Theoretical Framework for Global Dynamics and Metastability Analysis of Heavy-Tailed Systems

Chang-Han Rhee

Northwestern University

2023 INFORMS Annual Meeting, Phoenix

October 16, 2023

Based on the joint works with

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

How Rare Events Characterize Global Dynamics of Heavy-Tailed Systems

Chang-Han Rhee

Northwestern University

2023 INFORMS Annual Meeting, Phoenix

October 16, 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

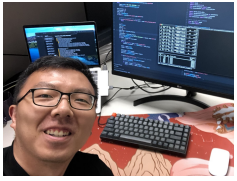
Jose Blanchet
Stanford

Bohan Chen
Munich Re

Sewoong Oh
U. of Washington

Zhe Su
Northwestern

Bert Zwart
CWI



Xingyu Wang
Northwestern

Team

Mihail Bazhba
U. of Amsterdam

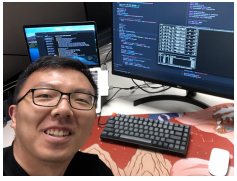
Jose Blanchet
Stanford

Bohan Chen
Munich Re

Sewoong Oh
U. of Washington

Zhe Su
Northwestern

Bert Zwart
CWI



Xingyu Wang
Northwestern

(2023 Nicholson Student Paper Prize, 2nd Place)

Generalization Mystery of Deep Learning

Empirical Success of Deep Neural Networks (DNNs)

"Deep Learning is eating the world."

- Jorge Nocedal

Empirical Success of Deep Neural Networks (DNNs)

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

Empirical Success of Deep Neural Networks (DNNs)

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

Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

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

Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

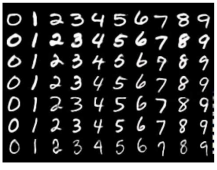
- Even with massive over-parametrization, DNNs generalize well.
- Training a DNN can be formulated as a non-convex loss optimization problem.
- **Algorithmic Regularazation**

Training Set

Test Set [image source]

Empirical Success of Deep Neural Networks (DNNs)

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



Training Set

Test Set image source

Empirical Success of Deep Neural Networks (DNNs)

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

Training Set

Test Set [image source]

Empirical Success of Deep Neural Networks (DNNs)

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

A Central Mystery of Deep Learning

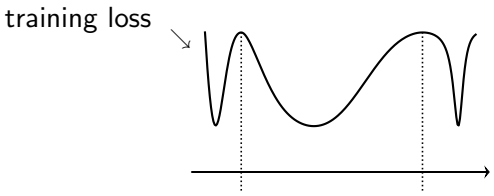
Heavy Tailed Large Deviations and Metastability

Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

Heavy Tailed Large Deviations and Metastability

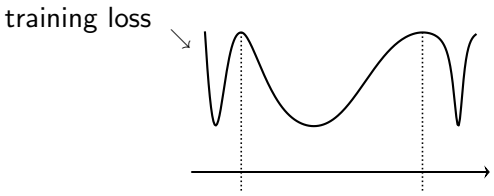
- Popular explanation: SGD somehow finds flat local minima.



Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

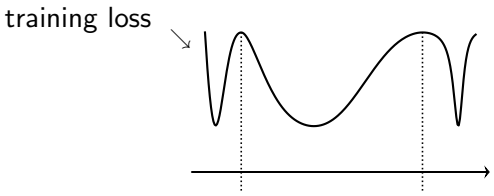


Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how?

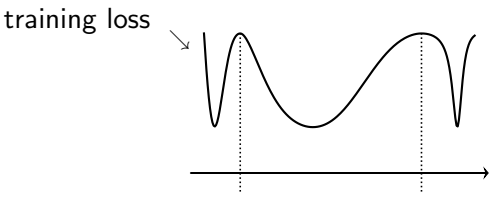


Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

- But how?



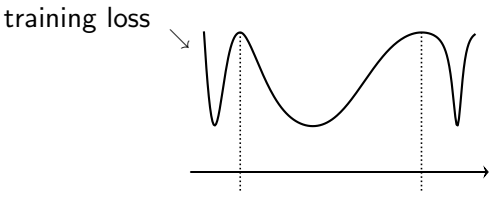
It takes more than 10^{400} time steps for SGD to escape from any of these.

Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

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



It takes more than 10^{400} time steps for SGD to escape from any of these.

Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

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

- Heavy-tails in SGD sheds a new light to the algorithmic regularization.

  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), Wang, Oh, Rhee (2022), etc

Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

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

- Heavy-tails in SGD sheds a new light to the algorithmic regularization.

  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), Wang, Oh, Rhee (2022), etc

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

Heavy Tailed Large Deviations and Metastability

- Popular explanation: SGD somehow finds flat local minima.

  ↖ tends to generalize well

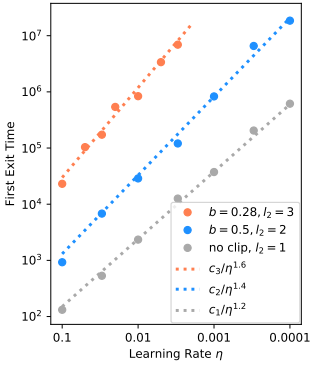
- But how? Previous theoretical attempts failed to explain how.

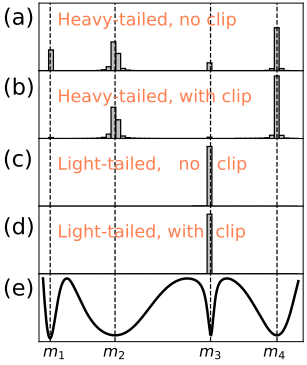
- Heavy-tails in SGD sheds a new light to the algorithmic regularization.

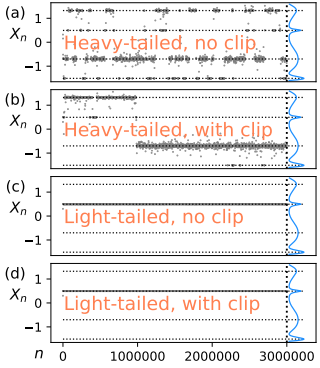
  eg. Simsekli et al. (2019), Hodgkinsons & Mahoney (2020), Wang, Oh, Rhee (2022), etc

Large deviations and metastability of SGD has fascinating connection this!

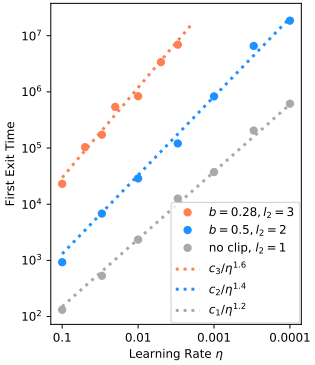
Entirely Different Global Dynamics Depending on Tail Behaviors

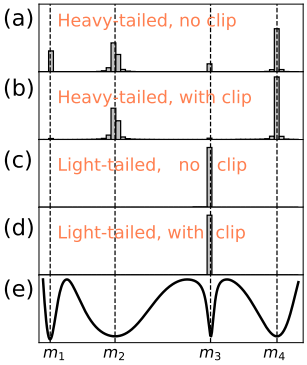


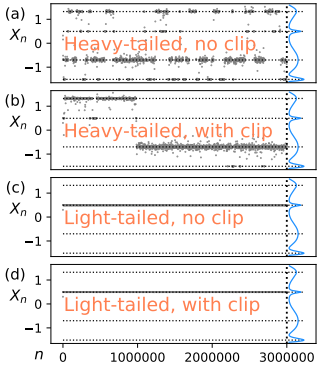




Entirely Different Global Dynamics Depending on Tail Behaviors







Explanation?

Heavy Tails and Catastrophe Principle

Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

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

Heavy-Tailed Distributions

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

Rare Events depend on "Tail Behaviors"

Light-Tailed Distributions

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

Heavy-Tailed Distributions

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

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

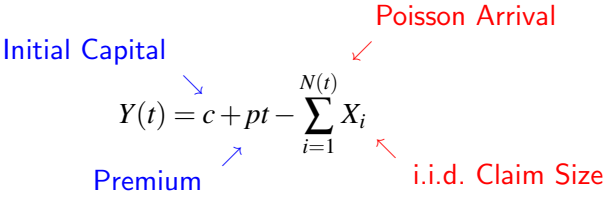
Initial Capital

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

Premium

i.i.d. Claim Size

Insurance Example: Capital Reserve



Insurance Example: Capital Reserve

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

Insurance Example: Capital Reserve

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

Insurance Example: Capital Reserve

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

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

Typical Scenario

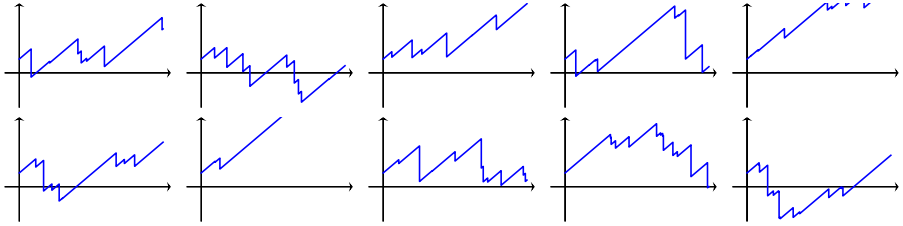
Typical Scenario

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

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

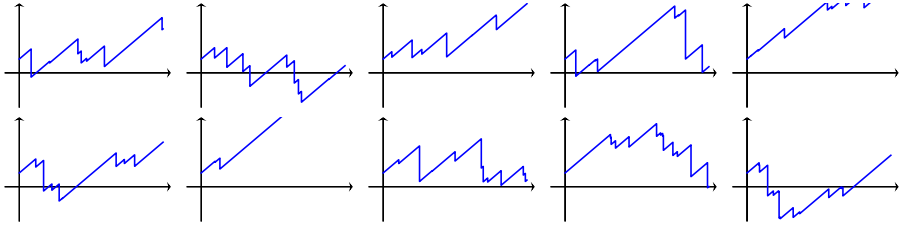
Typical Scenario

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



Typical Scenario

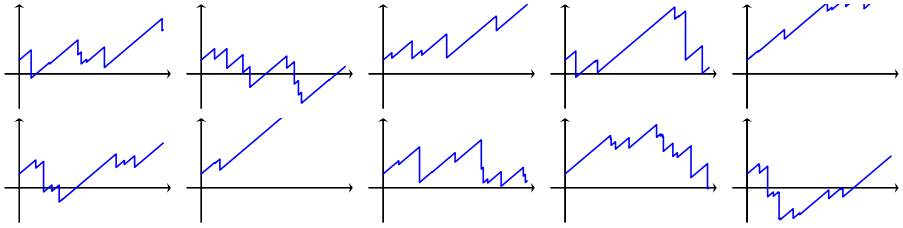
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **light-tailed**



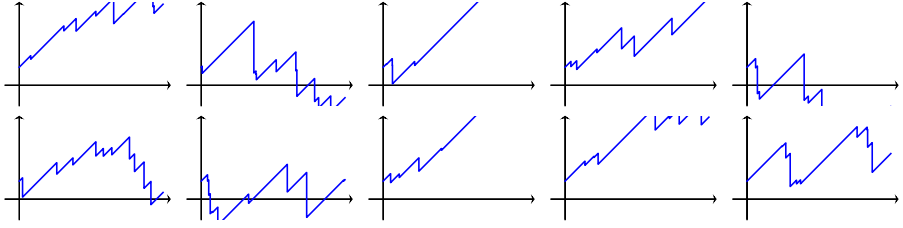
Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

Typical Scenario

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

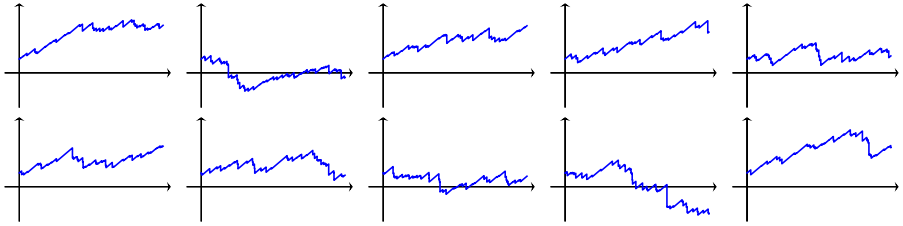


Sample paths of $\bar{Y}_n$: n=10 & claim sizes are **heavy-tailed**

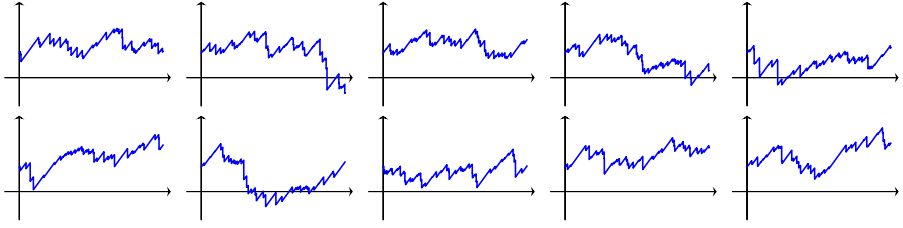


Typical Scenario

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

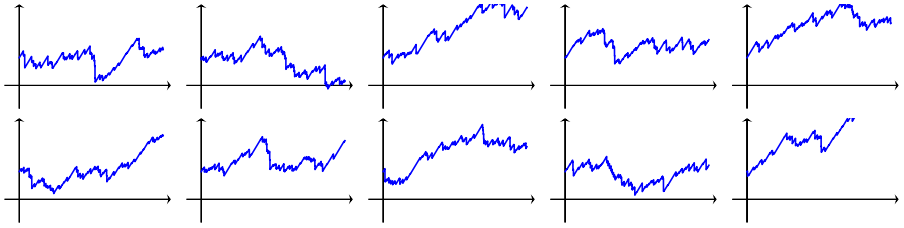


Sample paths of $\bar{Y}_n$: n=50 & claim sizes are **heavy-tailed**

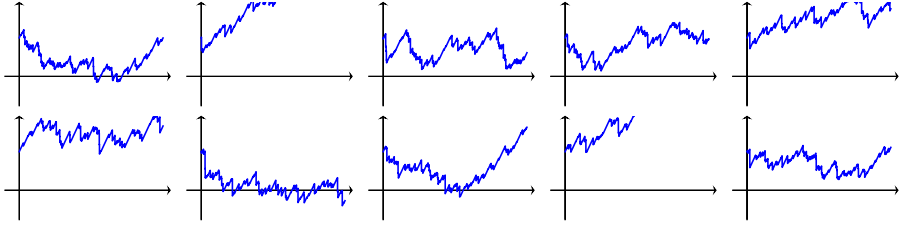


Typical Scenario

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

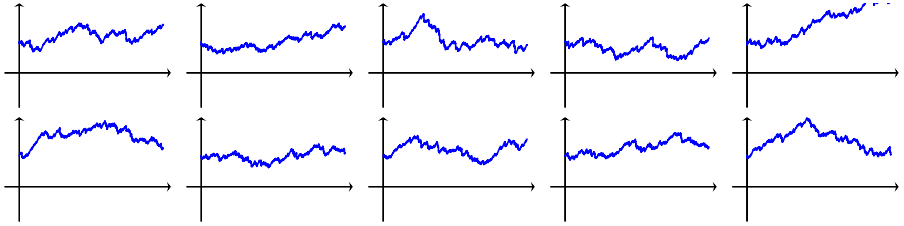


Sample paths of $\bar{Y}_n$: n=100 & claim sizes are **heavy-tailed**

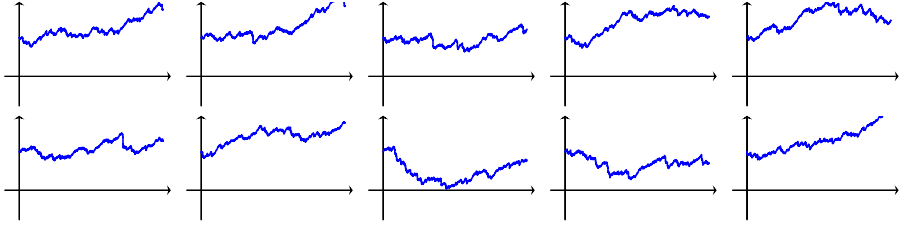


Typical Scenario

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

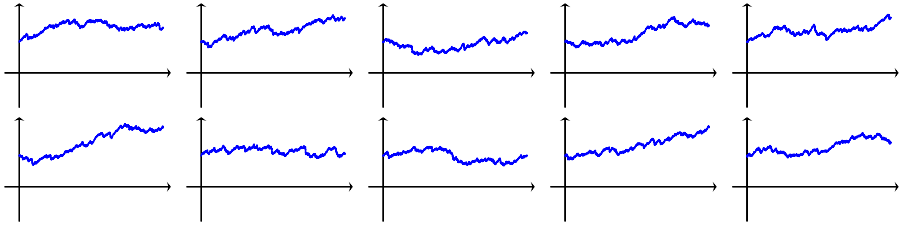


Sample paths of $\bar{Y}_n$: n=500 & claim sizes are **heavy-tailed**

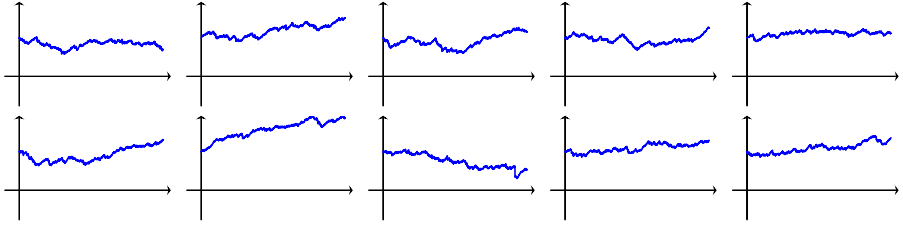


Typical Scenario

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

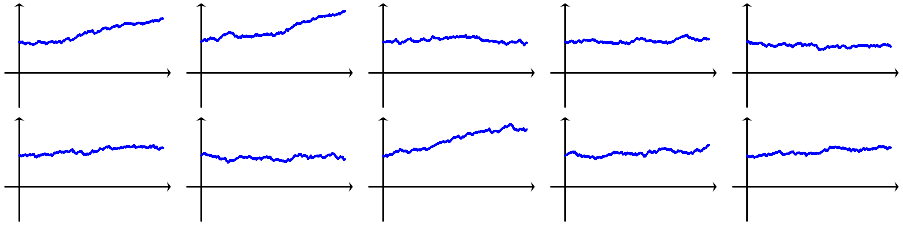


Sample paths of $\bar{Y}_n$: n=1000 & claim sizes are **heavy-tailed**

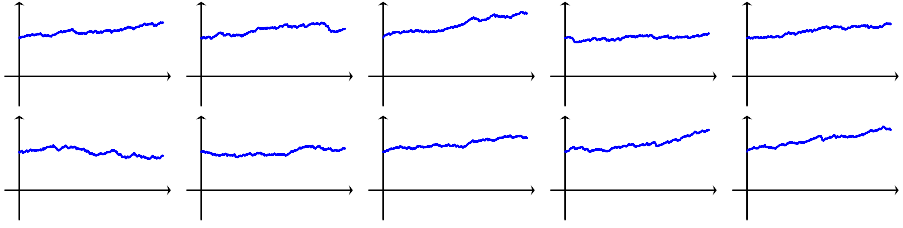


Typical Scenario

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

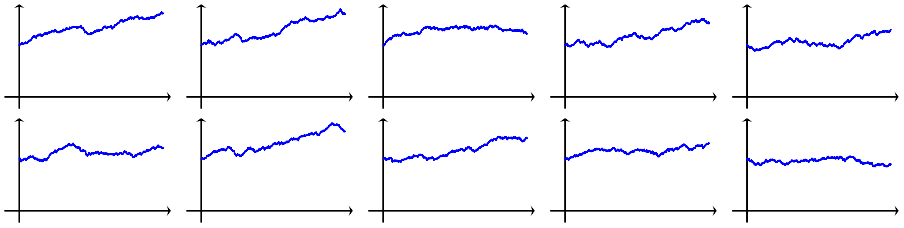


Sample paths of $\bar{Y}_n$: n=2500 & claim sizes are **heavy-tailed**

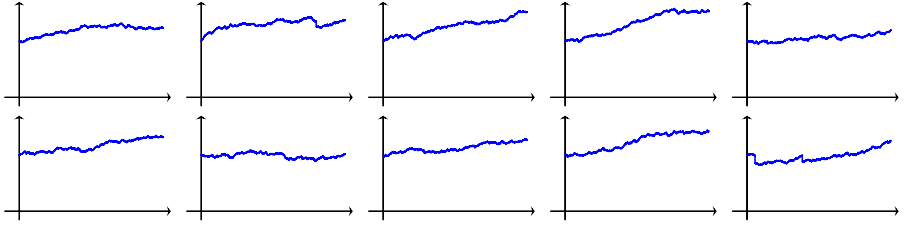


Typical Scenario

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



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

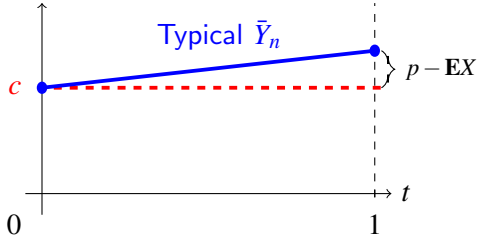


Typical Scenario

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

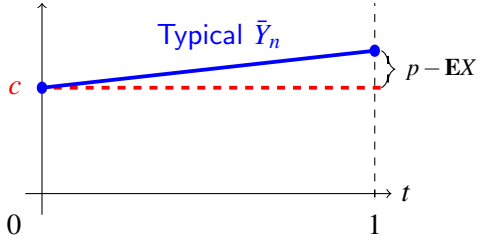
Typical Scenario

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



Typical Scenario

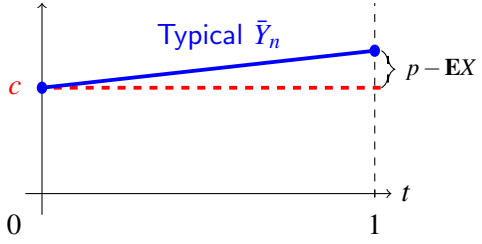
That is, $\bar{Y}_n(t) \approx c + (p - \mathbf{E}X)t$ for large n.



