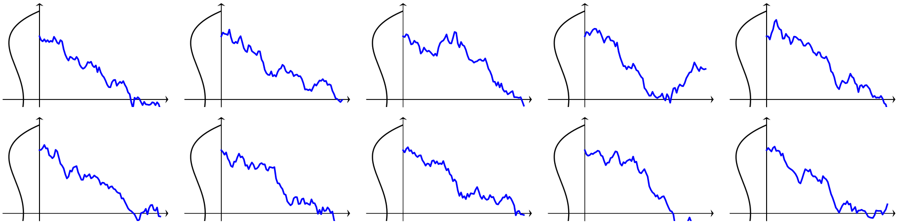


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/50$

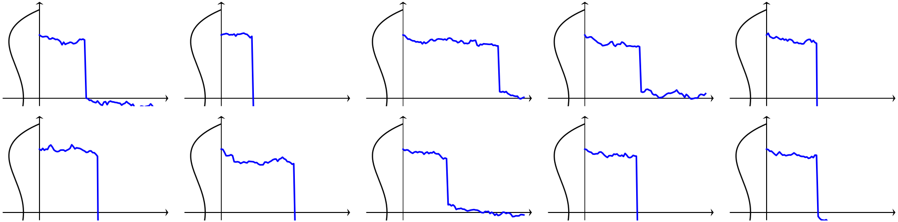


Catastrophe Principle Dictates SGD's Escape Route

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

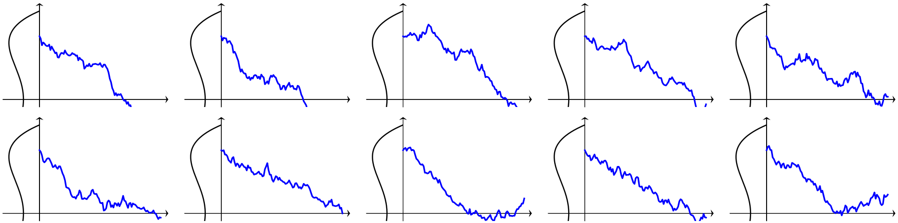


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

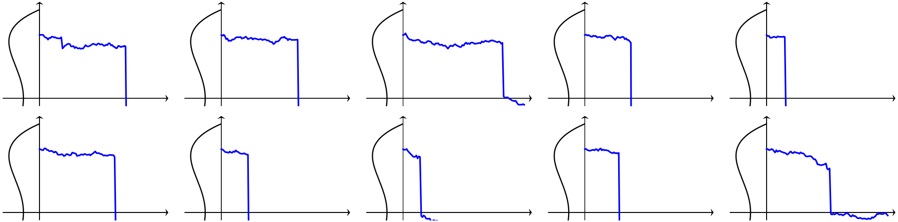


Catastrophe Principle Dictates SGD's Escape Route

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

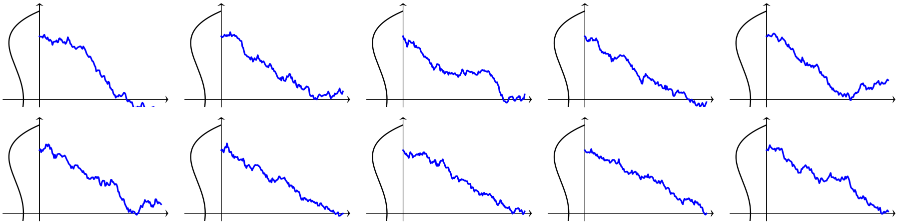


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

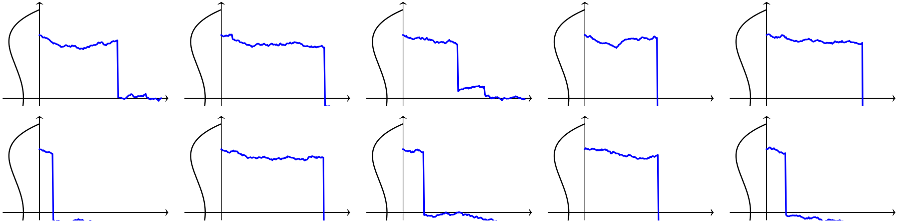


Catastrophe Principle Dictates SGD's Escape Route

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