Typically, your insurance business will flourish

Typical Scenario

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

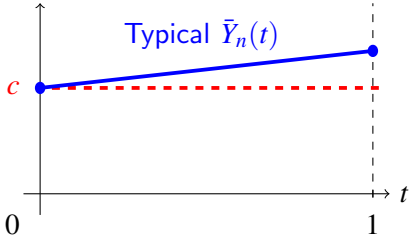


Typically, your insurance business will flourish

regardless of the tail distributions.

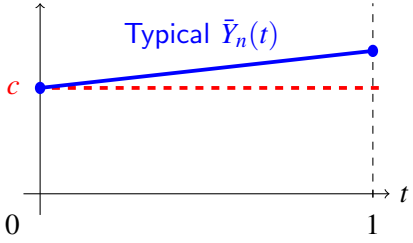
What about atypical cases?

A Rare Event: Bankruptcy



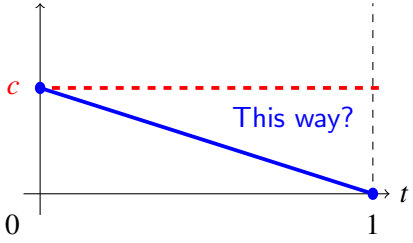
A Rare Event: Bankruptcy

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



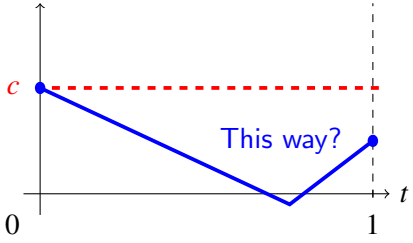
A Rare Event: Bankruptcy

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



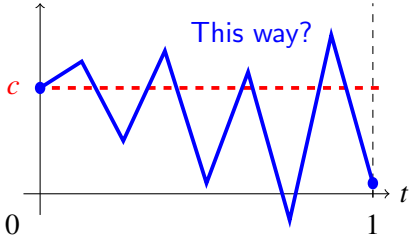
A Rare Event: Bankruptcy

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



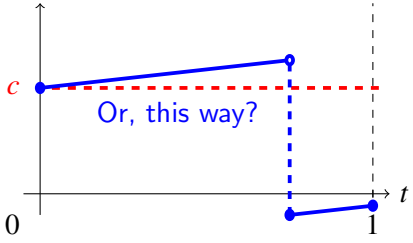
A Rare Event: Bankruptcy

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



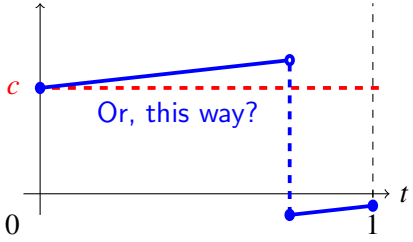
A Rare Event: Bankruptcy

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



A Rare Event: Bankruptcy

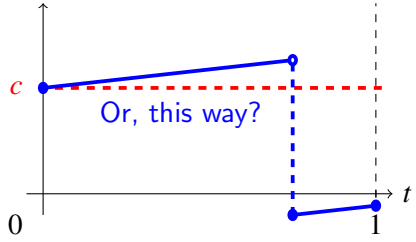
Consider $B \triangleq \{\ \bar{Y}_n$ falls below 0 on $[0,1]\ \}$. (i.e., Bankruptcy)



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)

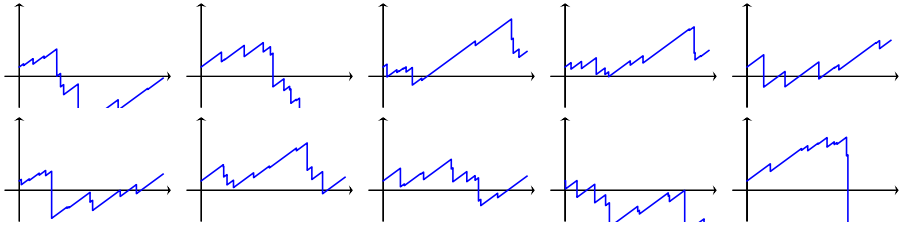


Are we going to see clear patterns?

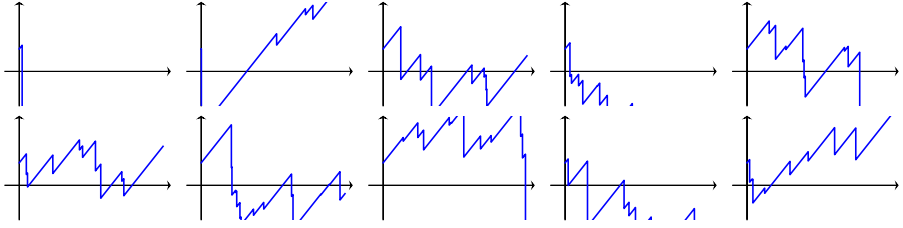
Do they depend on the tail distributions?

A Rare Event: Bankruptcy

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



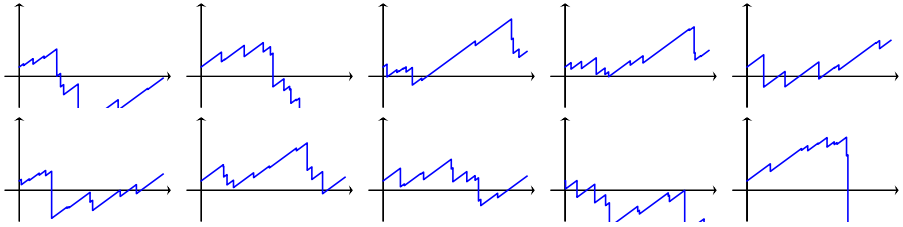
Sample paths of $\bar{Y}_{10}$ conditional on B for **heavy-tailed** claims:



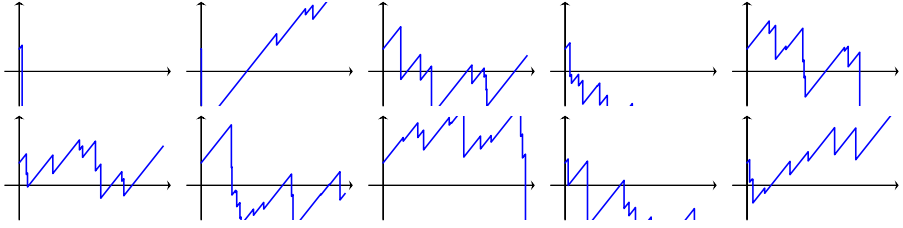
A Rare Event: Bankruptcy

Bankruptcy

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



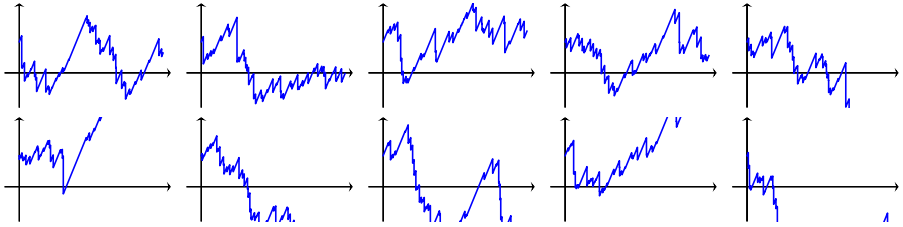
Sample paths of $\bar{Y}_{10}$ conditional on B for **heavy-tailed** claims:



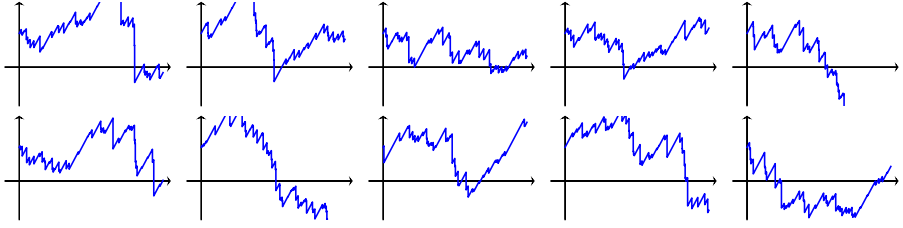
A Rare Event: Bankruptcy

Bankruptcy

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



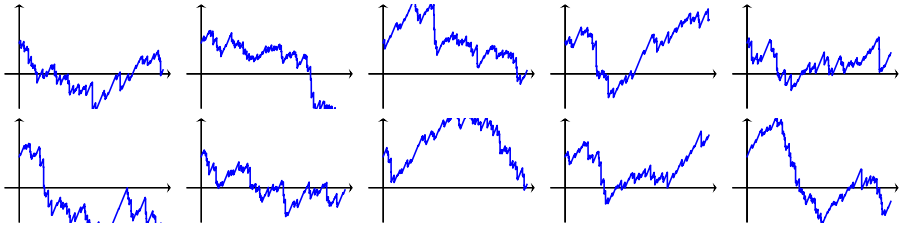
Sample paths of $\bar{Y}_{50}$ conditional on B for **heavy-tailed** claims:



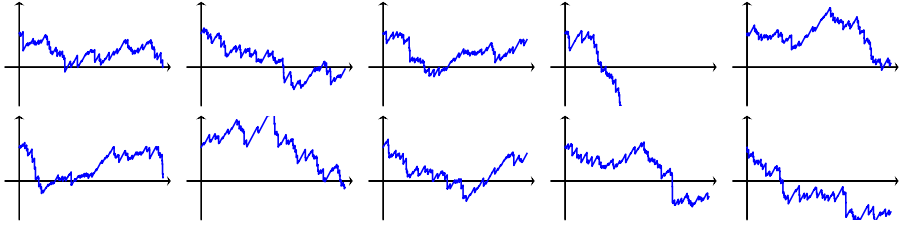
A Rare Event: Bankruptcy

Bankruptcy

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



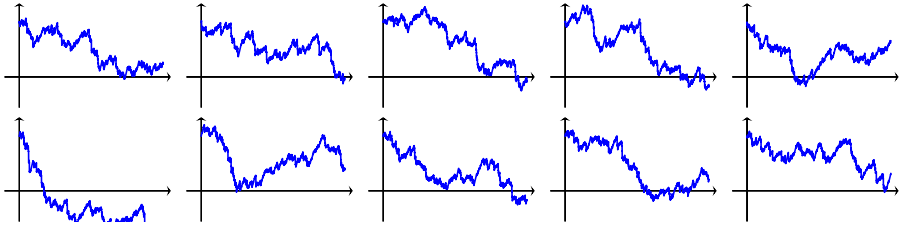
Sample paths of $\bar{Y}_{100}$ conditional on B for **heavy-tailed** claims:



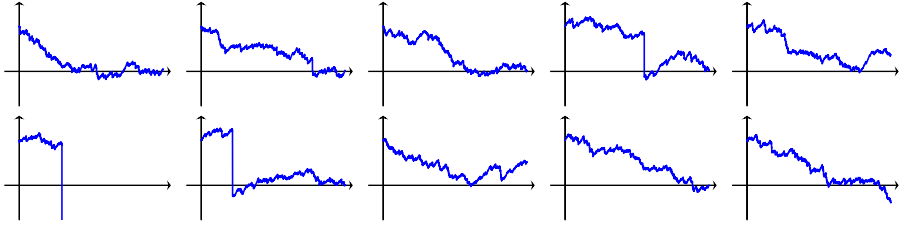
A Rare Event: Bankruptcy

Bankruptcy

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



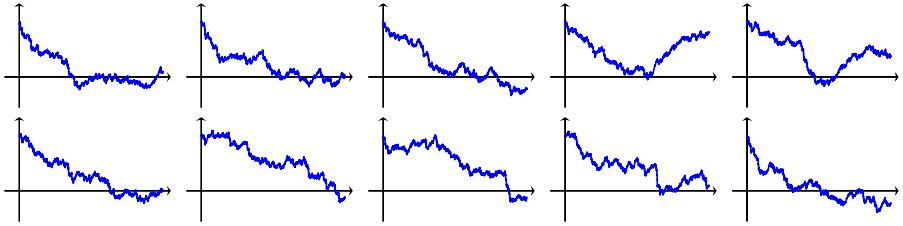
Sample paths of $\bar{Y}_{500}$ conditional on B for **heavy-tailed** claims:



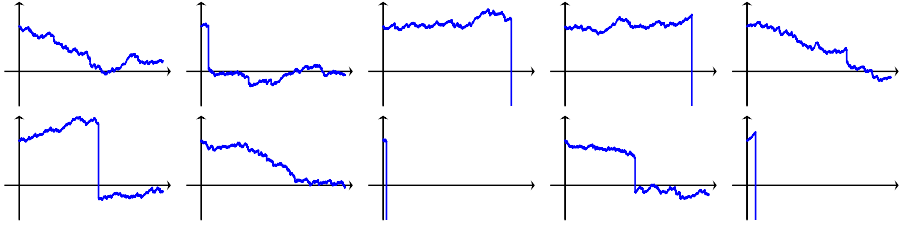
A Rare Event: Bankruptcy

Bankruptcy

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



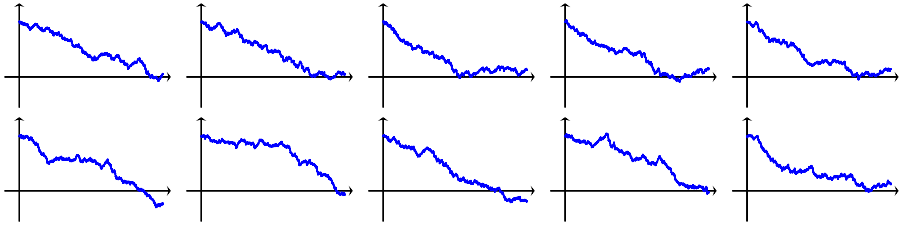
Sample paths of $\bar{Y}_{1000}$ conditional on B for **heavy-tailed** claims:



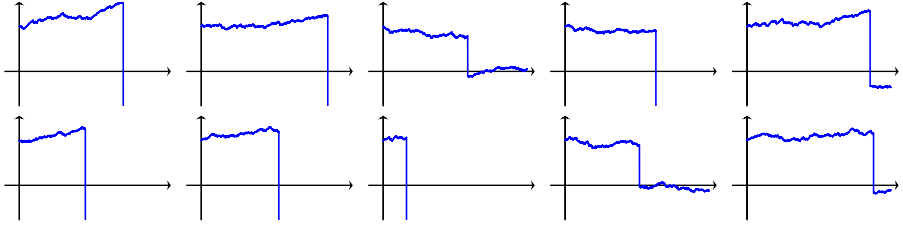
A Rare Event: Bankruptcy

Bankruptcy

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



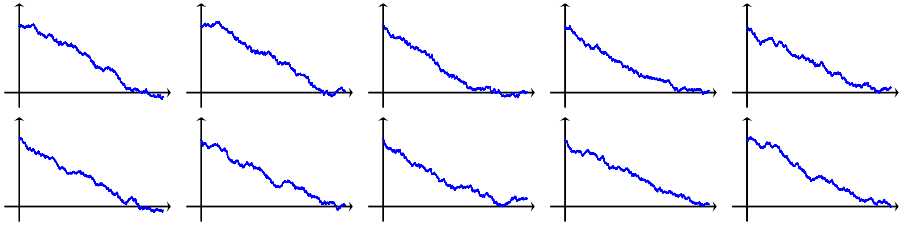
Sample paths of $\bar{Y}_{2500}$ conditional on B for **heavy-tailed** claims:



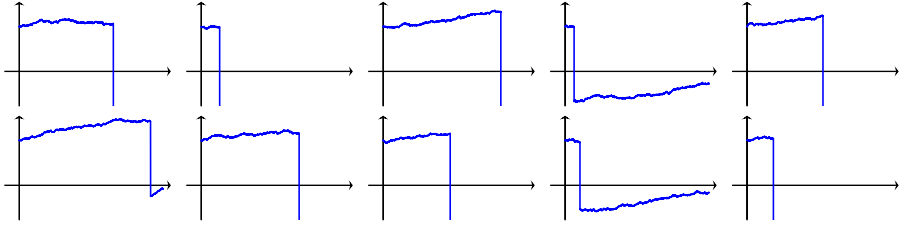
A Rare Event: Bankruptcy

Bankruptcy

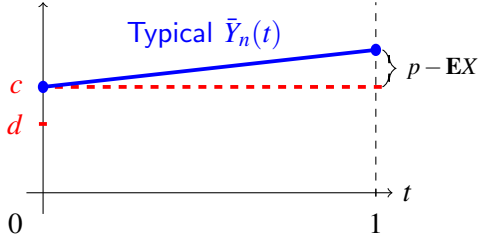
Sample paths of $\bar{Y}_{5000}$ conditional on B for **light-tailed** claims:



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

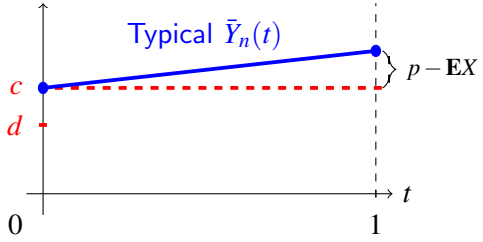


Bankruptcy Despite of Reinsurance



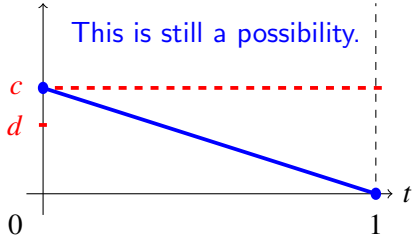
Bankruptcy Despite of Reinsurance

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



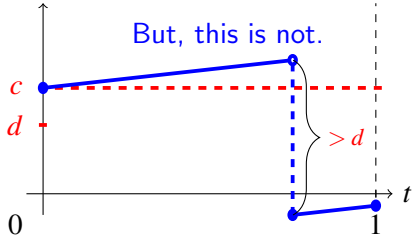
Bankruptcy Despite of Reinsurance

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



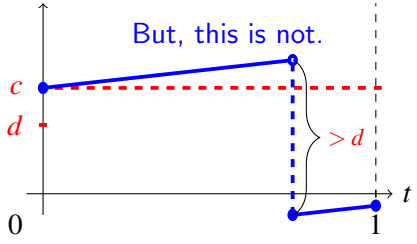
Bankruptcy Despite of Reinsurance

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



Bankruptcy Despite of Reinsurance

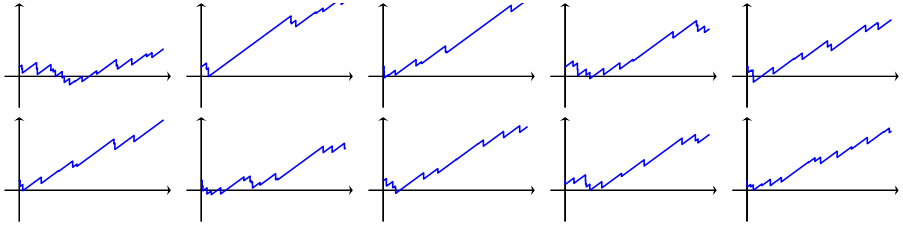
Consider $R \triangleq \{\ \bar{Y}_n$ falls below 0 on [0,1], jump sizes $\leq d\ \}$



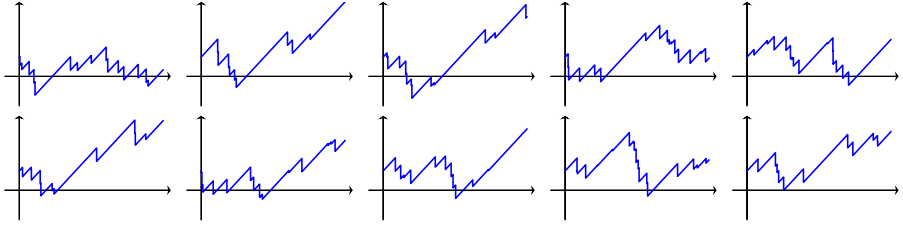
How does the pattern change in this case?

Bankruptcy Despite of Reinsurance

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

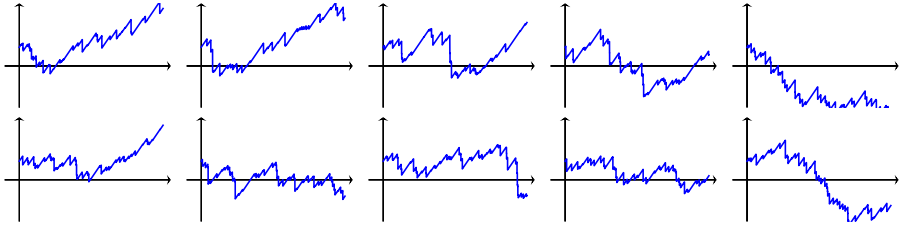


Sample paths of $\bar{Y}_{10}$ conditional on R for **heavy-tailed** claims:

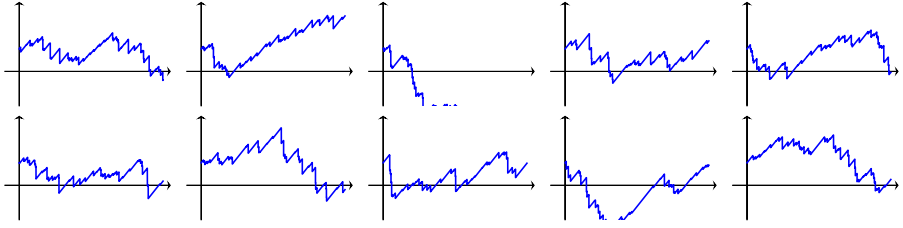


Bankruptcy Despite of Reinsurance

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

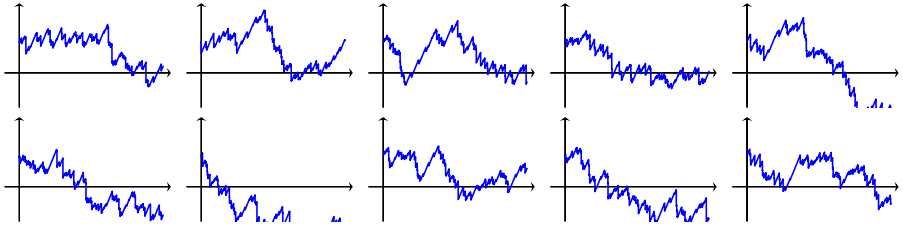


Sample paths of $\bar{Y}_{50}$ conditional on R for **heavy-tailed** claims:

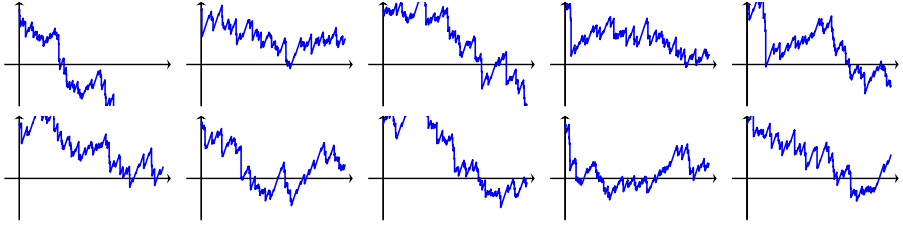


Bankruptcy Despite of Reinsurance

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

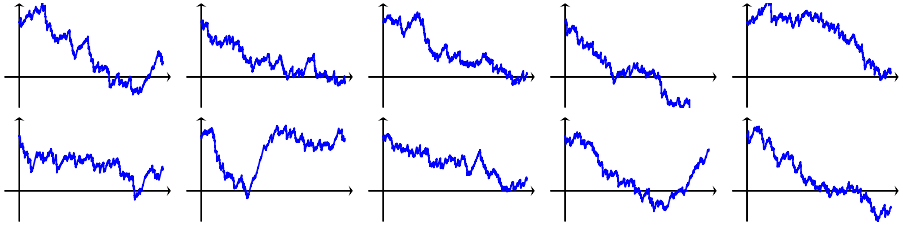


Sample paths of $\bar{Y}_{100}$ conditional on R for **heavy-tailed** claims:

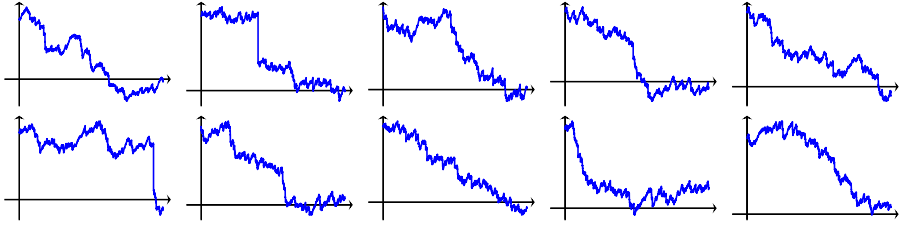


Bankruptcy Despite of Reinsurance

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

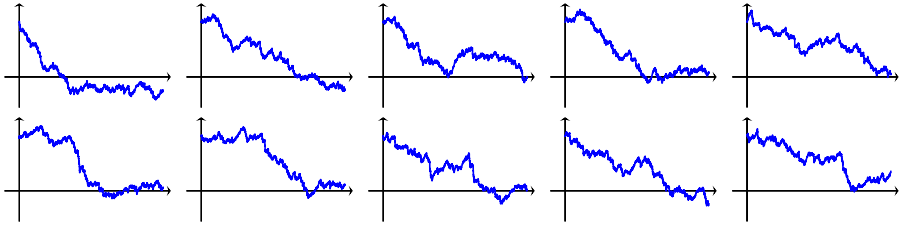


Sample paths of $\bar{Y}_{500}$ conditional on R for **heavy-tailed** claims:

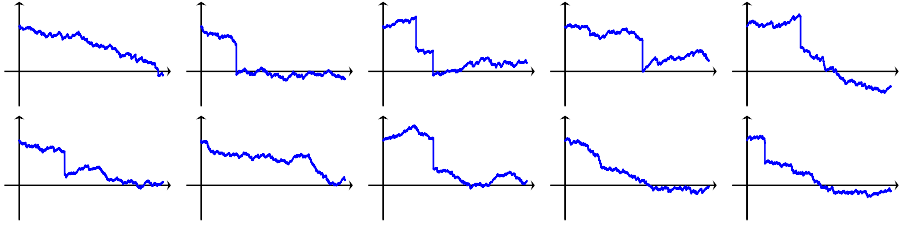


Bankruptcy Despite of Reinsurance

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

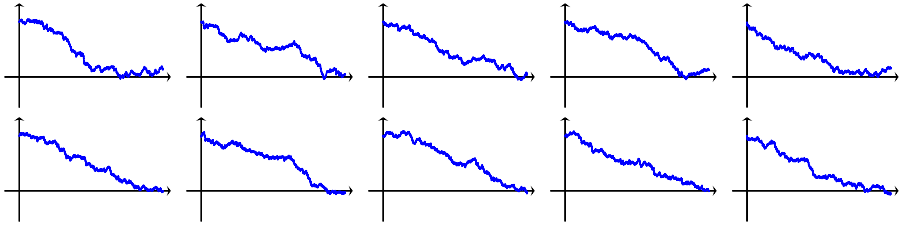


Sample paths of $\bar{Y}_{1000}$ conditional on R for **heavy-tailed** claims:

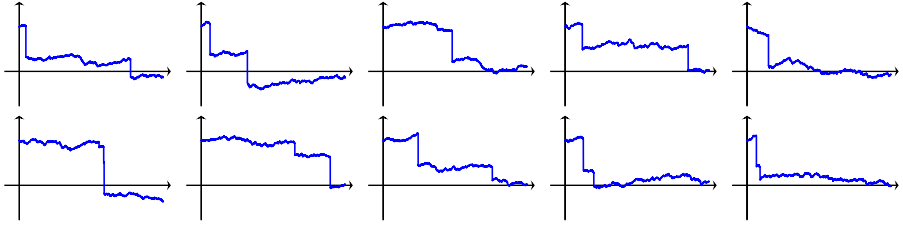


Bankruptcy Despite of Reinsurance

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

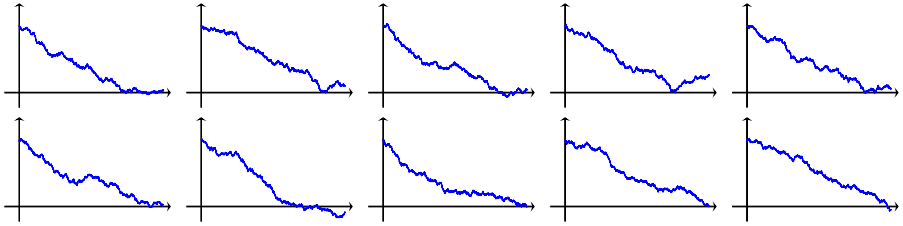


Sample paths of $\bar{Y}_{2500}$ conditional on R for **heavy-tailed** claims:

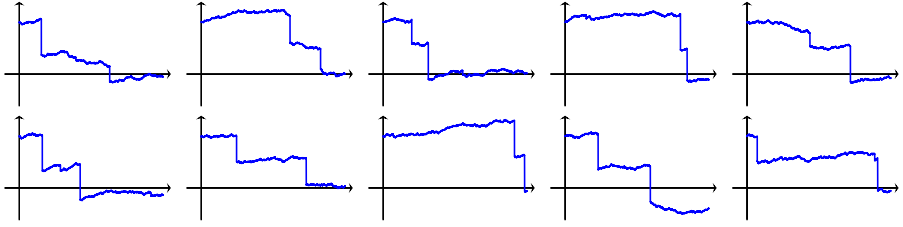


Bankruptcy Despite of Reinsurance

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

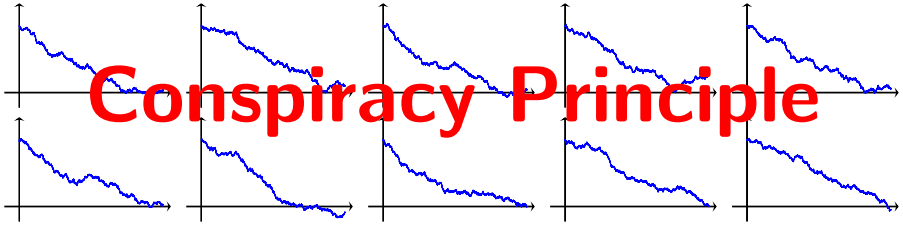


Sample paths of $\bar{Y}_{5000}$ conditional on R for **heavy-tailed** claims:

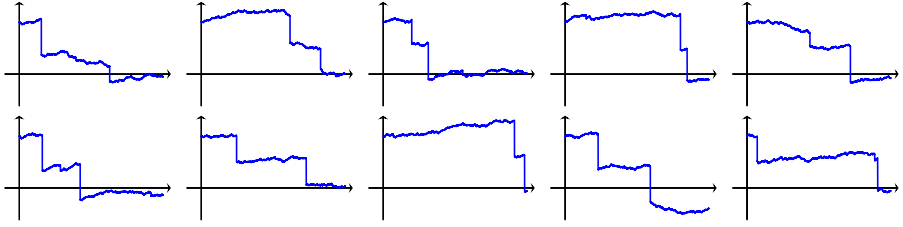


Bankruptcy Despite of Reinsurance

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

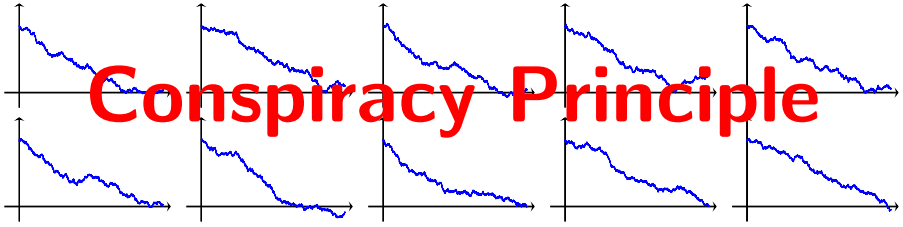


Sample paths of $\bar{Y}_{5000}$ conditional on R for **heavy-tailed** claims:

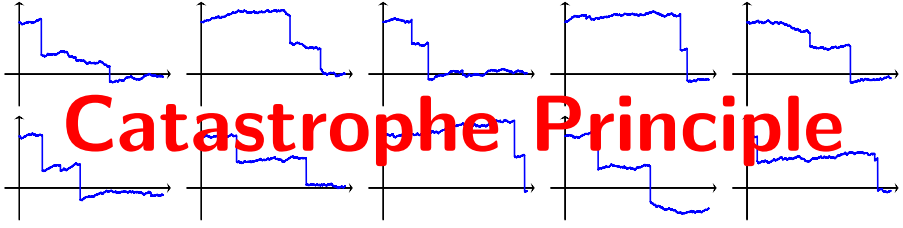


Bankruptcy Despite of Reinsurance

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

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

Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

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

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

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

Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

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

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

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

Heavy-Tailed Large Deviations

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

Theorem (R., Blanchet, Zwart, 2019)

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

LD power index

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

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

Implication: The Catastrophe Principle

Under certain regularity conditions on A,

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

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

Implication: The Catastrophe Principle

Under certain regularity conditions on A,

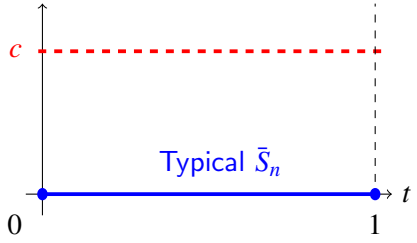
$$\mathcal{L}(\bar{S}_n | \bar{S}_n \in A) \to \mathcal{L}(\bar{S}_{|A})$$

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

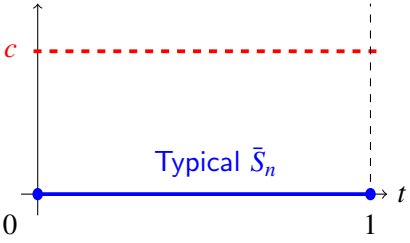
Rigorous Characterization of Catastrophe Principle

Conspiracy vs Catastrophe

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



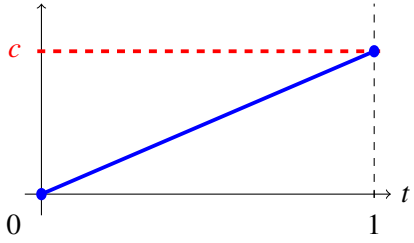
Light-Tailed Claim Size



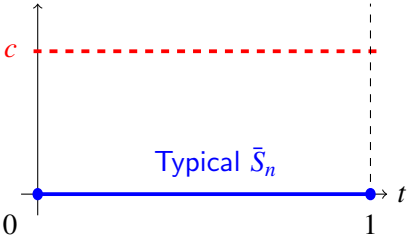
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



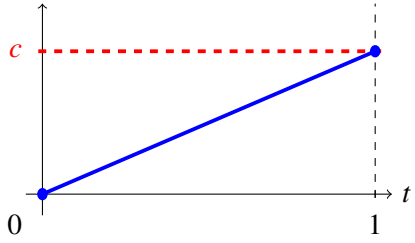
Light-Tailed Claim Size



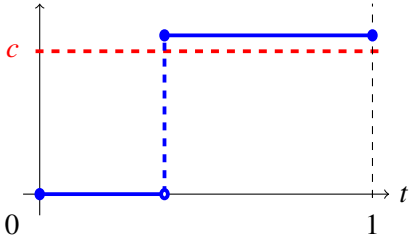
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



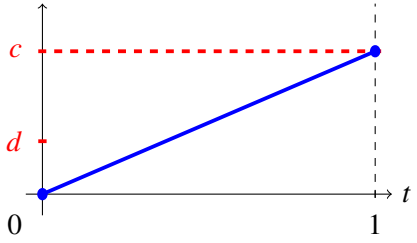
Light-Tailed Claim Size



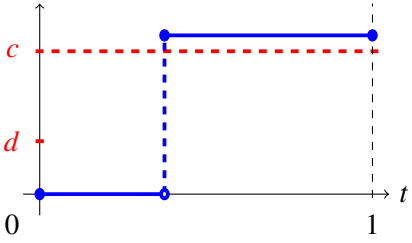
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



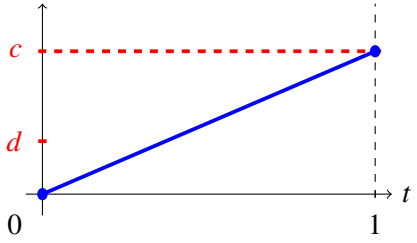
Light-Tailed Claim Size



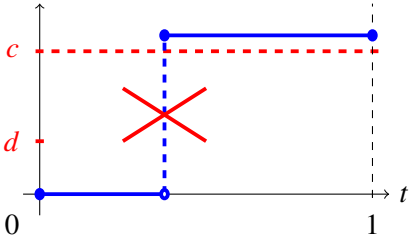
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



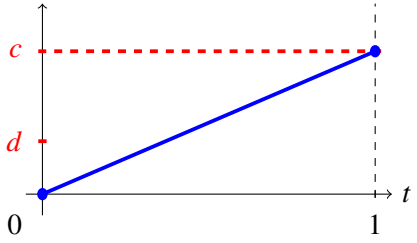
Light-Tailed Claim Size



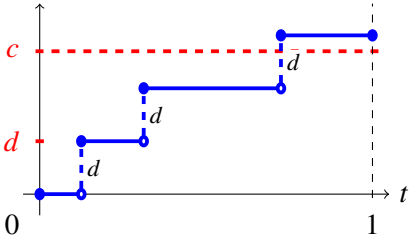
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



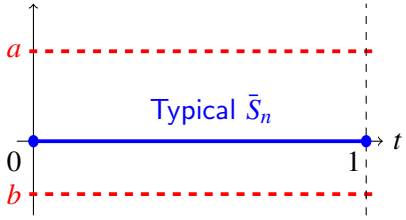
Light-Tailed Claim Size



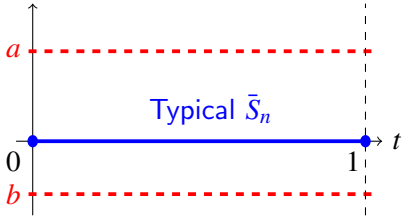
Heavy-Tailed Claim Size

Conspiracy vs Catastrophe

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



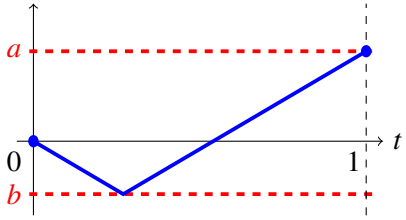
Light-Tailed Increments



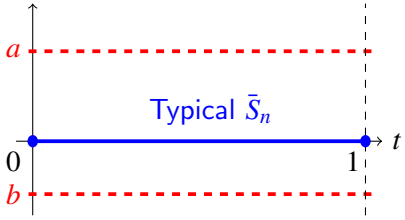
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



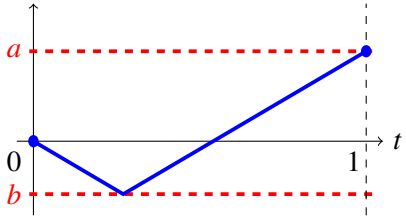
Light-Tailed Increments



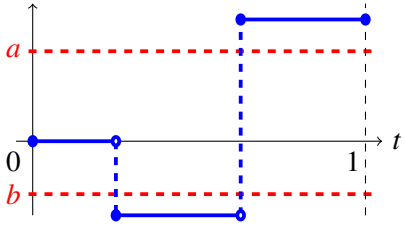
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



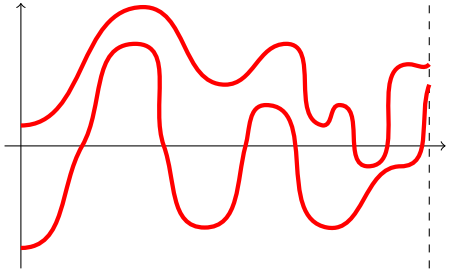
Light-Tailed Increments



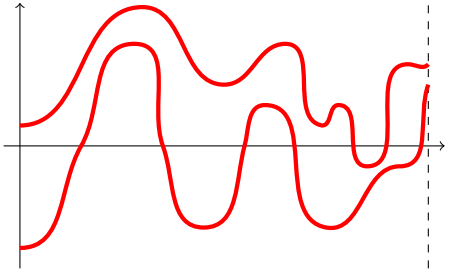
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



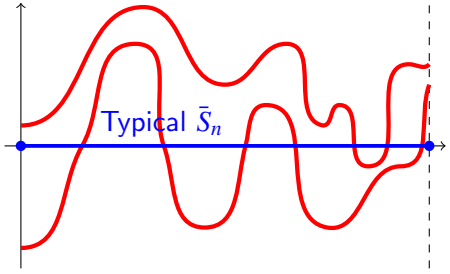
Light-Tailed Increments



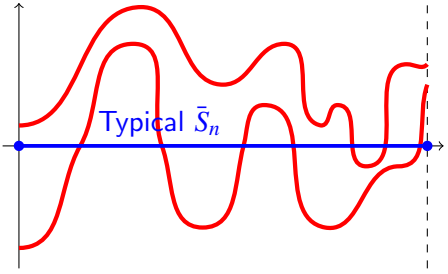
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



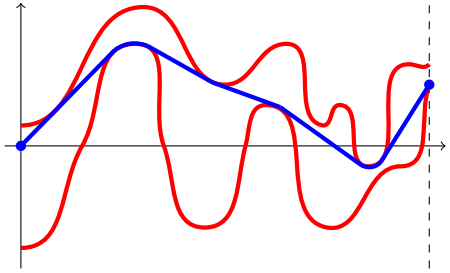
Light-Tailed Increments



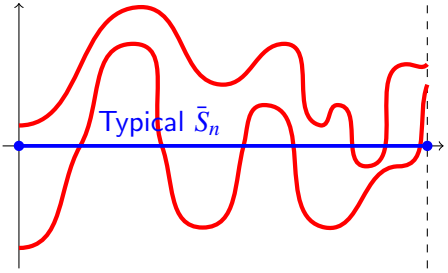
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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