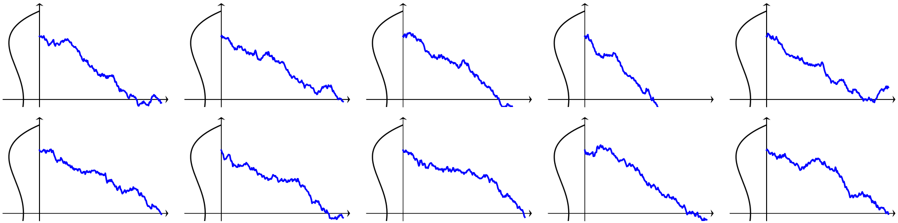


Trajectory of SGD X^{η} conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

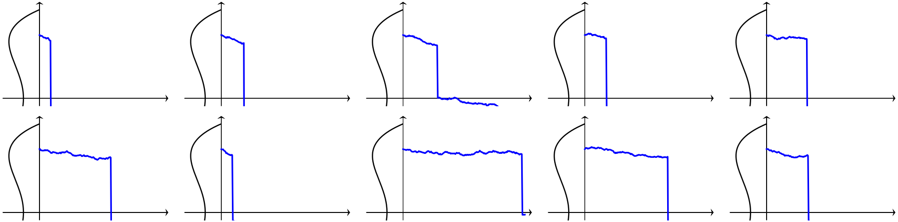


Catastrophe Principle Dictates SGD's Escape Route

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



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



Truncated Version of Stochastic Gradient Descent

SGD

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

Then, again,

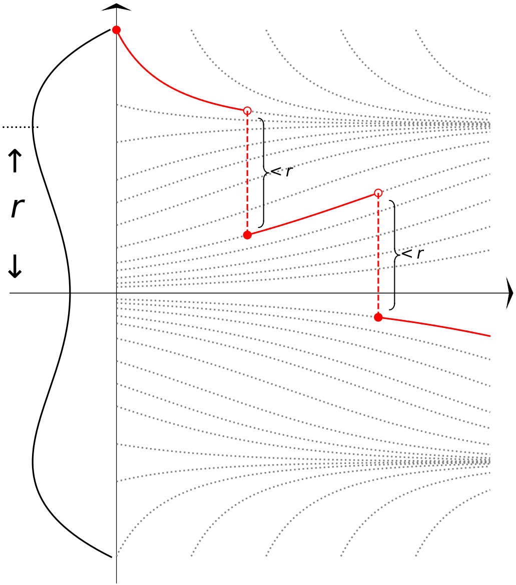
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

where

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

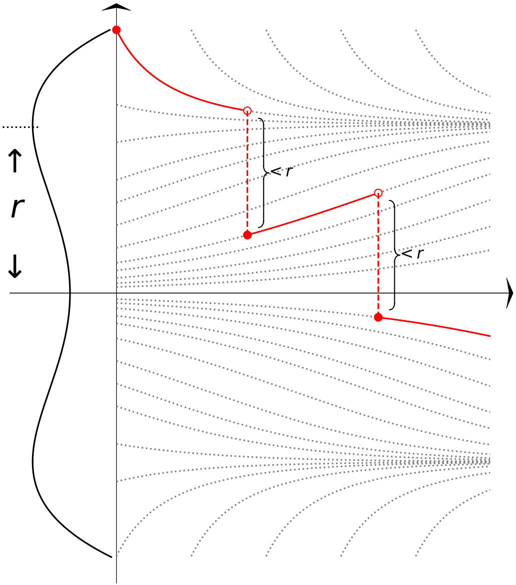
How does $\mathcal{J}(A)$ change?

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



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

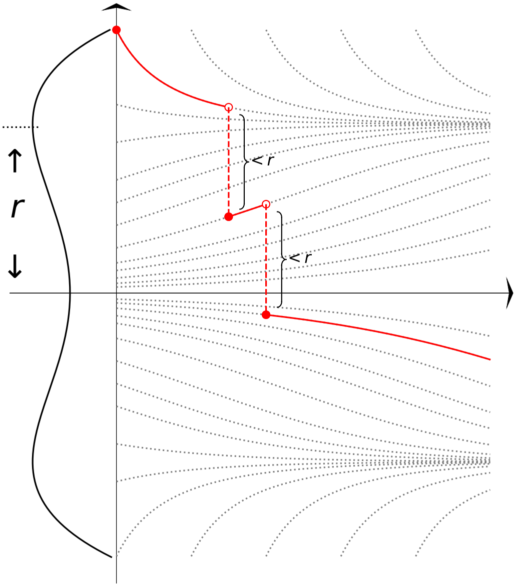
If $r \in (c, 2c)$



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

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

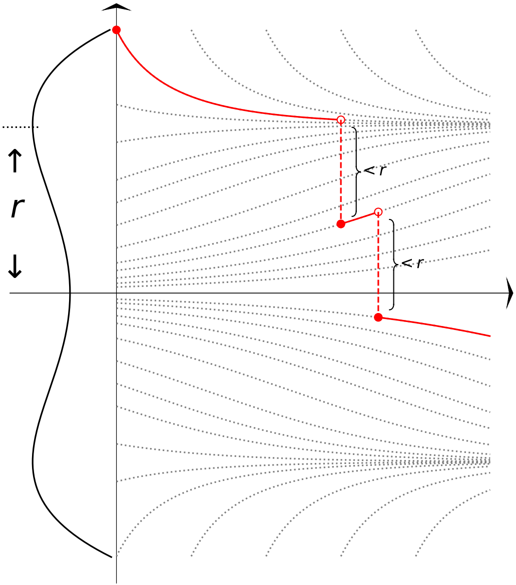
If $r \in (c, 2c)$



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

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

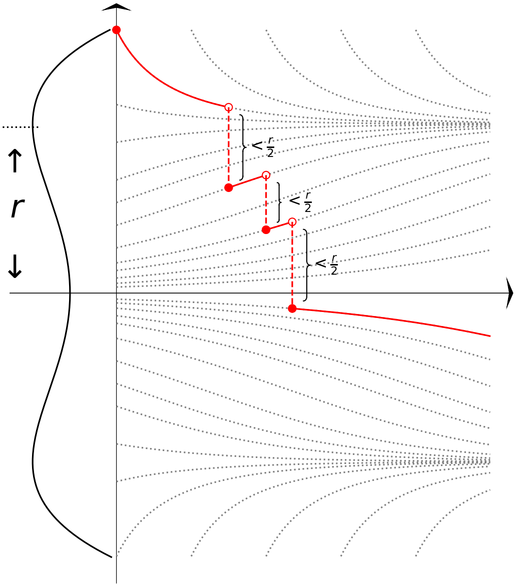
If $r \in (c, 2c)$



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

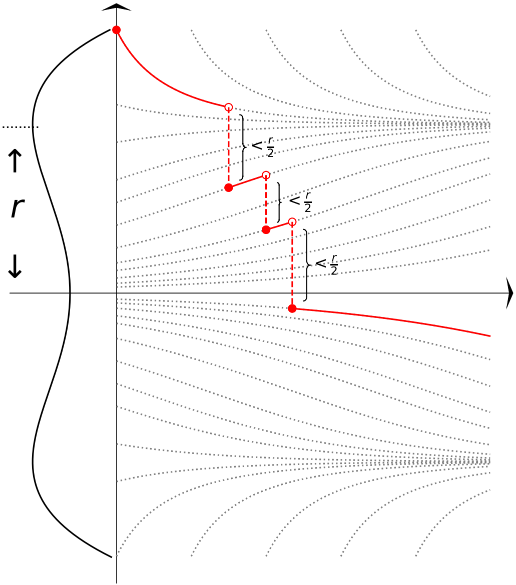
How does $\mathcal{J}(A)$ change?

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



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

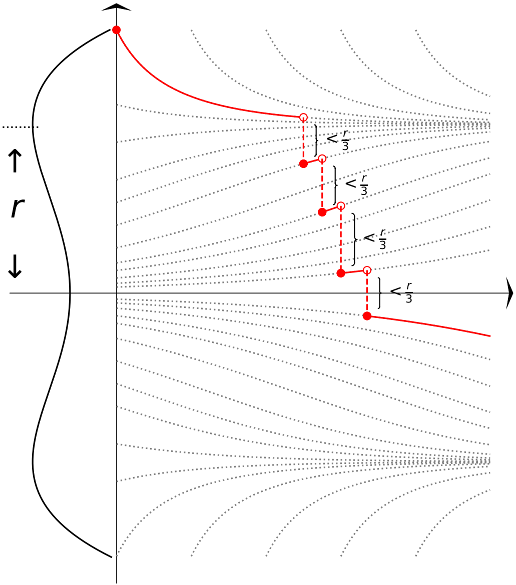
If $r \in (2c, 3c)$



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

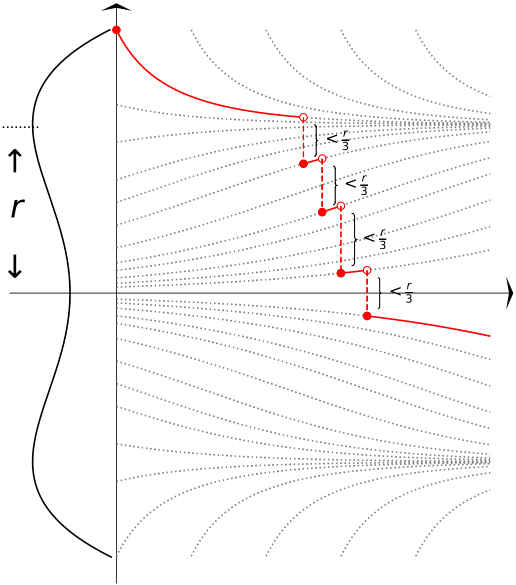
How does $\mathcal{J}(A)$ change?