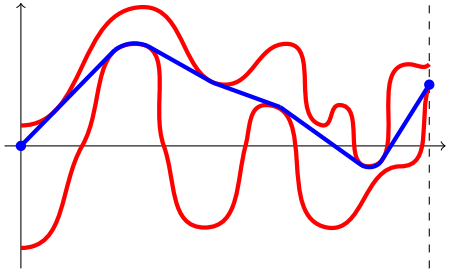
Light-Tailed Increments



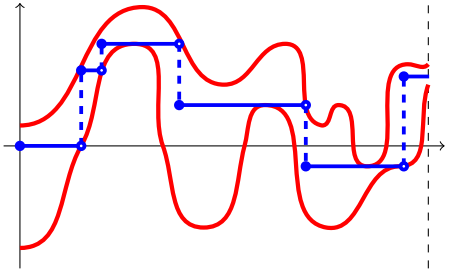
Heavy-Tailed Increments

Conspiracy vs Catastrophe

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

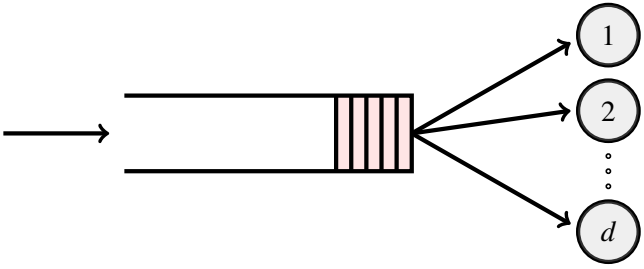


Light-Tailed Increments

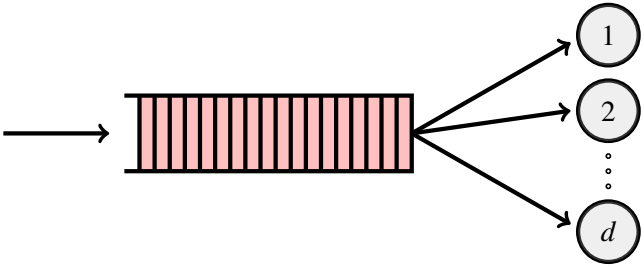


Heavy-Tailed Increments

Congestion of Multiple Server Queue:

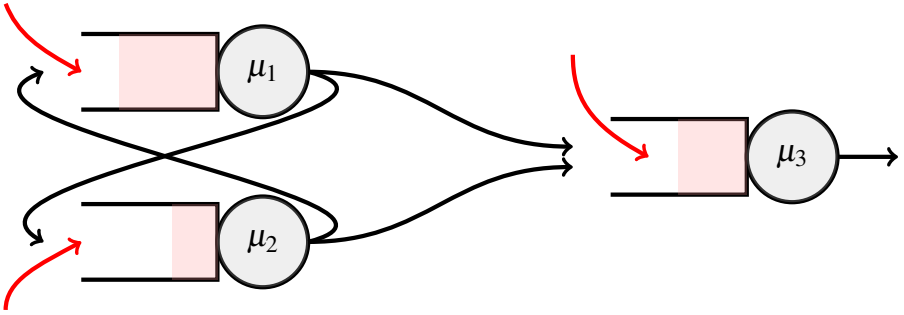


Congestion of Multiple Server Queue:



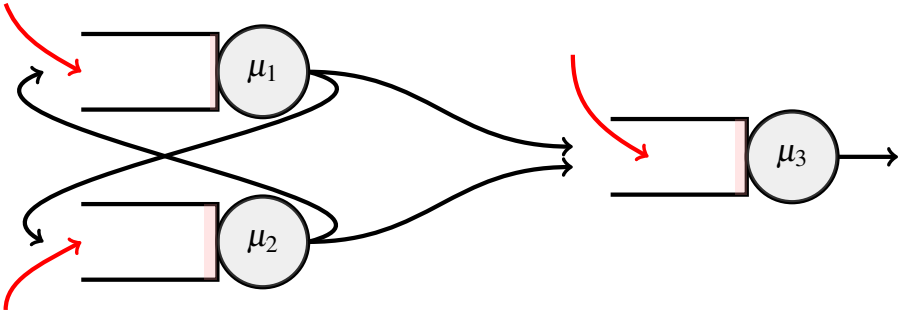
Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



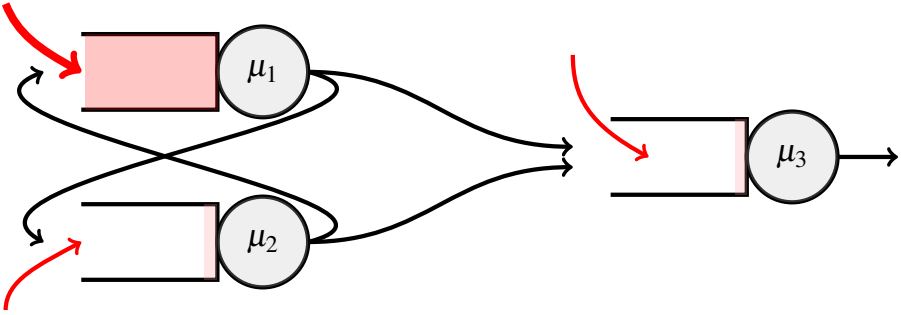
Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:



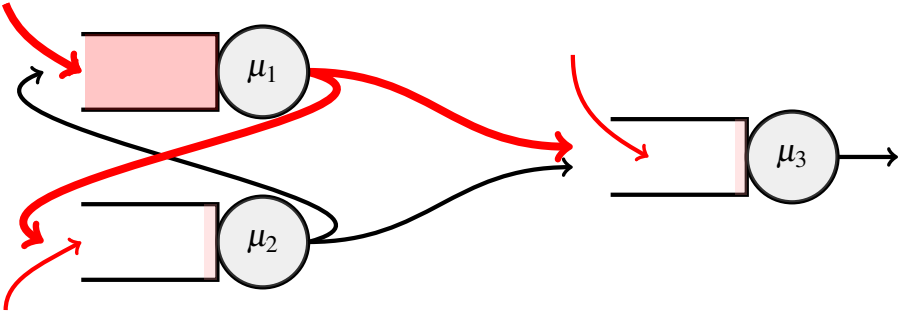
Conspiracy vs Catastrophe

Congestion of Stochastic Fluid Networks:

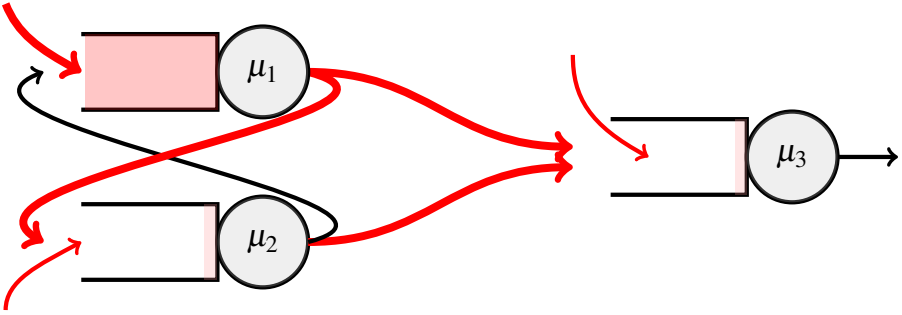


Conspiracy vs Catastrophe

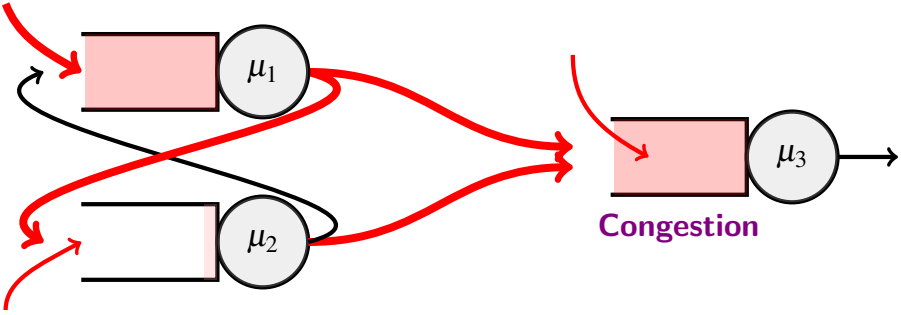
Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



Congestion of Stochastic Fluid Networks:



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

Open Problems Solved by Heavy-Tailed Large Deviations

- Design of strongly efficient rare-event simulation algorithm
- Multiple server queue length asymptotics
- Ruin probability under the presence of reinsurance contracts

From Large Deviations to Metastability

Stochastic Gradient Descent (SGD)

Minimizing loss function f:

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

Stochastic Gradient Descent (SGD)

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 (SGD)

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 (SGD)

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 (SGD)

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 (SGD)

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 (SGD)

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 (SGD)

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 (SGD)

Minimizing loss function f:

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

Then

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

Gradient Flow

where

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

Heavy-Tailed Large Deviations for SGD

Theorem (Wang, R., 2023+)

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

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

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

Heavy-Tailed Large Deviations for SGD

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

Theorem (Wang, R., 2023+)

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

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

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

Heavy-Tailed Large Deviations for SGD

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

Theorem (Wang, R., 2023+)

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

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

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

Heavy-Tailed Large Deviations for SGD

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

Theorem (Wang, R., 2023+)

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

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

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

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

$\mathbb{M}$-Convergence

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

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

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

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

$\mathbb{M}$-Convergence

ε-fattening of $\mathbb{C}$

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

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

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

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

$\mathbb{M}$-Convergence

ε-fattening of $\mathbb{C}$

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

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

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

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

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

$\mathbb{M}$-Convergence

ε-fattening of $\mathbb{C}$

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

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

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

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

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

Definition (Uniform $\mathbb{M}$-convergence; Wang, R., 2023+)

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

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

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

Theorem (Wang, R., 2023+)

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

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

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

Theorem (Wang, R., 2023+)

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

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

Then the following statements are equivalent:

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

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

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

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

Theorem (Wang, R., 2023+)

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

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

Then the following statements are equivalent:

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

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

Furthermore, they both imply

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\tau_{I(\varepsilon)^\complement}$: exit time from $I(\varepsilon)$

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

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

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

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

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

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

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

$\tau_{I(\varepsilon)^\complement}$: exit time from $I(\varepsilon)$

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

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

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

Also,

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

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

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

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

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

$\tau_{I(\varepsilon)^\complement}$: exit time from $I(\varepsilon)$

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

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

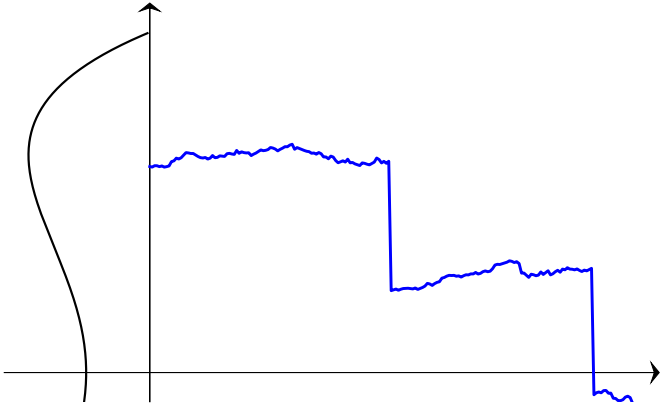
$$\lim_{\varepsilon \to 0} \lim_{T \to \infty} \delta_B(\varepsilon, T) = 0.$$

Also,

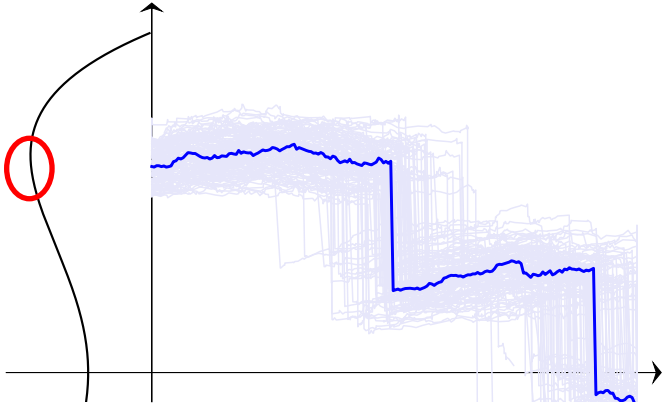
$$\limsup_{\eta \downarrow 0} \frac{\sup_{x \in I(\varepsilon)} \mathbf{P}\big(\tau^\eta_{(I(\varepsilon) \setminus A(\varepsilon))^\complement}(x) > T/\eta\big)}{\gamma(\eta)T/\eta} = 0;$$

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

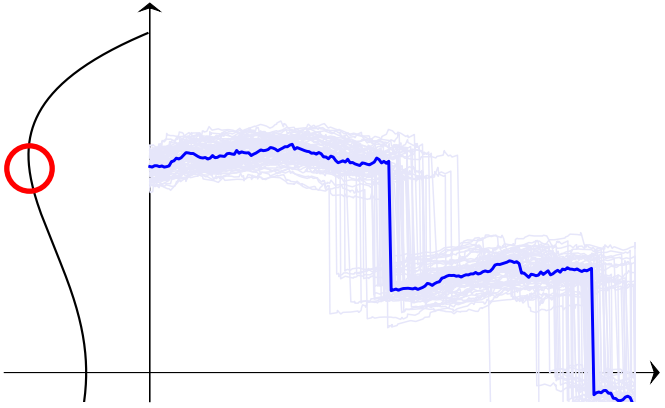
Asymptotic Atoms



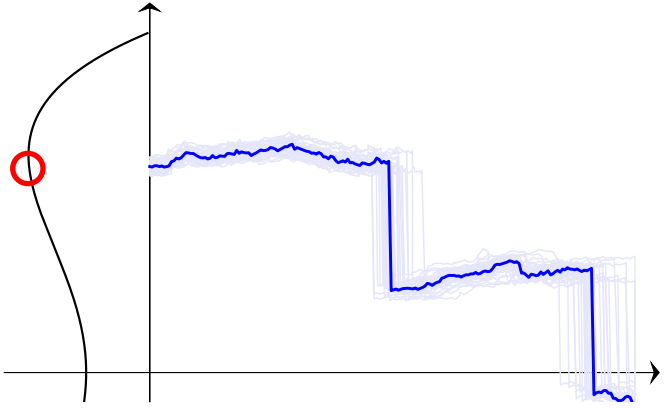
Asymptotic Atoms



Asymptotic Atoms



Asymptotic Atoms



Exit Time and Location under the Presence of Asymptotic Atom

Theorem (Wang, R., 2023+)

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

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

Truncated Version of Stochastic Gradient Descent

SGD

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

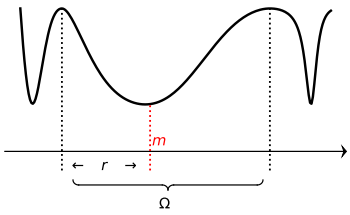
Then, again,

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

where

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

First Exit Time Analysis



gradient noise

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

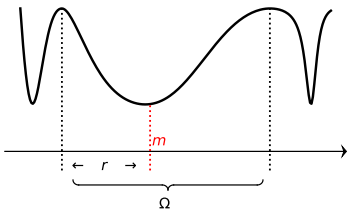
$$\mathbf{P}(|Z_i| \geq x) = x^{-(\alpha+1)}, \quad \alpha > 0$$

Theorem (Wang, R., 2023+)

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

$$\tau(\eta)\gamma(\eta) \Rightarrow \mathrm{Exp}(1)$$

First Exit Time Analysis



gradient noise

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

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

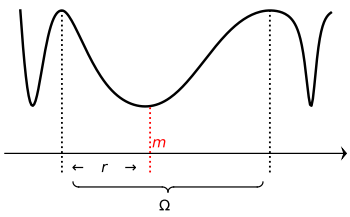
Theorem (Wang, R., 2023+)

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

First Exit Time

$$\tau(\eta)\gamma(\eta) \Rightarrow \mathrm{Exp}(1)$$

First Exit Time Analysis



gradient noise

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

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

Theorem (Wang, R., 2023+)

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

First Exit Time

$$\tau(\eta)\gamma(\eta) \Rightarrow \mathrm{Exp}(1)$$

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

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, R., 2023+)

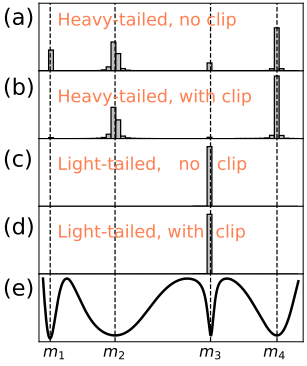
Under certain structural conditions, for any $t > 0$ and $\beta > 1 + \alpha \cdot l^*$,

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

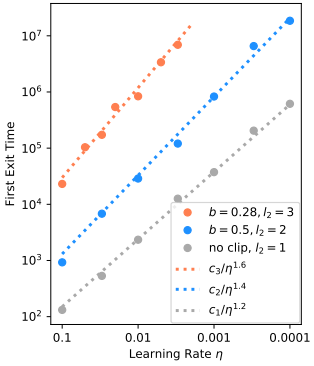
In fact, $W^\eta_{\lfloor t/\eta^{1+\alpha \cdot l^*} \rfloor}$ converges to a Markov jump processes

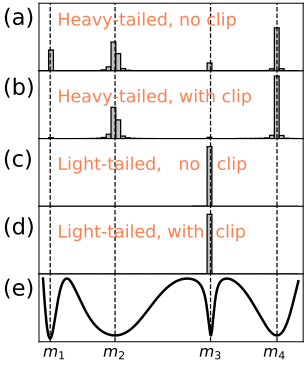
whose state space consists of wide local minima only.

Eliminating Sharp Local Minima with Truncated Heavy-Tails

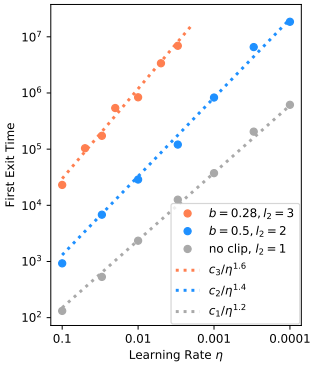


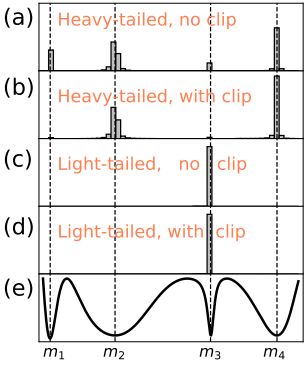
Eliminating Sharp Local Minima with Truncated Heavy-Tails

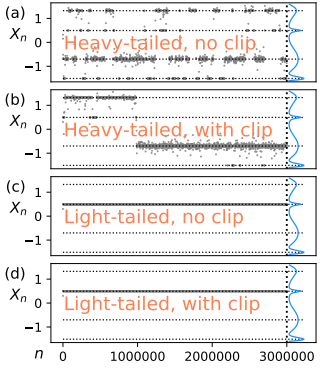




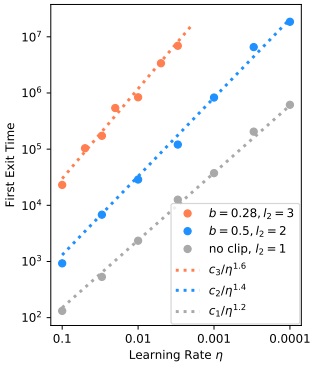
Eliminating Sharp Local Minima with Truncated Heavy-Tails

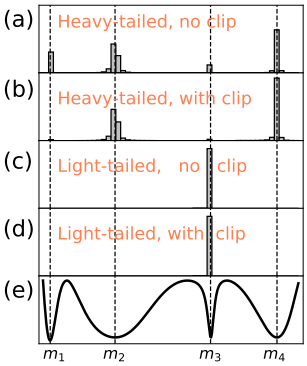


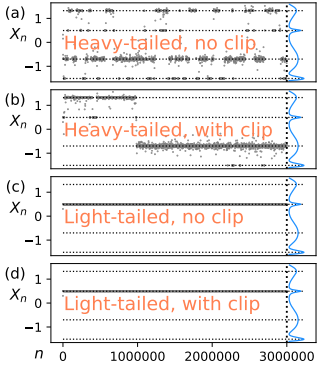




Eliminating Sharp Local Minima with Truncated Heavy-Tails

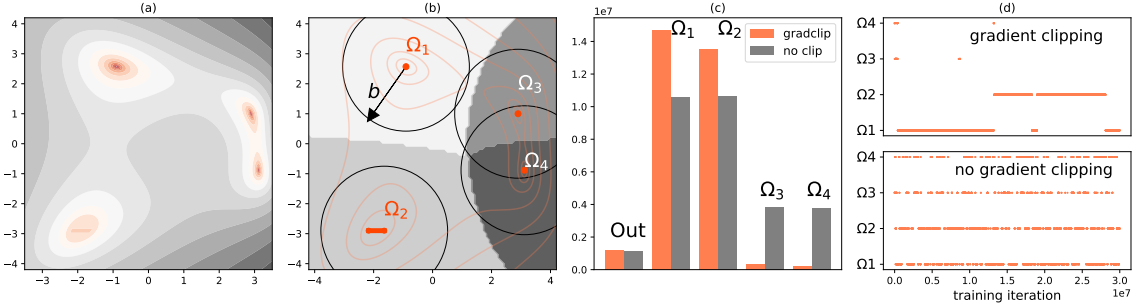




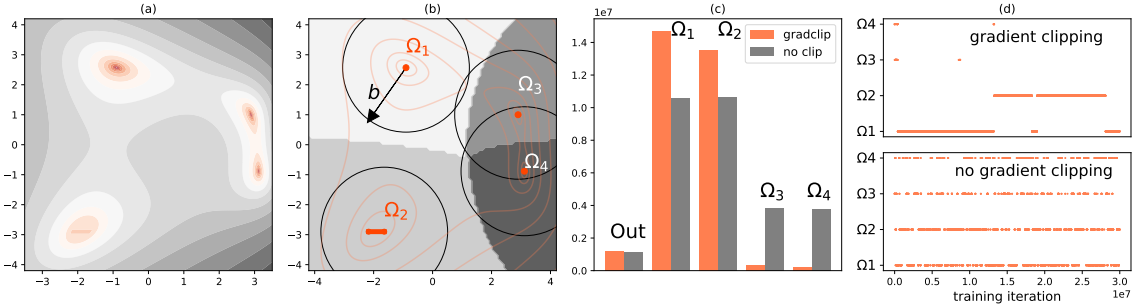


Consistent with what theory predicts!