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



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

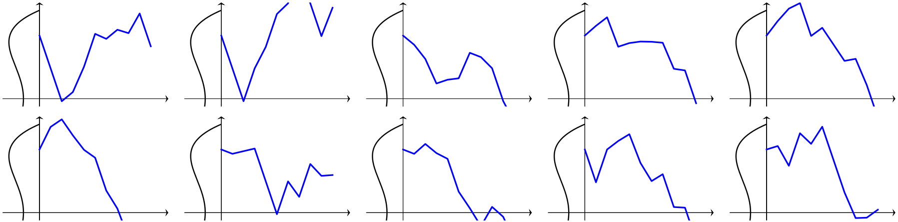
If $r \in (3c, 4c)$



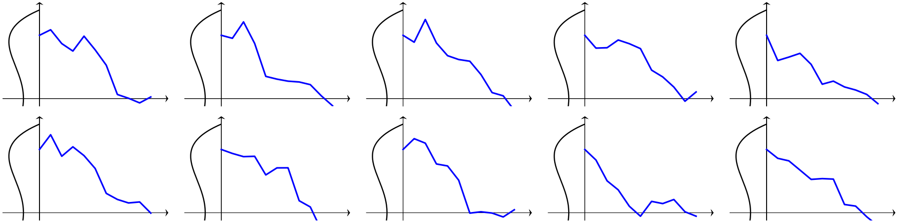
$\mathcal{J}(A) = 4$

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

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

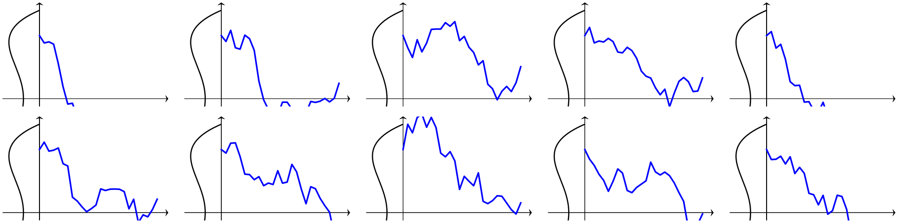


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

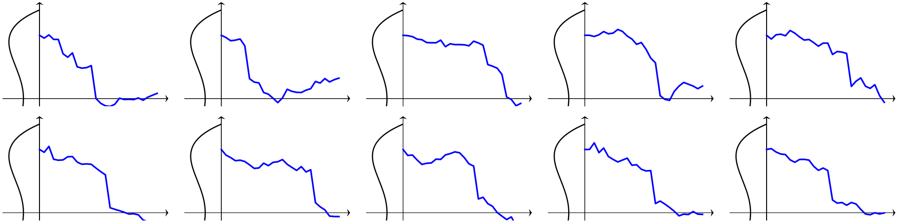


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

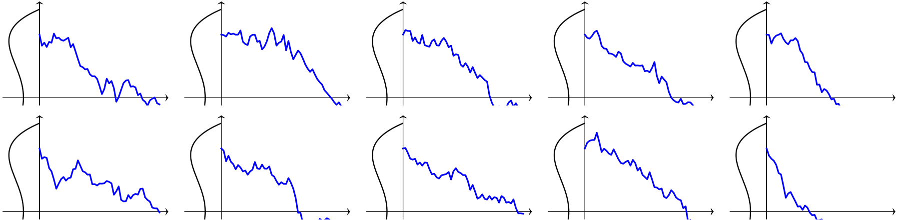


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

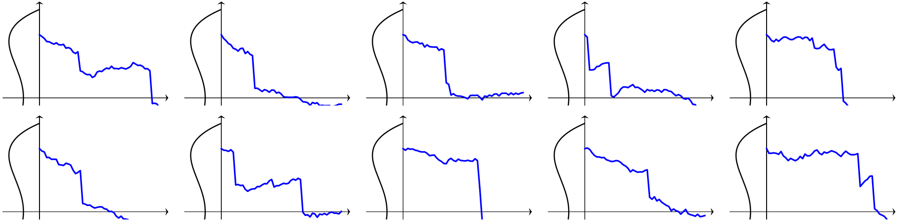


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

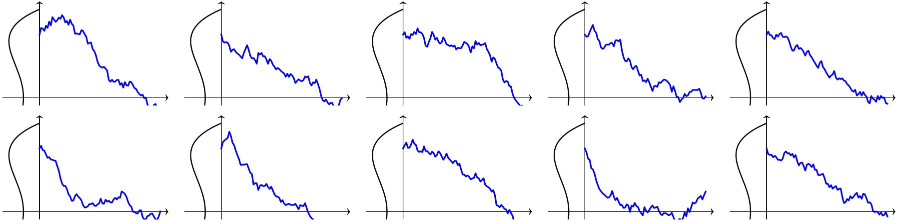


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

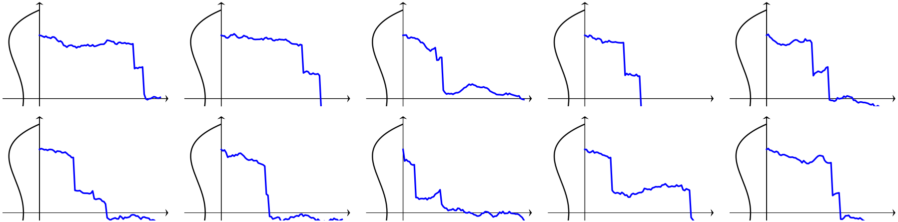


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

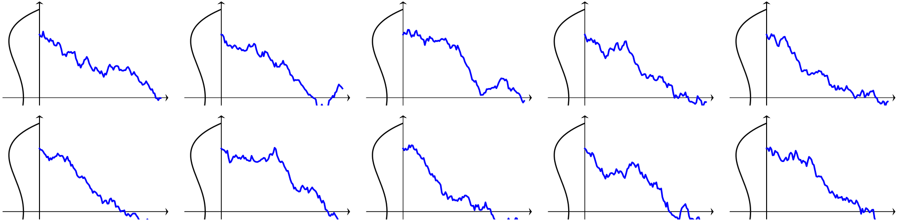


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

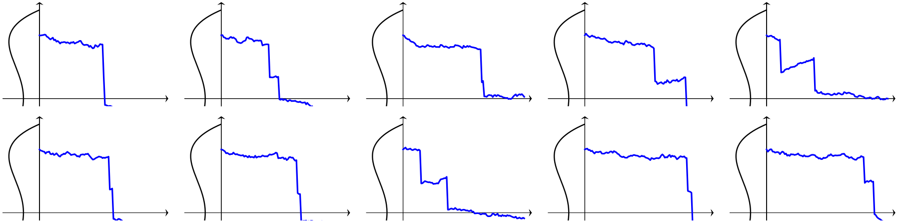


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