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!

In Deep Neural Network

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

In Deep Neural Network

Stochastic Gradient

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

In Deep Neural Network

Stochastic Gradient

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

In Deep Neural Network

Stochastic Gradient

Pareto RV

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

In Deep Neural Network

Heavy-Tailed Stochastic Gradient

Pareto RV

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

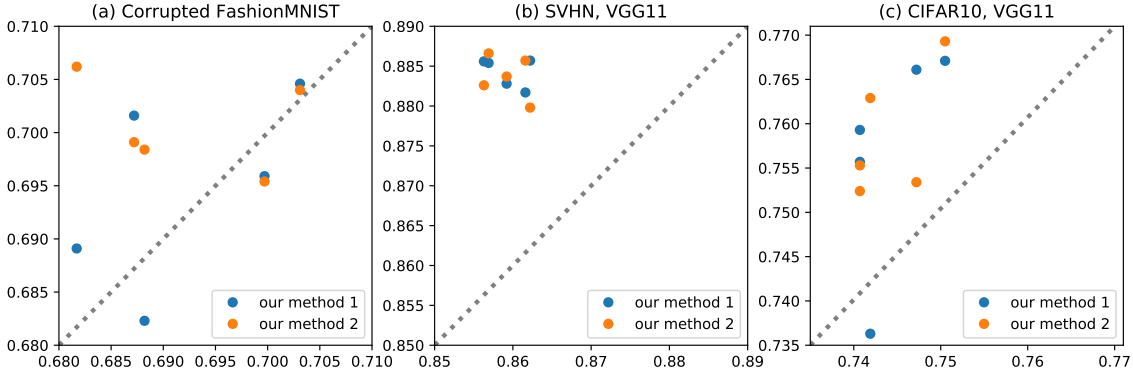
In Deep Neural Network

Heavy-Tailed Stochastic Gradient Pareto RV

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

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

Improvement is Consistent



Summary

Summary

- Heavy-tailed large deviations crisply characterize Catastrophe Principle

Summary

- Heavy-tailed large deviations crisply characterize Catastrophe Principle

- New locally uniform large deviations formulation

Summary

- Heavy-tailed large deviations crisply characterize Catastrophe Principle
- New locally uniform large deviations formulation
- Uniform $\mathbb{M}$-convergence, asymptotic atom, portmanteau theorem

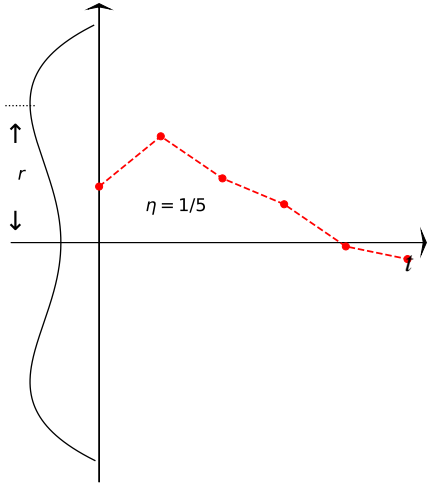
Summary

- Heavy-tailed large deviations crisply characterize Catastrophe Principle
- New locally uniform large deviations formulation
- Uniform $\mathbb{M}$-convergence, asymptotic atom, portmanteau theorem
- These tools reveal the global dynamics of heavy-tailed dynamical systems

Summary

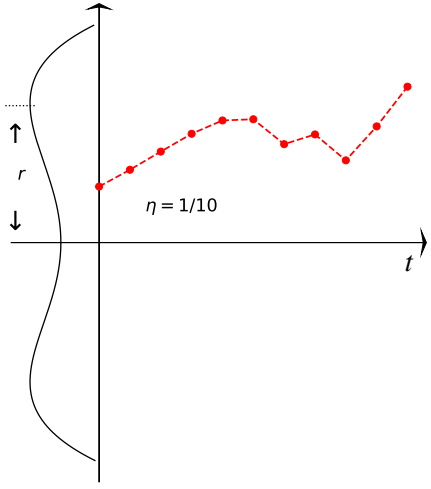
- Heavy-tailed large deviations crisply characterize Catastrophe Principle
- New locally uniform large deviations formulation
- Uniform $\mathbb{M}$-convergence, asymptotic atom, portmanteau theorem
- These tools reveal the global dynamics of heavy-tailed dynamical systems
- Tail-Inflation-Truncation strategy improves SGDs' generalization performance in DNN

Gradient Flow: Law of Large Numbers



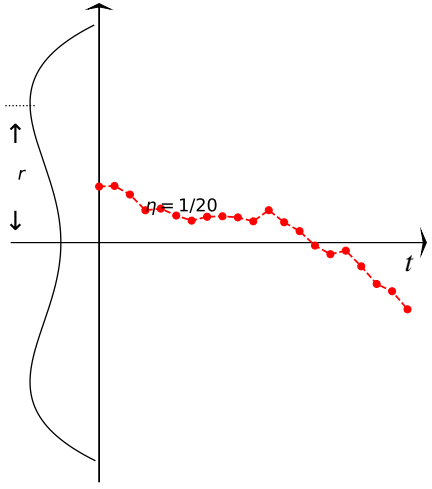
(90° rotation of) f

Gradient Flow: Law of Large Numbers



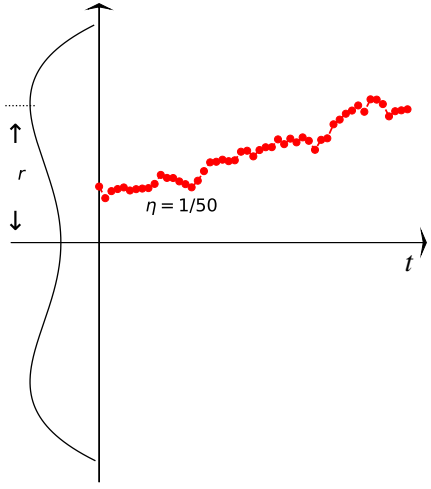
(90° rotation of) f

Gradient Flow: Law of Large Numbers



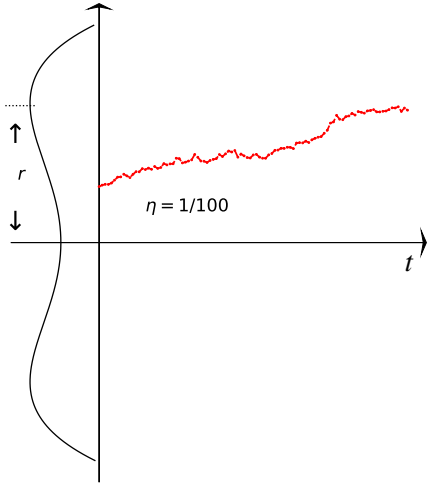
(90° rotation of) f

Gradient Flow: Law of Large Numbers



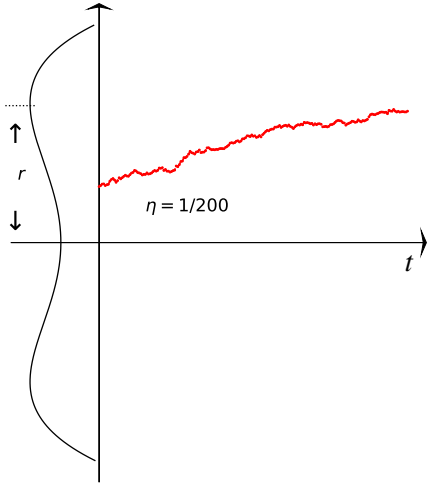
(90° rotation of) f

Gradient Flow: Law of Large Numbers



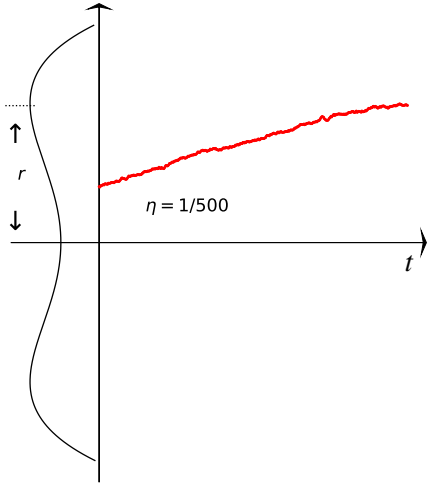
(90° rotation of) f

Gradient Flow: Law of Large Numbers



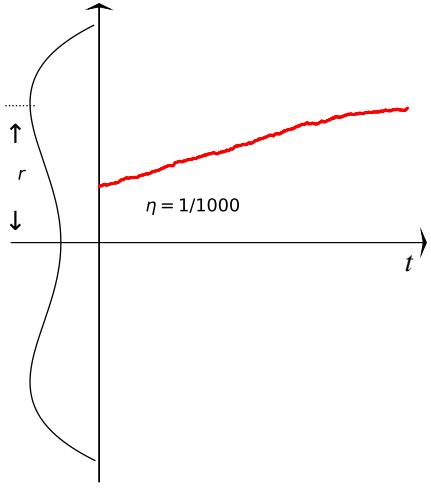
(90° rotation of) f

Gradient Flow: Law of Large Numbers



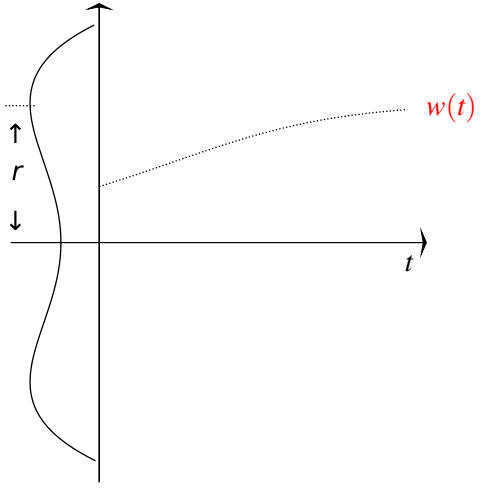
(90° rotation of) f

Gradient Flow: Law of Large Numbers



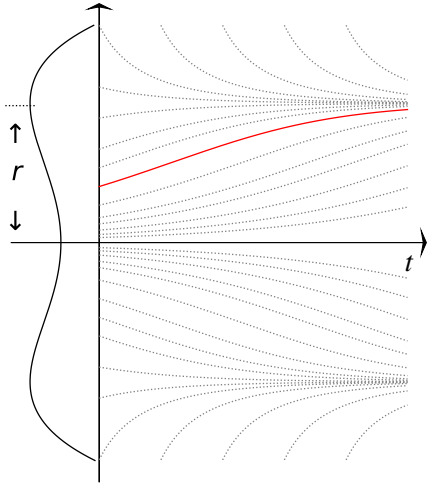
(90° rotation of) f

Gradient Flow: Law of Large Numbers



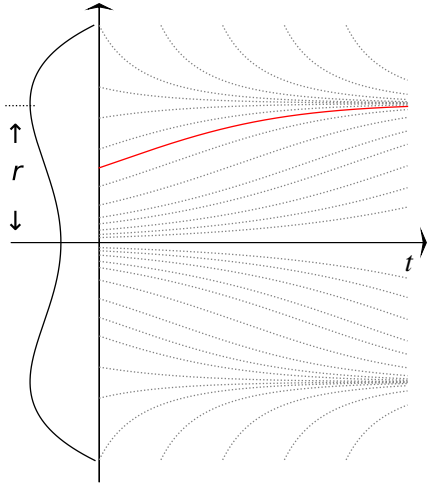
(90° rotation of) f

Gradient Flow: Law of Large Numbers



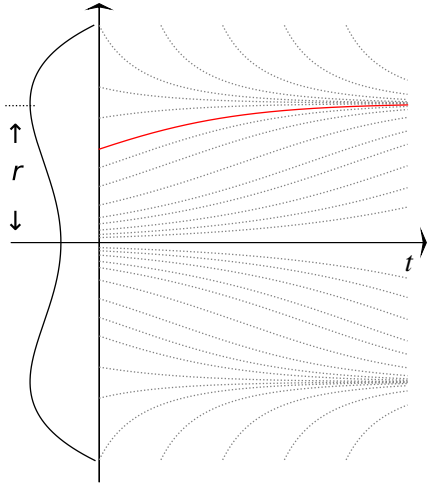
(90° rotation of) f

Gradient Flow: Law of Large Numbers



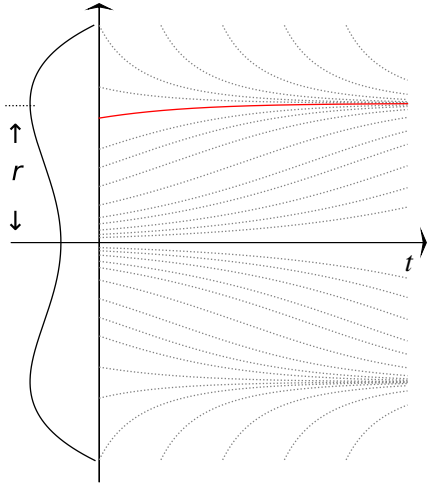
(90° rotation of) f

Gradient Flow: Law of Large Numbers



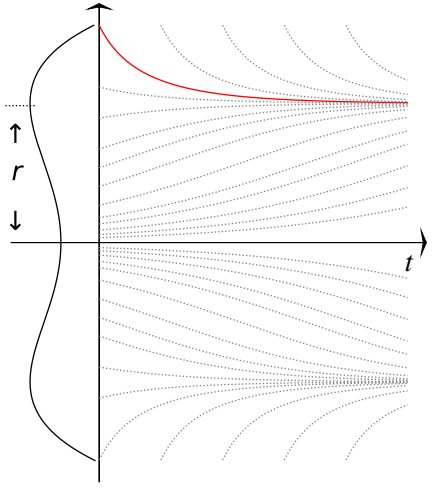
(90° rotation of) f

Gradient Flow: Law of Large Numbers



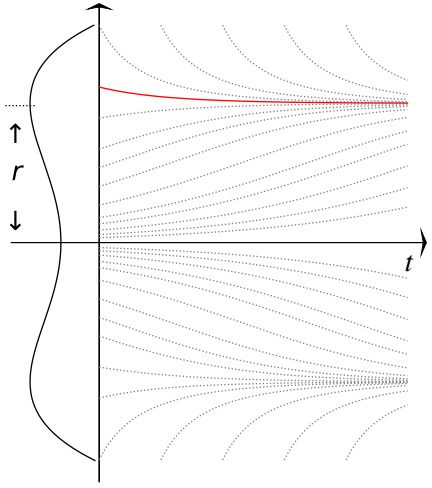
(90° rotation of) f

Gradient Flow: Law of Large Numbers



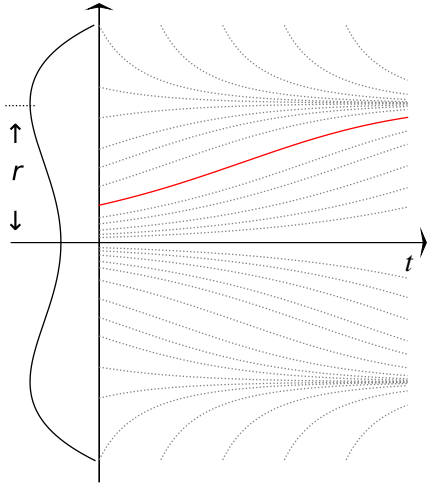
(90° rotation of) f

Gradient Flow: Law of Large Numbers



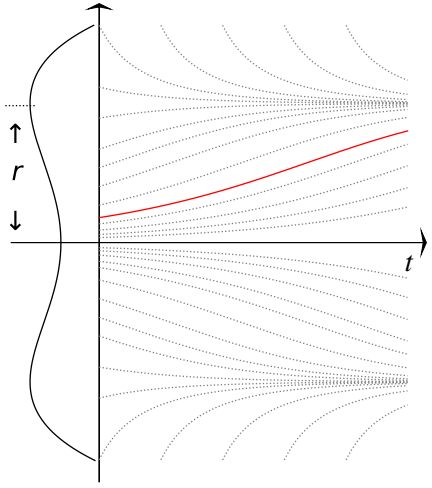
(90° rotation of) f

Gradient Flow: Law of Large Numbers



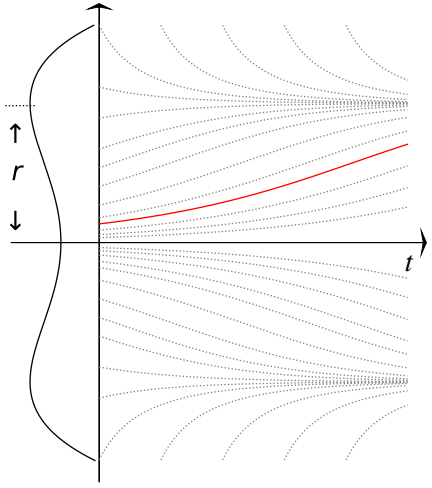
(90° rotation of) f

Gradient Flow: Law of Large Numbers



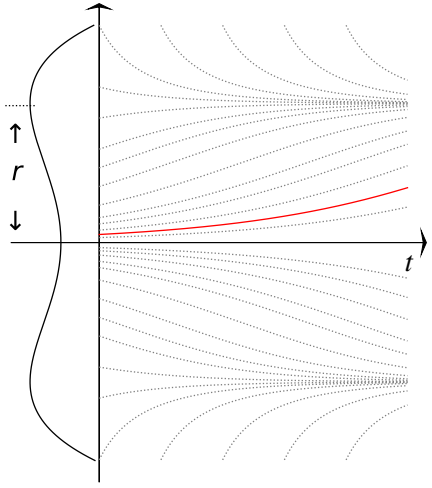
(90° rotation of) f

Gradient Flow: Law of Large Numbers



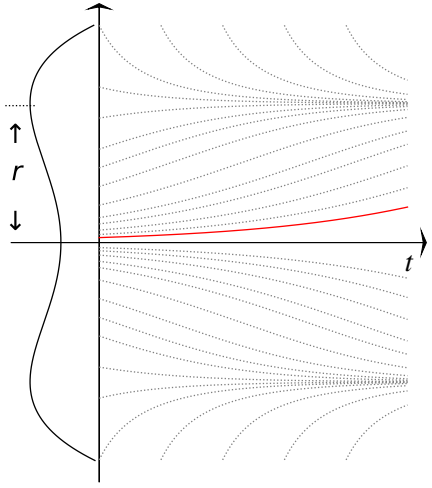
(90° rotation of) f

Gradient Flow: Law of Large Numbers



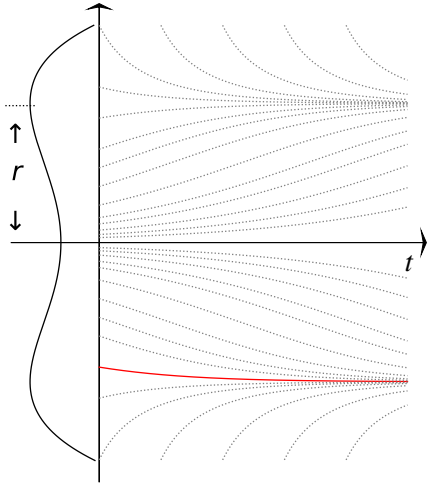
(90° rotation of) f

Gradient Flow: Law of Large Numbers



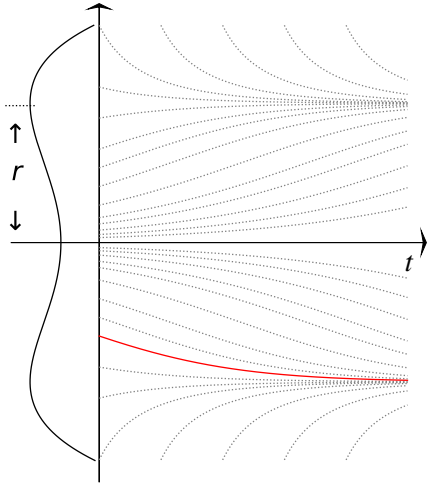
(90° rotation of) f

Gradient Flow: Law of Large Numbers



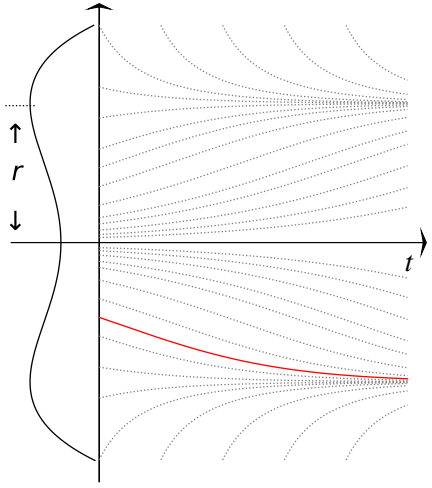
(90° rotation of) f

Gradient Flow: Law of Large Numbers



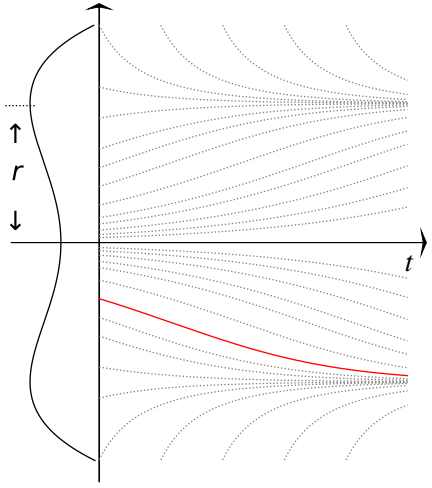
(90° rotation of) f

Gradient Flow: Law of Large Numbers



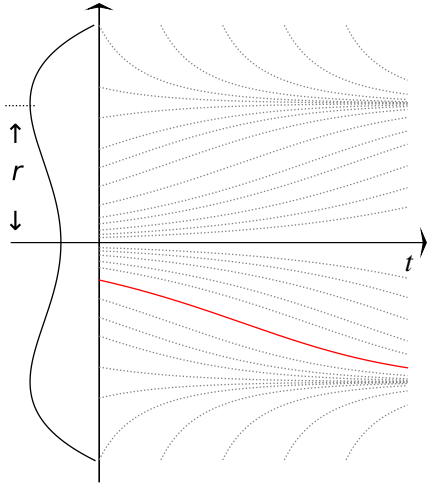
(90° rotation of) f

Gradient Flow: Law of Large Numbers



(90° rotation of) f

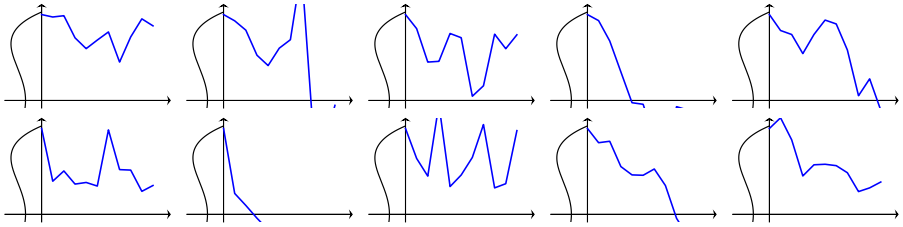
Gradient Flow: Law of Large Numbers



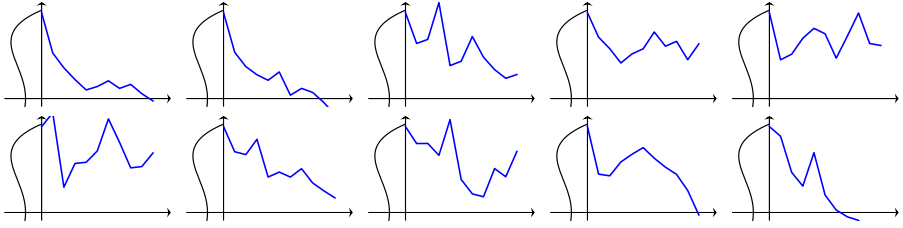
(90° rotation of) f

Typical Scenario

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

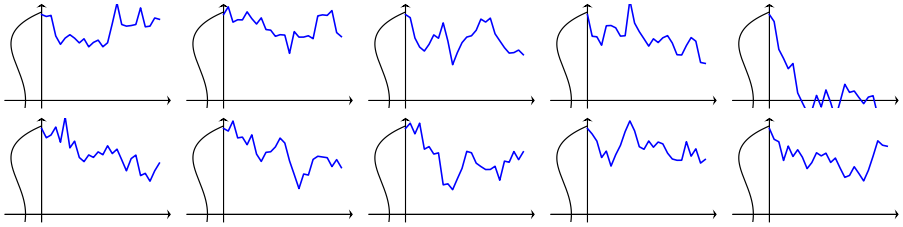


Trajectory of SGD W^η: $\eta = 1/10$ & noises are **heavy-tailed**

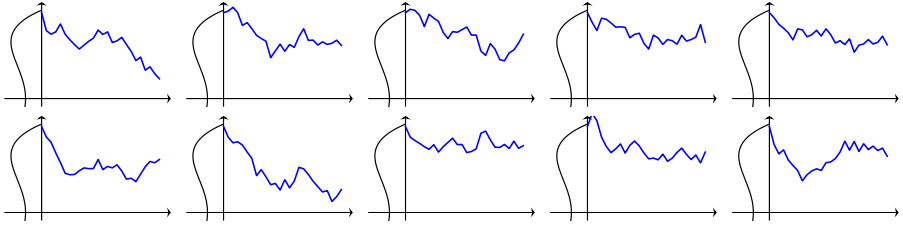


Typical Scenario

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

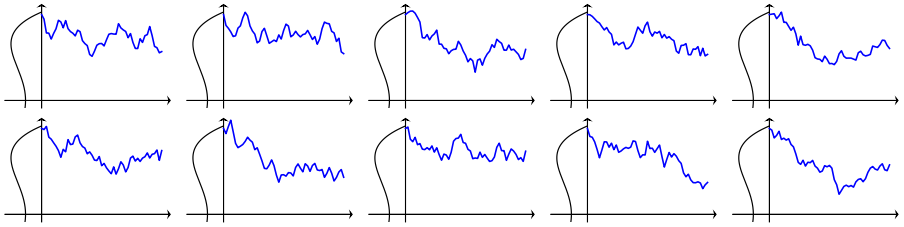


Trajectory of SGD W^η: $\eta = 1/25$ & noises are **heavy-tailed**

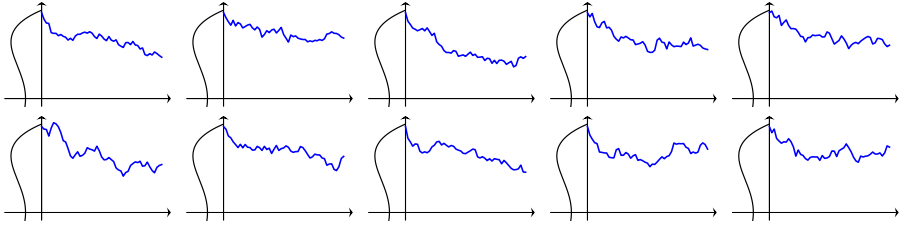


Typical Scenario

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

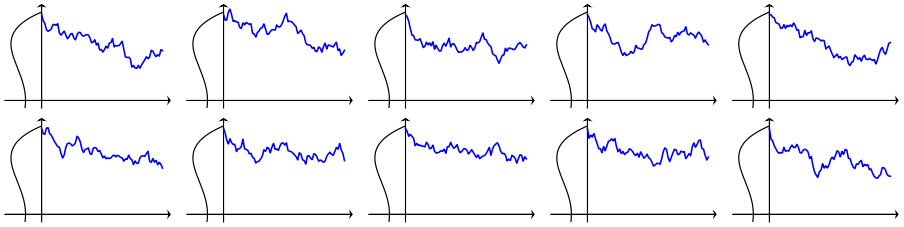


Trajectory of SGD W^η: $\eta = 1/50$ & noises are **heavy-tailed**

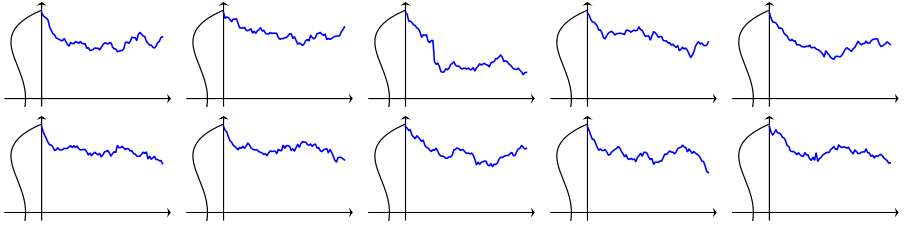


Typical Scenario

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

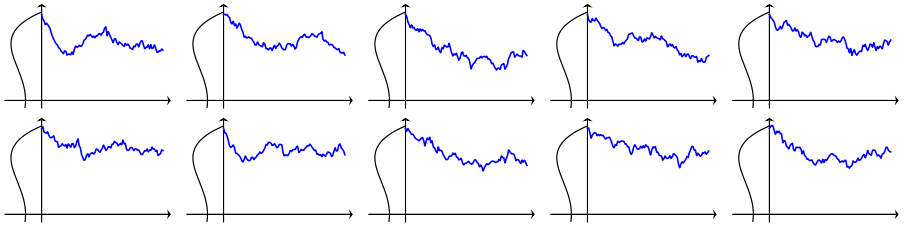


Trajectory of SGD W^η: $\eta = 1/75$ & noises are **heavy-tailed**

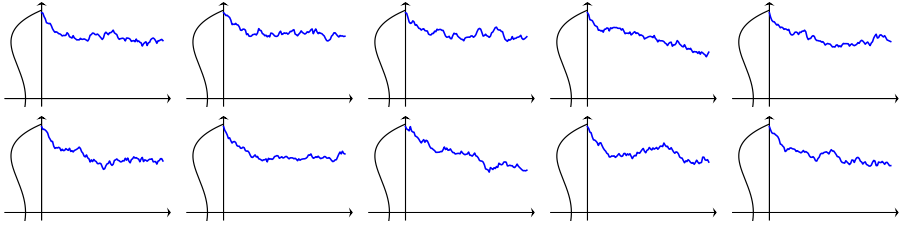


Typical Scenario

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

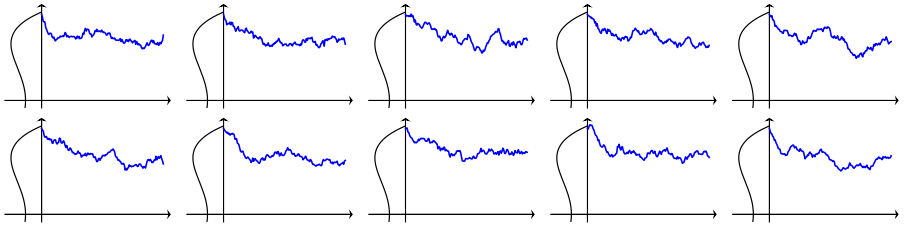


Trajectory of SGD W^η: $\eta = 1/100$ & noises are **heavy-tailed**

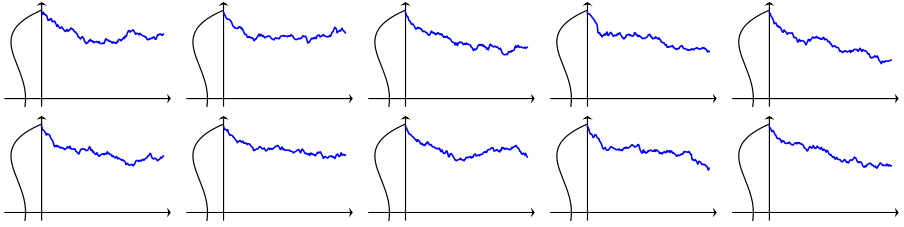


Typical Scenario

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

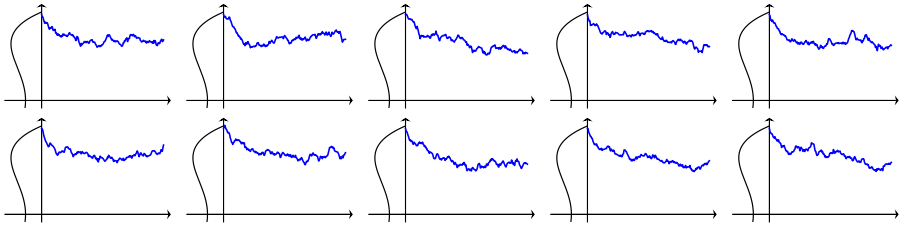


Trajectory of SGD W^η: $\eta = 1/150$ & noises are **heavy-tailed**

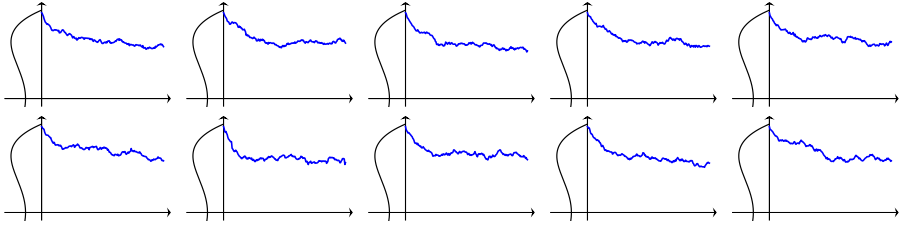


Typical Scenario

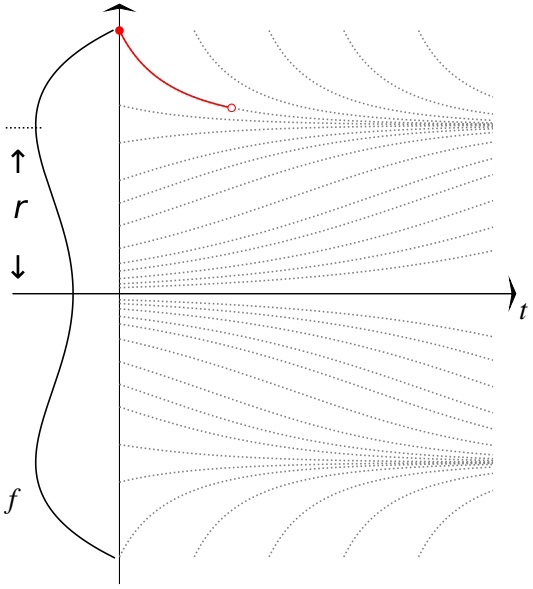
Trajectory of SGD W^η: $\eta = 1/200$ & noises are **light-tailed**



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

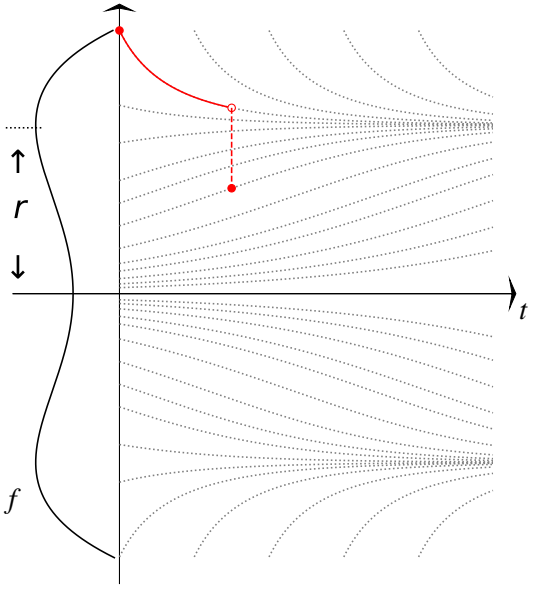


Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



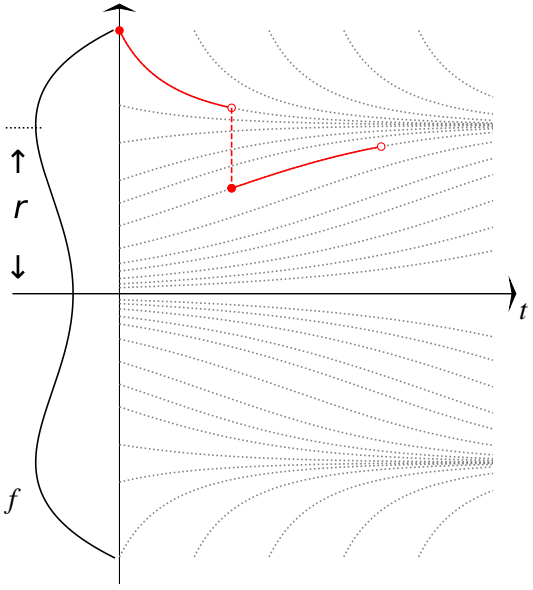
(90° rotation of) f

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



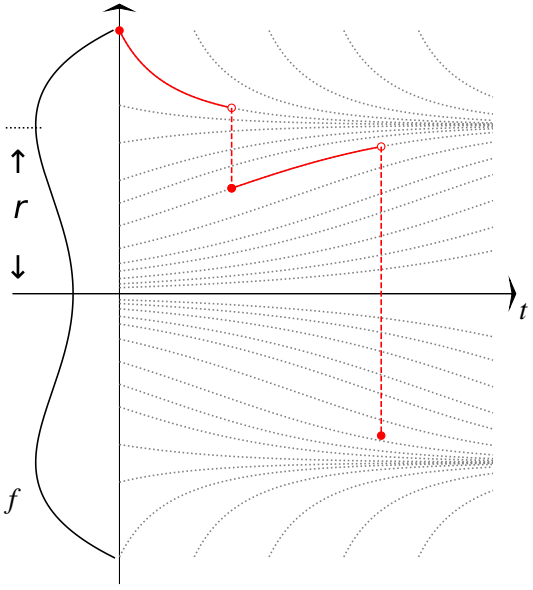
(90° rotation of) f

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



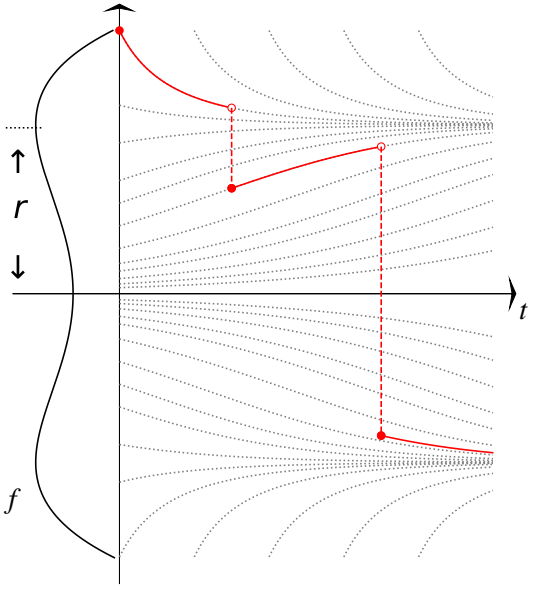
(90° rotation of) f

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



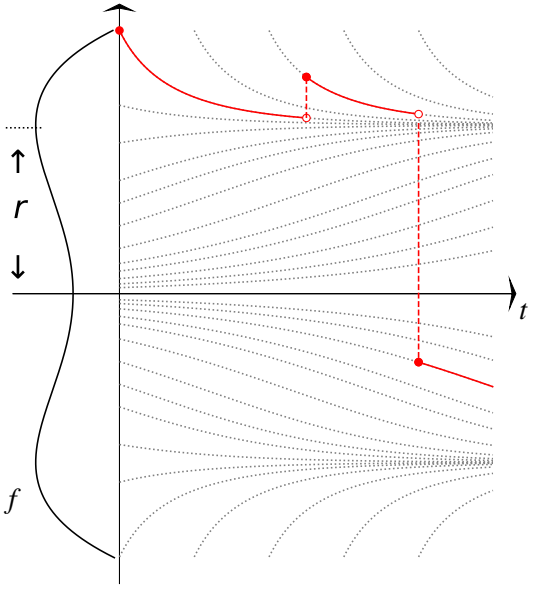
(90° rotation of) f

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



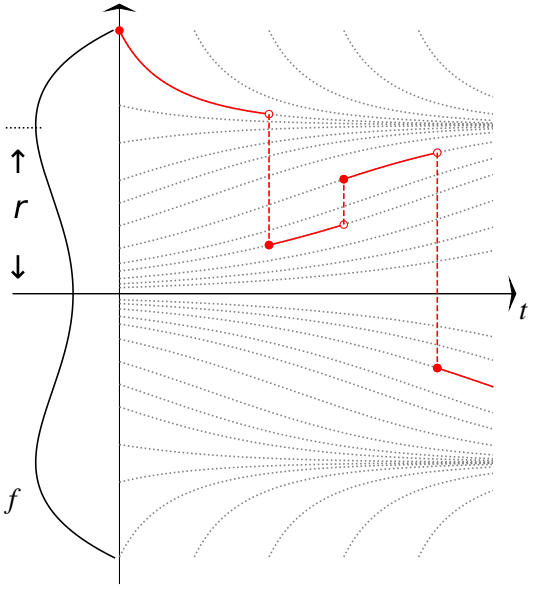
(90° rotation of) f

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



(90° rotation of) f

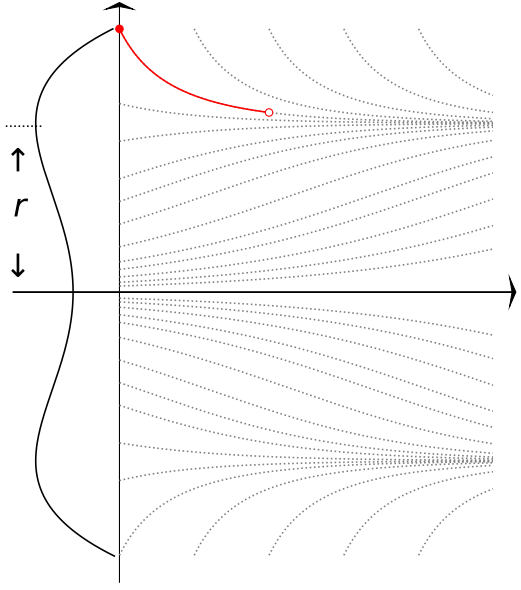
Adding Jumps to $w(\cdot)$: Piecewise Gradient Flow