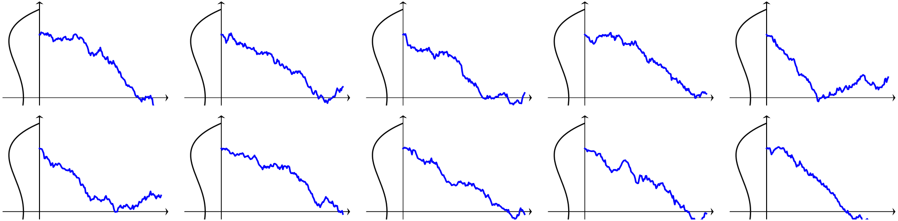


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

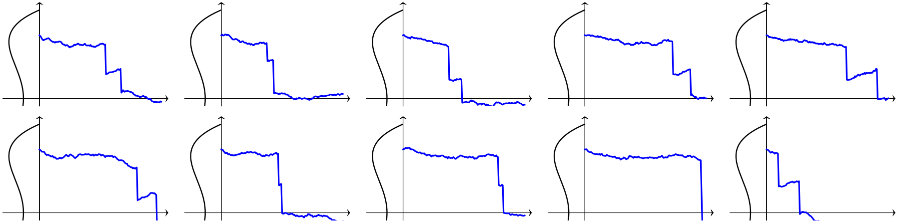


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

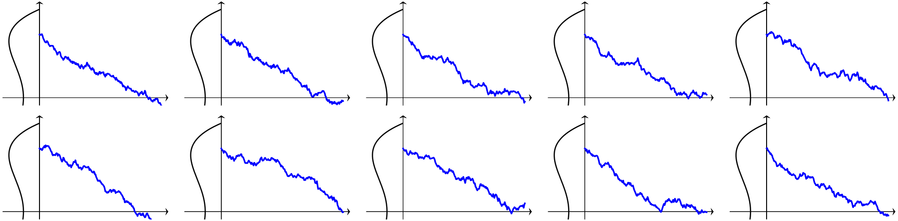


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

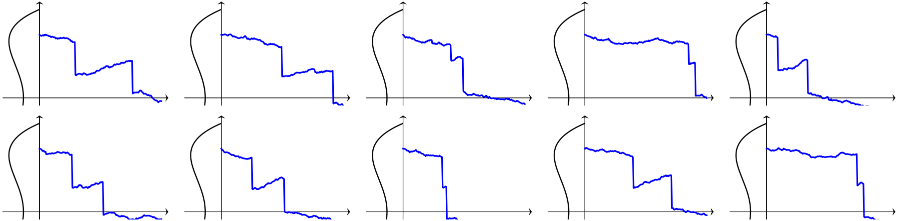


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

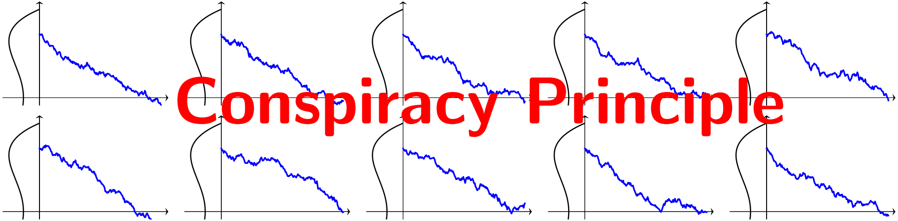


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



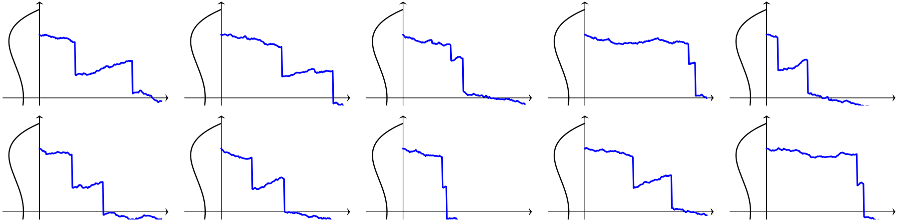
SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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



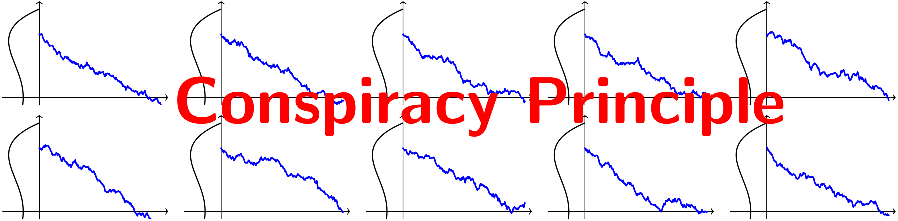
Conspiracy Principle

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



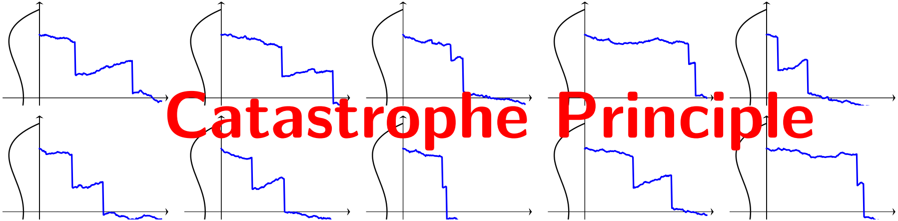
SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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



Conspiracy Principle

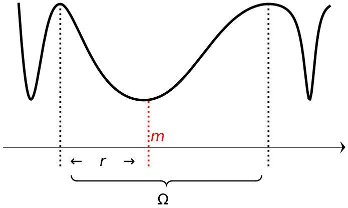
Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



Catastrophe Principle

Metastability of SGD

First Exit Time Analysis



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

gradient noise

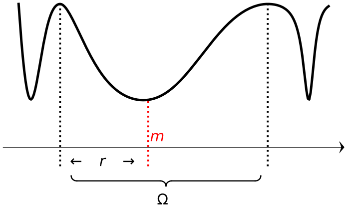
$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$

Theorem (Wang, Oh, R., 2022)

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

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

First Exit Time Analysis



gradient noise

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

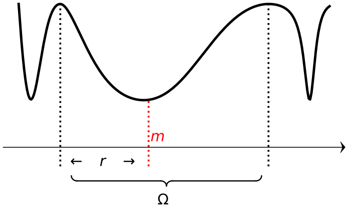
Theorem (Wang, Oh, R., 2022)

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

First Exit Time

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

First Exit Time Analysis



gradient noise

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

Theorem (Wang, Oh, R., 2022)

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

First Exit Time

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

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

Eliminating Sharp Local Minima with Truncated Heavy-Tails

gradient noise

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

Theorem (Wang, Oh, R., 2022)

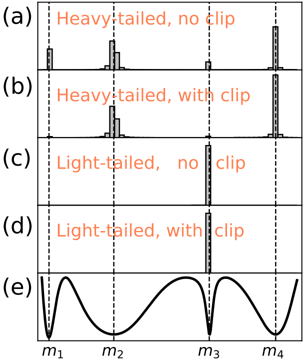
Under certain structural conditions, for any $t > 0$ and $\beta > 1 + \alpha \cdot l^*$,