(90° rotation of) f

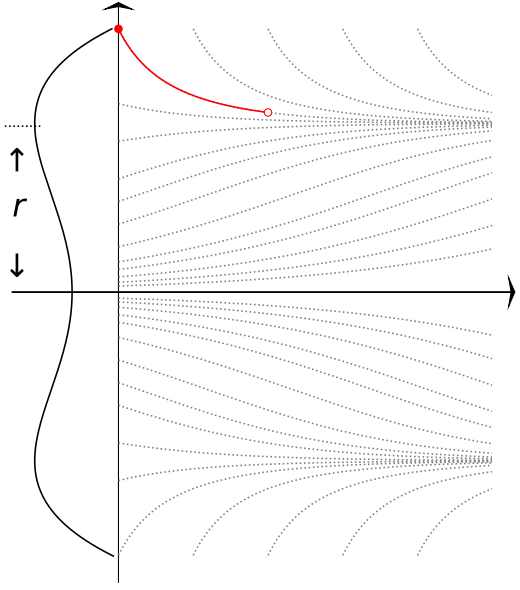
Catastrophe Principle: Most Likely Escape Route

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



Catastrophe Principle: Most Likely Escape Route

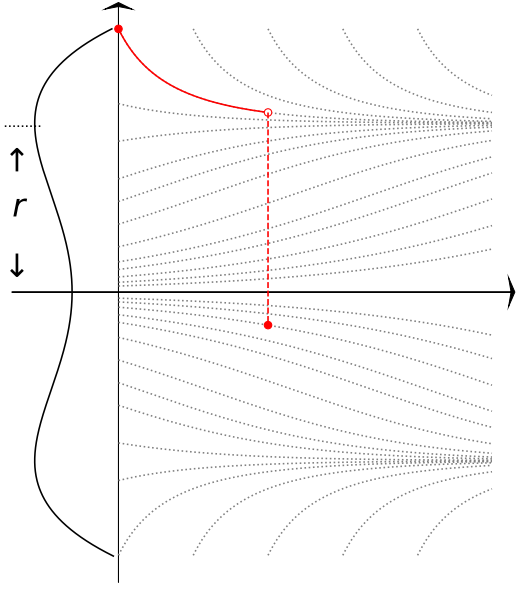
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

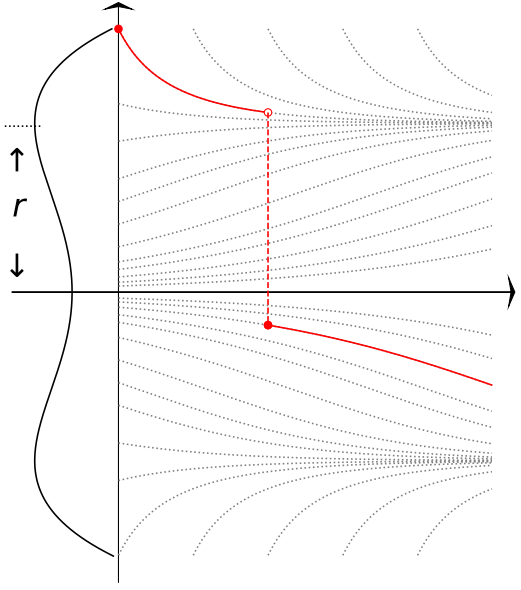
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

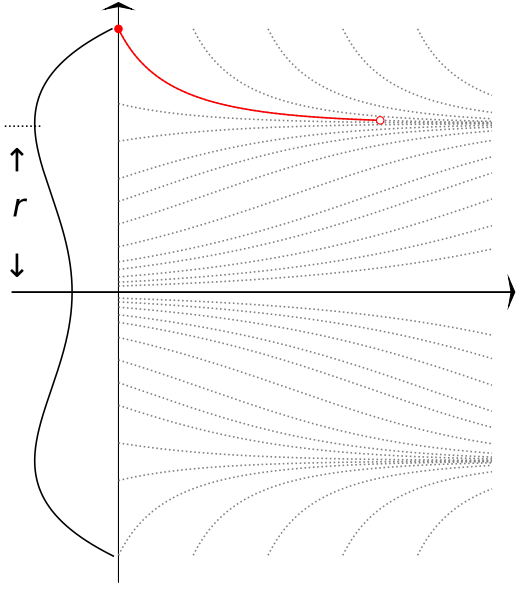
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

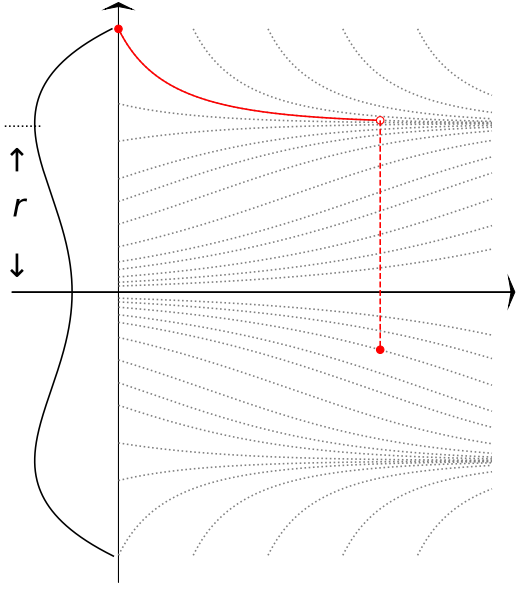
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

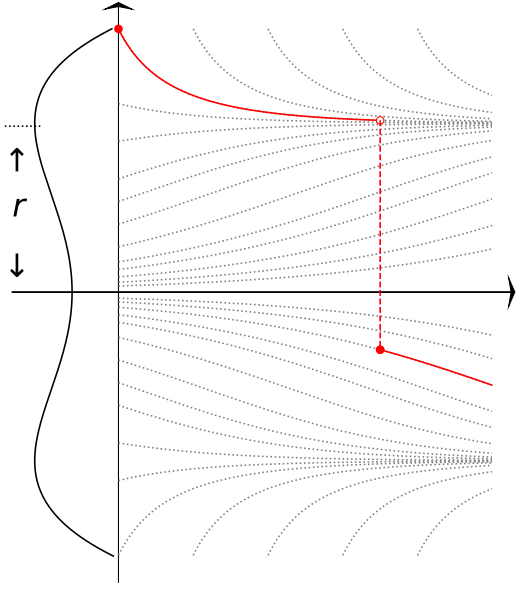
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

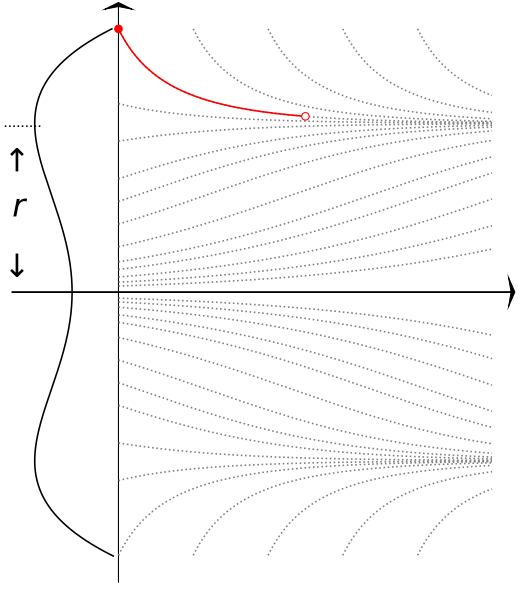
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

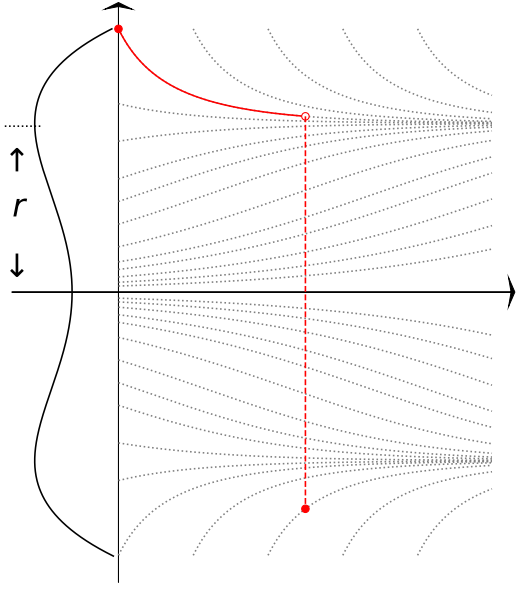
Most Likely Paths for $\{ W^{\eta}$ escapes the local minimum $\}$



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

Catastrophe Principle: Most Likely Escape Route

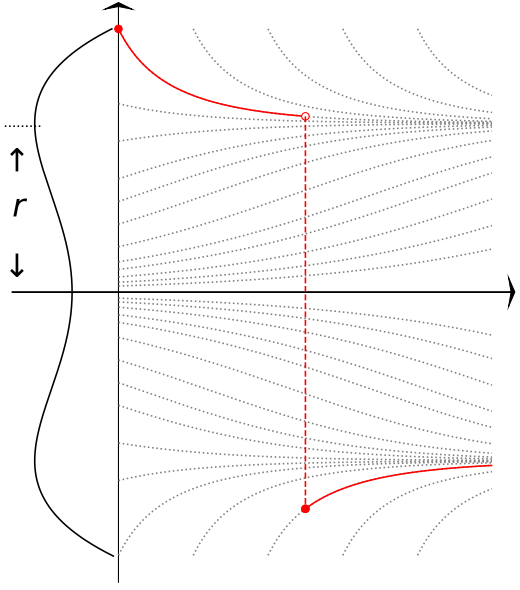
Most Likely Paths for $\{W^{\eta}$ escapes the local minimum$\}$



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

Catastrophe Principle: Most Likely Escape Route

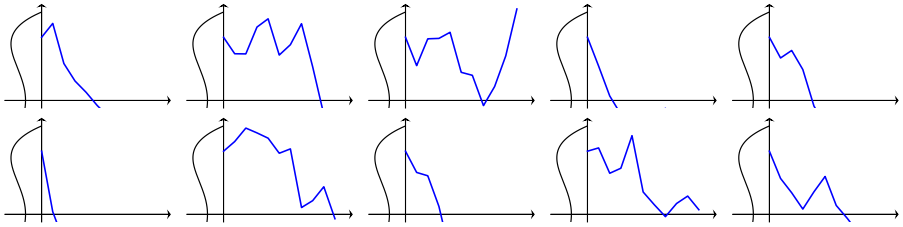
Most Likely Paths for $\{W^\eta$ escapes the local minimum$\}$



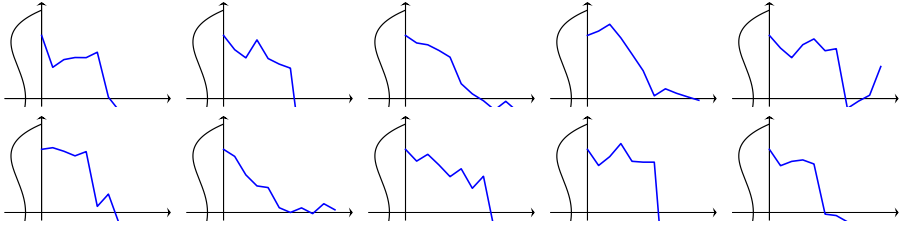
$\mathcal{J}(A) = 1$

Catastrophe Principle Dictates SGD's Escape Route

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

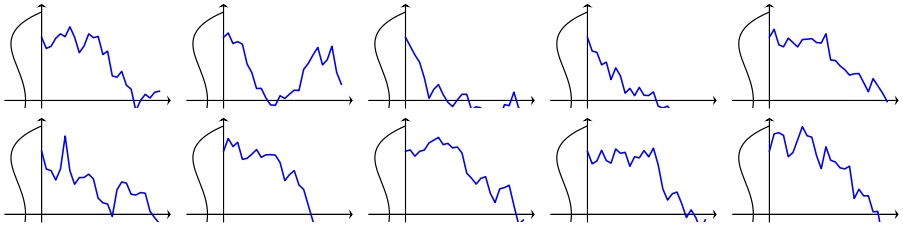


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

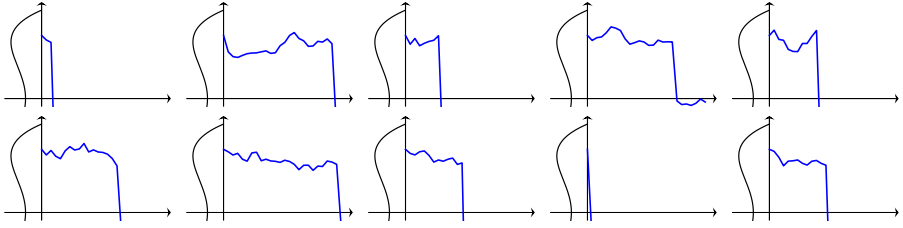


Catastrophe Principle Dictates SGD's Escape Route

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

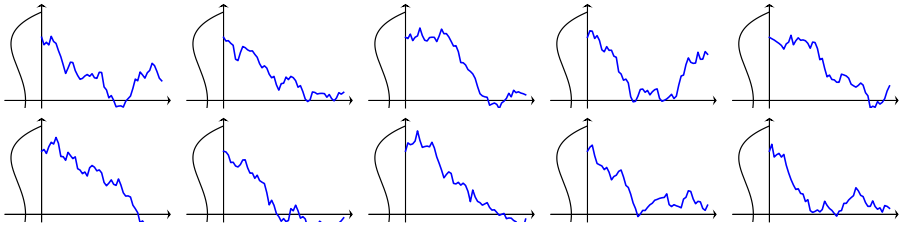


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

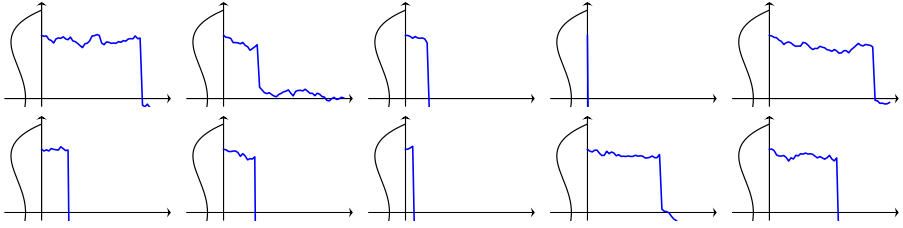


Catastrophe Principle Dictates SGD's Escape Route

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

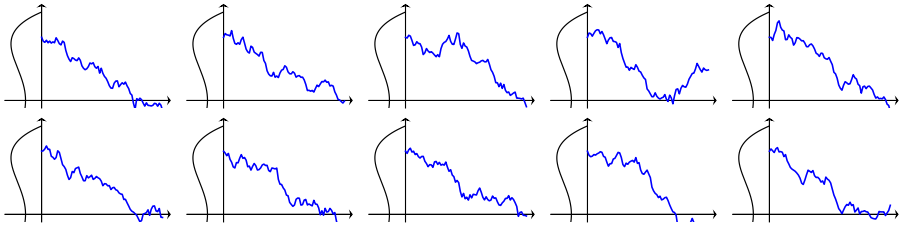


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/50$

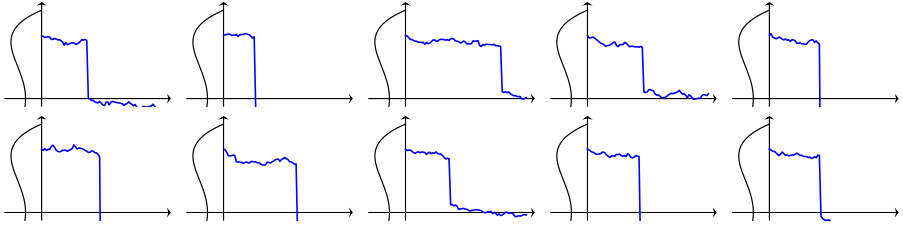


Catastrophe Principle Dictates SGD's Escape Route

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

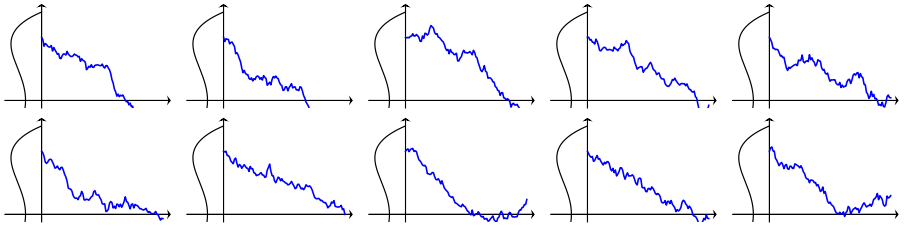


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

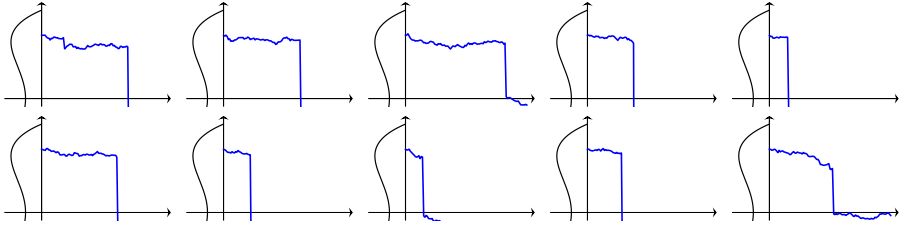


Catastrophe Principle Dictates SGD's Escape Route

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

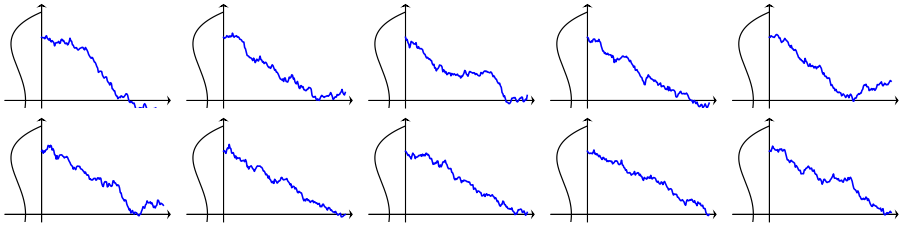


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

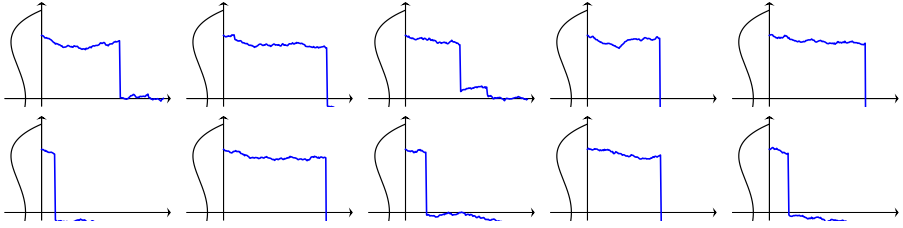


Catastrophe Principle Dictates SGD's Escape Route

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

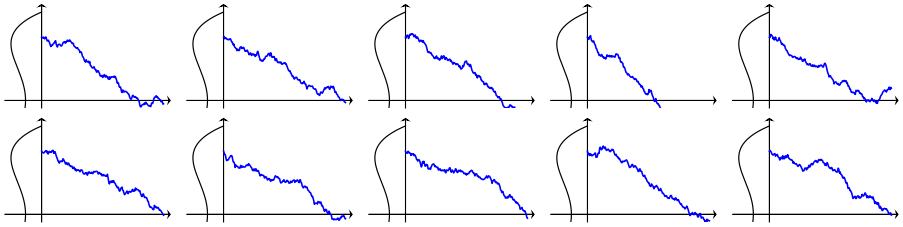


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

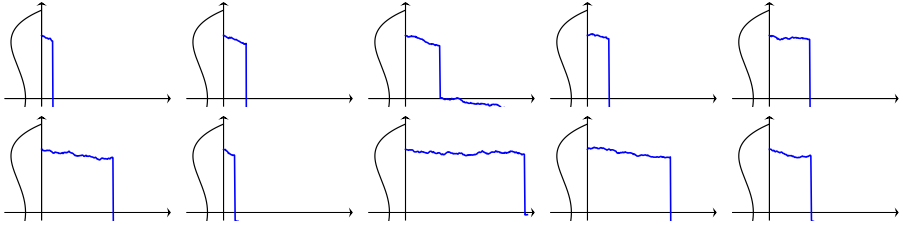


Catastrophe Principle Dictates SGD's Escape Route

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

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

Truncated Version of Stochastic Gradient Descent

SGD with Gradient Clipping

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

where

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

Then, again,

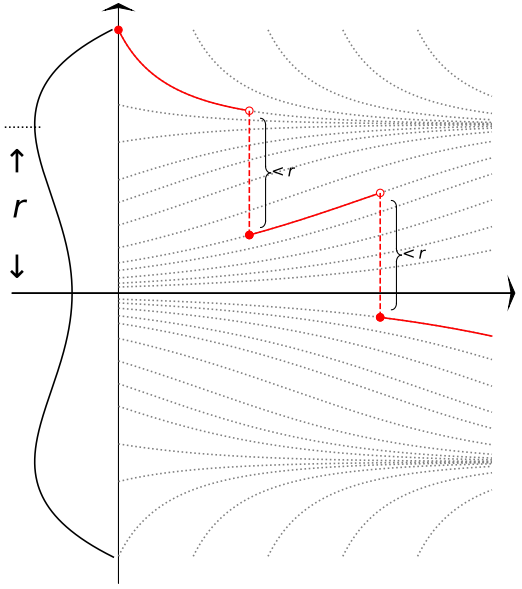
$$W^{\eta}(\cdot) \to w(\cdot) \qquad \text{as} \qquad \eta \to 0$$

where

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

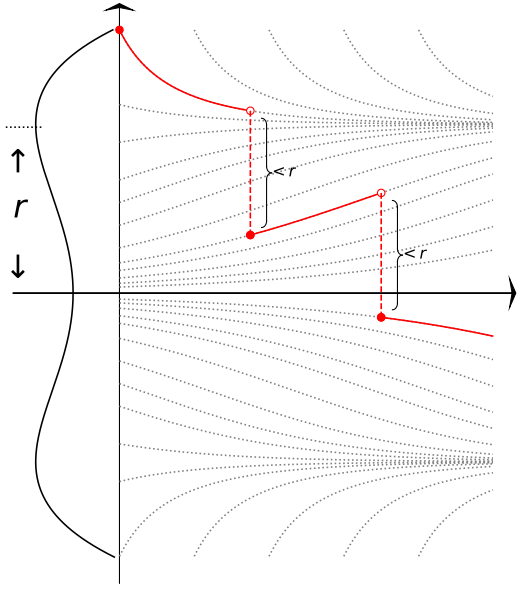
How does $\mathcal{J}(A)$ change?

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



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

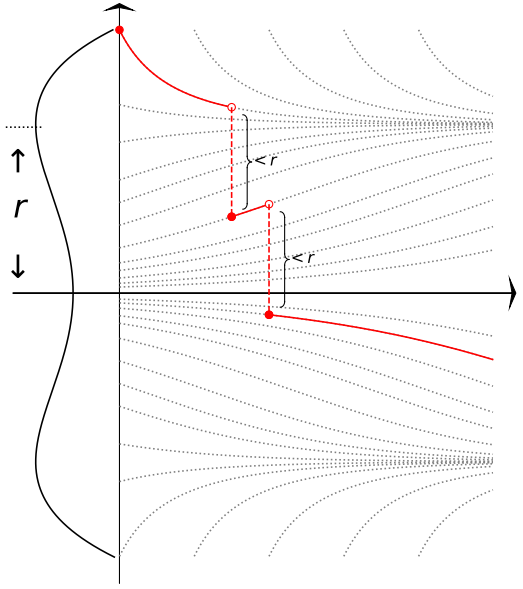
If $r \in (c, 2c)$



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

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

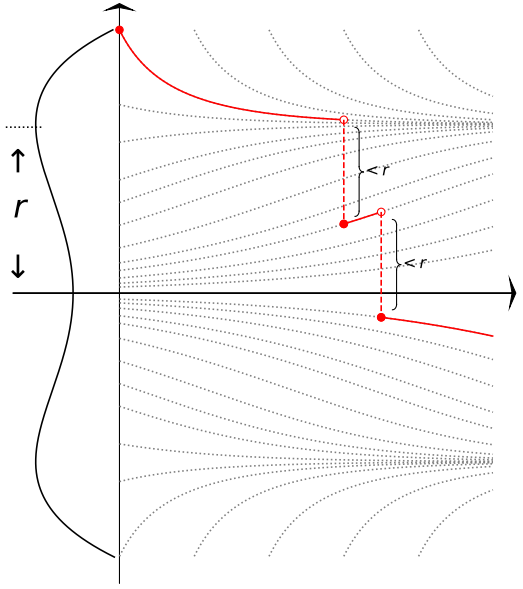
If $r \in (c, 2c)$



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

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

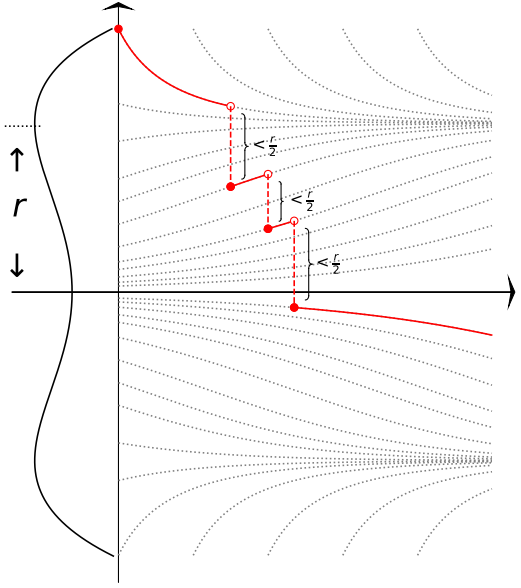
If $r \in (c, 2c)$



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

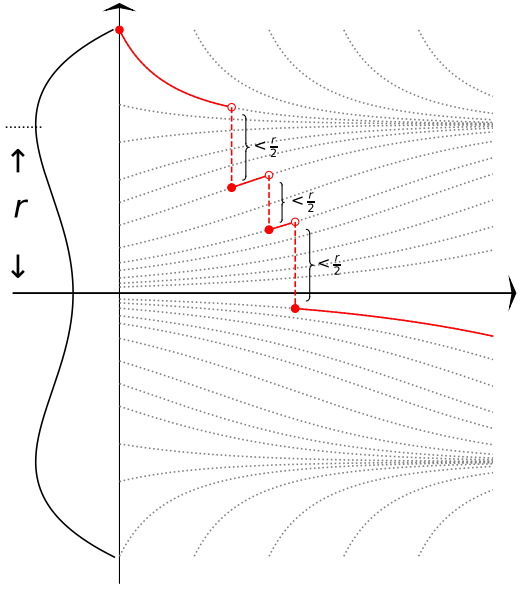
How does $\mathcal{J}(A)$ change?

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



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

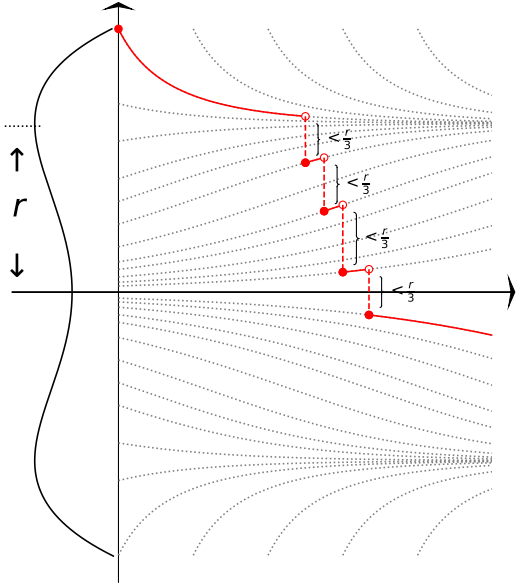
If $r \in (2c, 3c)$



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

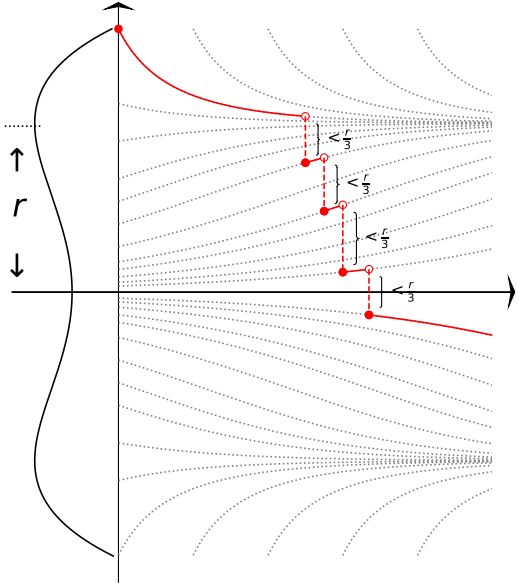
How does $\mathcal{J}(A)$ change?

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



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

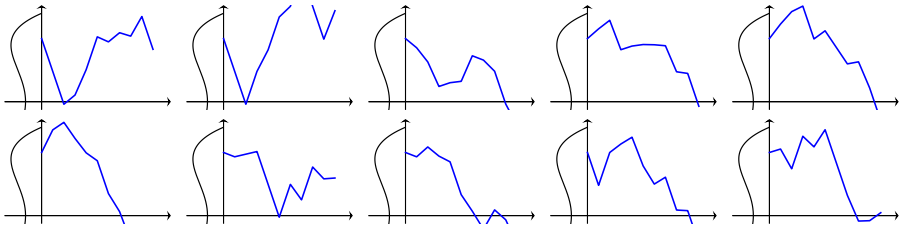
If $r \in (3c, 4c)$



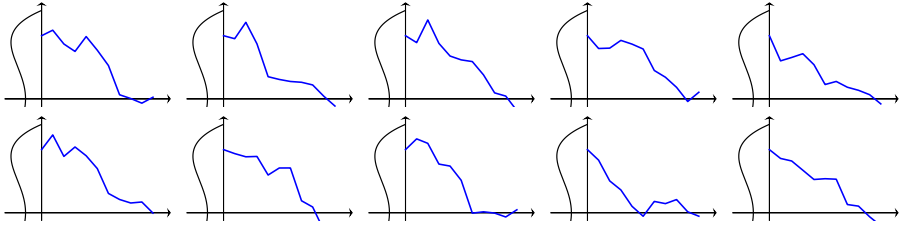
$\mathcal{J}(A) = 4$

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

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

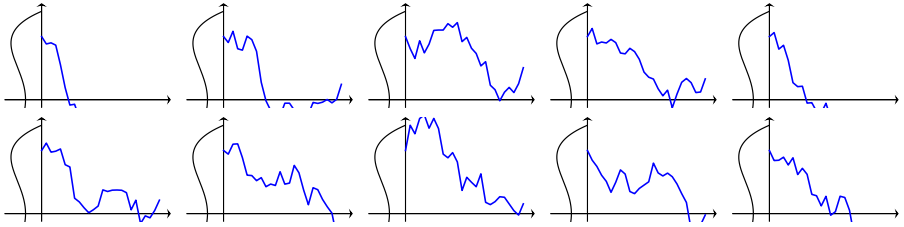


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

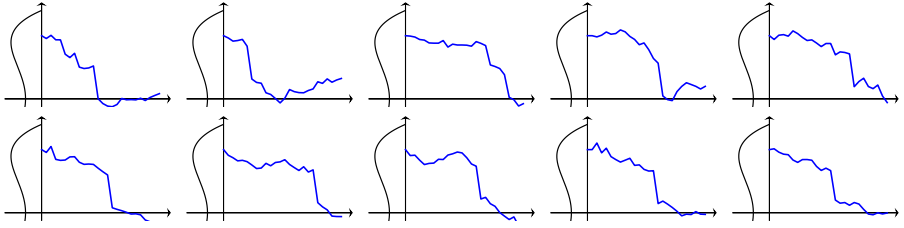


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

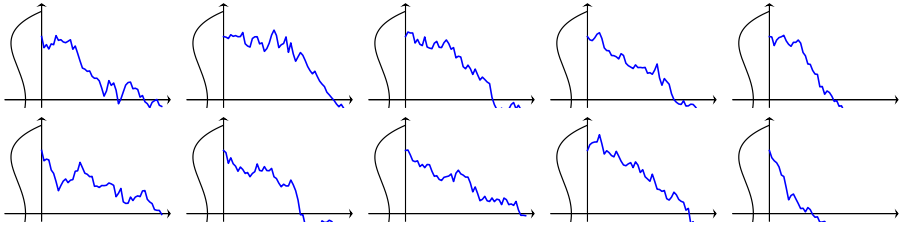


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/25$

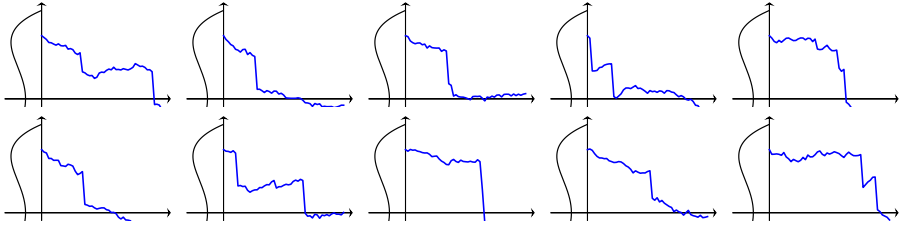


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

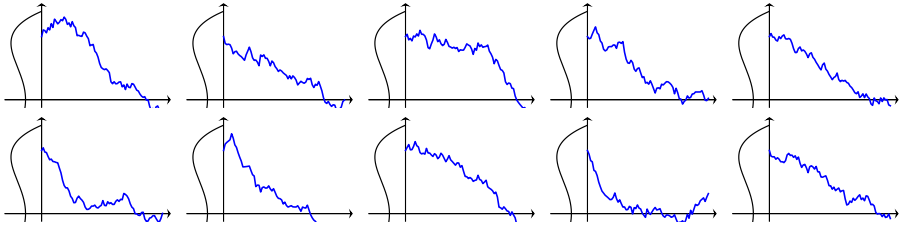


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/10$

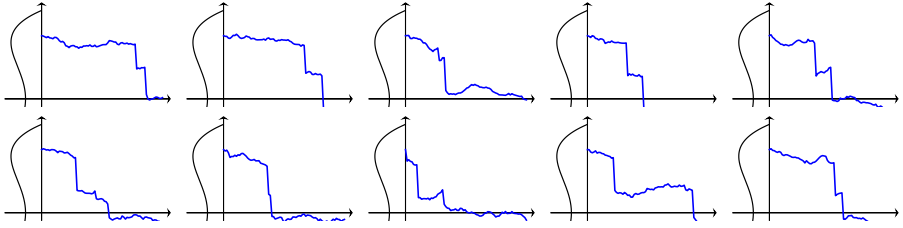


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

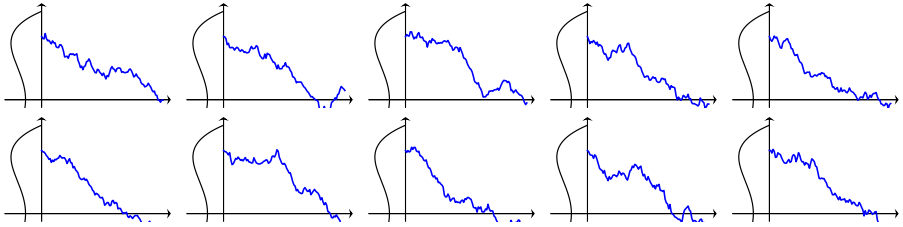


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/75$

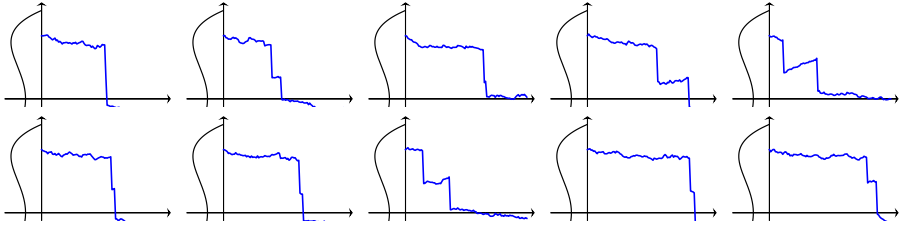


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

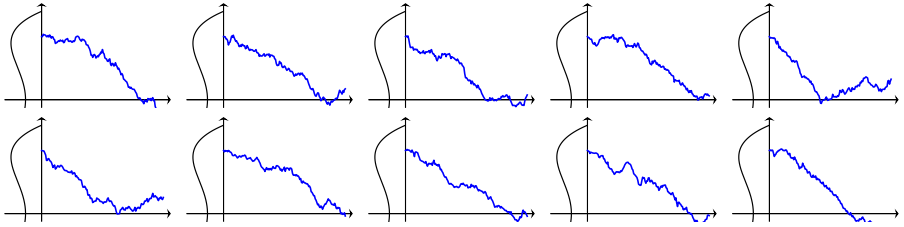


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/100$

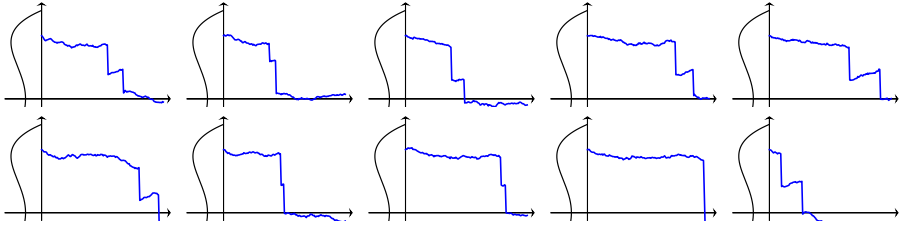


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

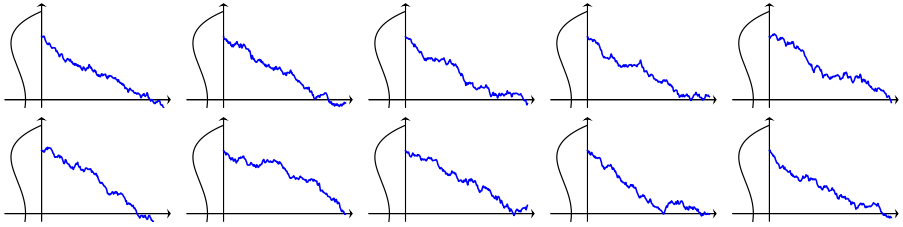


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/150$

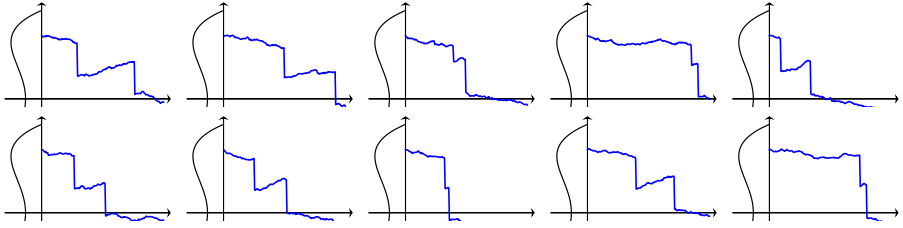


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

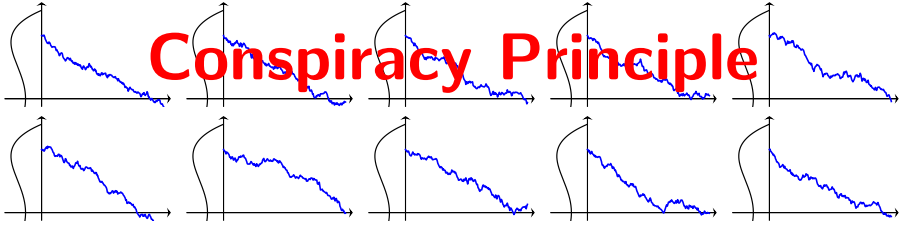


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$

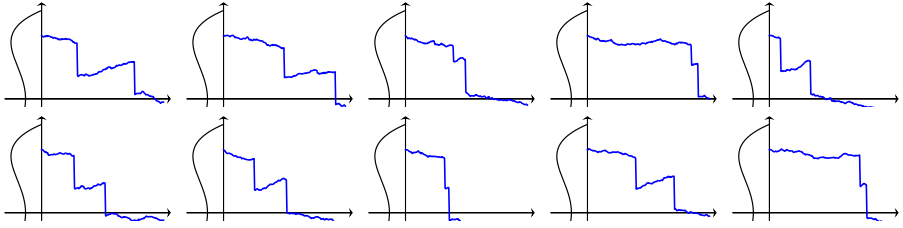


SGD's Escaping Route under Gradient Clipping with $\mathcal{J}(A) = 2$

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

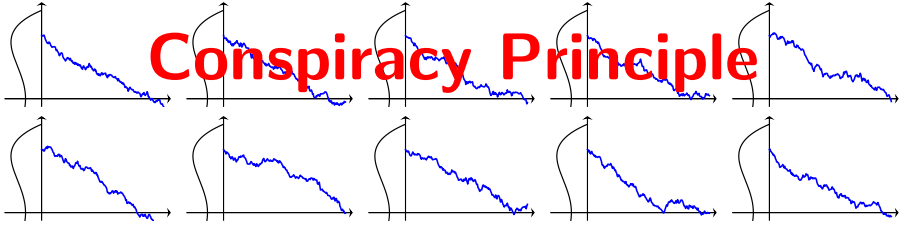


Trajectory of SGD X^η conditional on exit: **heavy-tailed** noises with $\eta = 1/200$



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

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



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