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

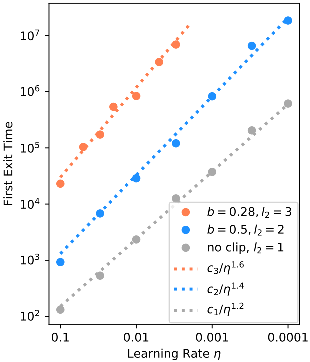
In fact, $W^\eta_{\lfloor t/\eta^{1+\alpha \cdot l^*} \rfloor}$ converges to a Markov jump processes

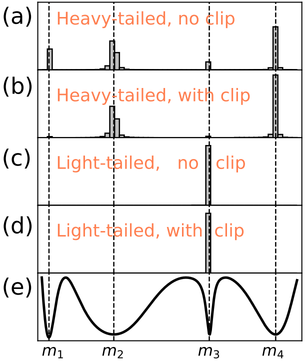
whose state space consists of wide local minima only.

Eliminating Sharp Local Minima with Truncated Heavy-Tails

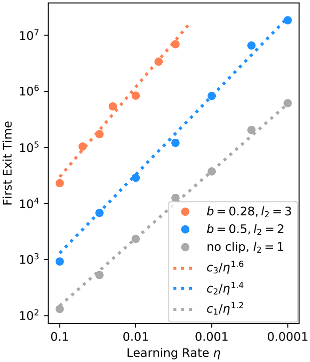


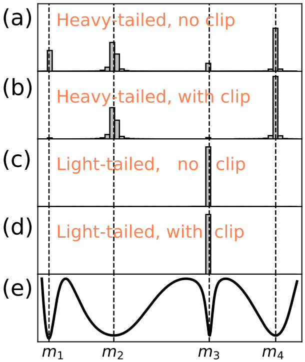
Eliminating Sharp Local Minima with Truncated Heavy-Tails

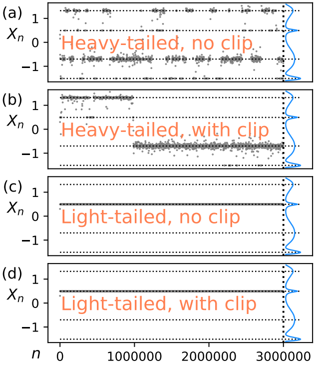




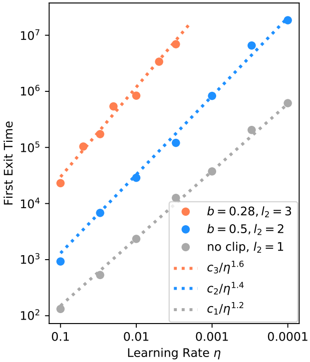
Eliminating Sharp Local Minima with Truncated Heavy-Tails

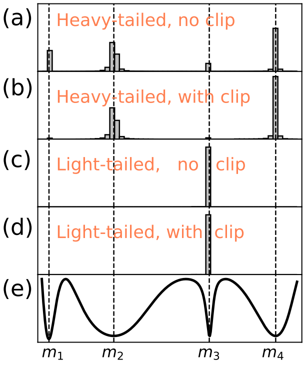


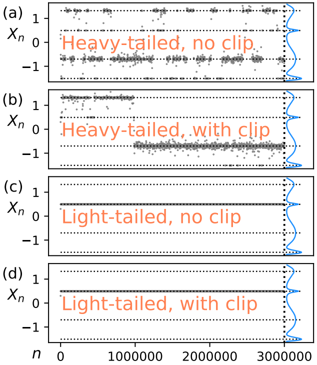




Eliminating Sharp Local Minima with Truncated Heavy-Tails

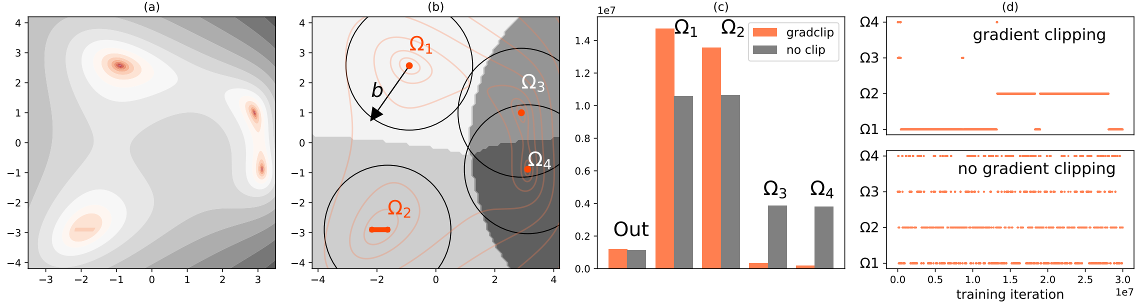




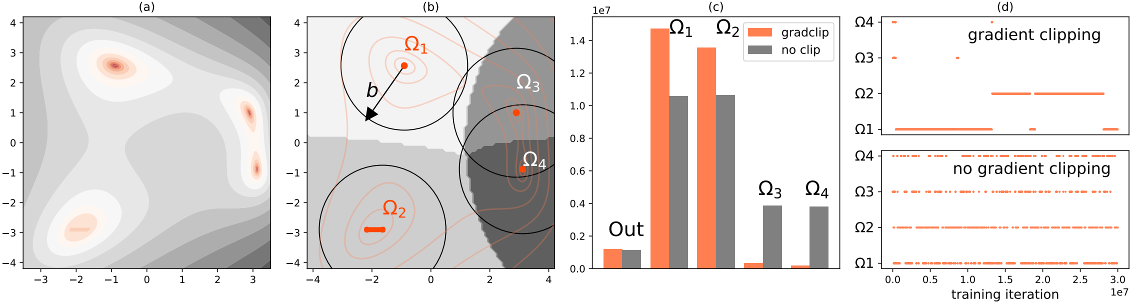


Consistent with what theory predicts!

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



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



Again, consistent with what theory predicts!

New Training Strategy: Tail-INflation-Truncation

Inflating Tail of Gradient Noise

$$Y = Y^{\text{Small Batch}}$$

Inflating Tail of Gradient Noise

Stochastic Gradient

$$\searrow$$

$$Y = Y^{\text{Small Batch}}$$

Inflating Tail of Gradient Noise

Stochastic Gradient

$$\searrow$$

$$Y = Y^{\text{Small Batch}} + R \cdot (Y^{\text{Large Batch}} - Y^{\text{Small Batch}})$$

Inflating Tail of Gradient Noise

Stochastic Gradient $\searrow$ Pareto RV $\swarrow$

$$Y = Y^{\text{Small Batch}} + R \cdot (Y^{\text{Large Batch}} - Y^{\text{Small Batch}})$$

Inflating Tail of Gradient Noise

Heavy-Tailed Stochastic Gradient

Pareto RV

$$Y = Y^{\text{Small Batch}} + R \cdot (Y^{\text{Large Batch}} - Y^{\text{Small Batch}})$$

Inflating Tail of Gradient Noise

Heavy-Tailed Stochastic Gradient

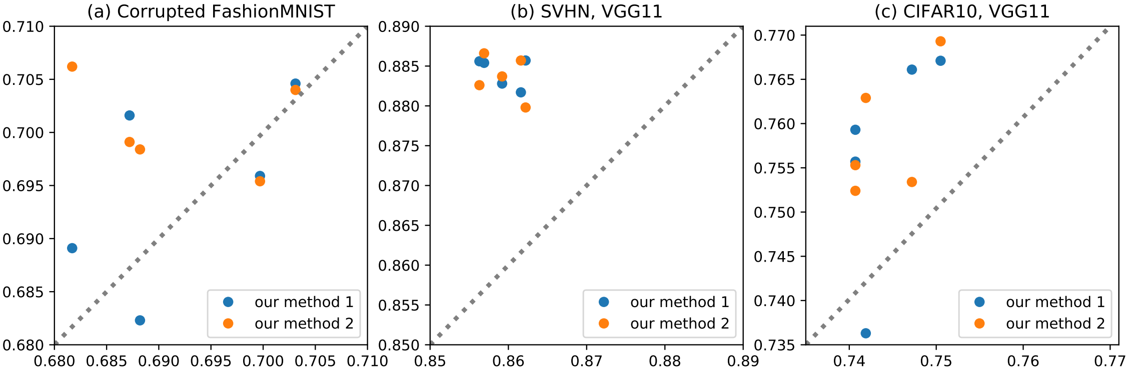
Pareto RV

$$Y = Y^{\text{Small Batch}} + R \cdot (Y^{\text{Large Batch}} - Y^{\text{Small Batch}})$$

Test accuracy	LB	SB	SB+Clip	SB+Noise	Our 1	Our 2
FashionMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**

Expected Sharpness	LB	SB	SB+Clip	SB+Noise	Our 1	Our 2
FashionMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

Improvement is Consistent



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

We will know in a few months:

- Moloco
  - top performing mobile ad tech company
  - 35B+ user impressions per month

- Launching a project to improve Moloco's conversion-rate prediction accuracy

Summary

- Rich mathematical structure in the global dynamics of heavy-tailed SGD
- Elimination of sharp local minima from SGD with truncated heavy-tailed gradient noise
- Tail-INflation-Truncation strategy improves SGD's generalization performance