An Introduction to Heavy Tails for ML Researchers

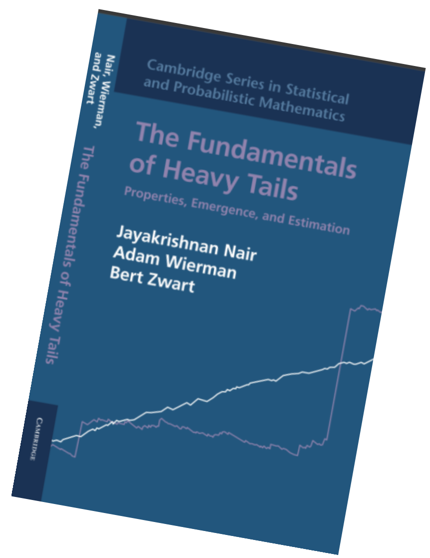
Conspiracies, Catastrophes, and the Principle of a Single Big Jump

Adam Wierman, Caltech

An Introduction to Heavy Tails for ML Researchers

Conspiracies, Catastrophes, and the Principle of a Single Big Jump

Material based on:



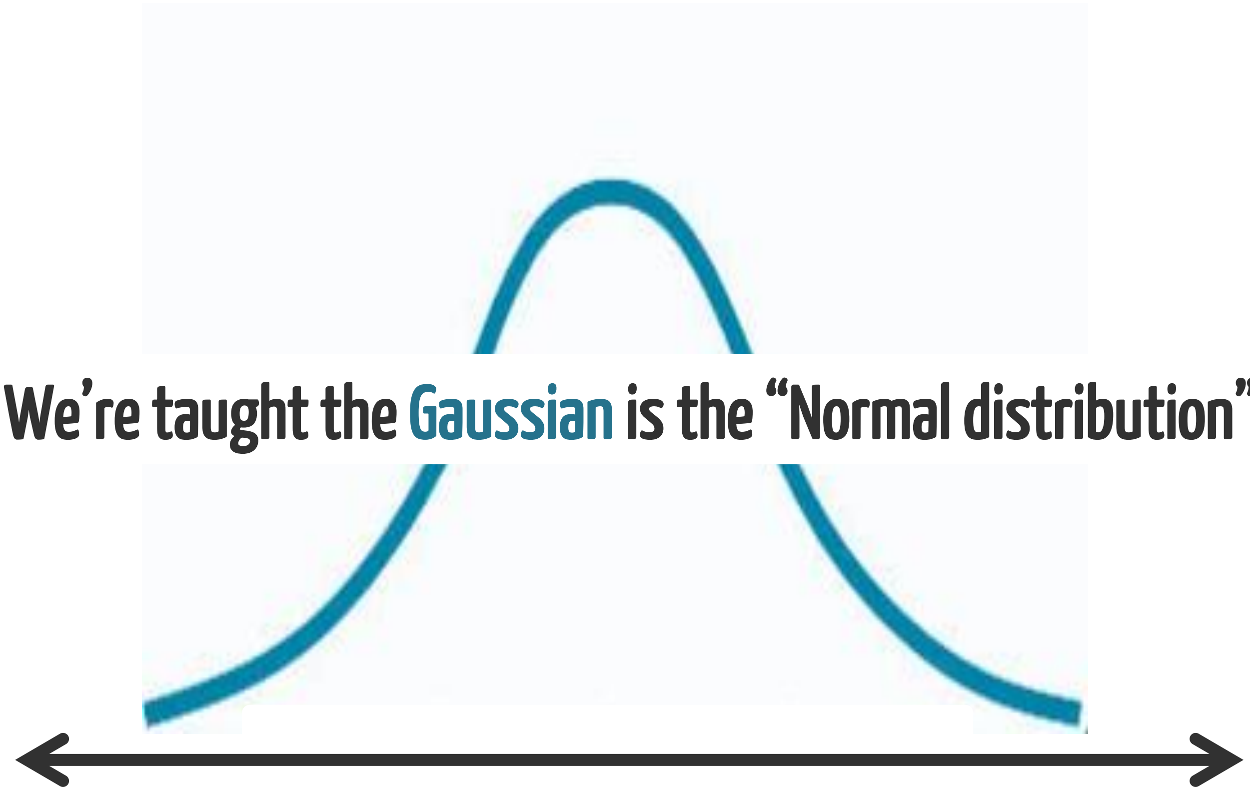
“The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking [...] as if anything but a bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world.”

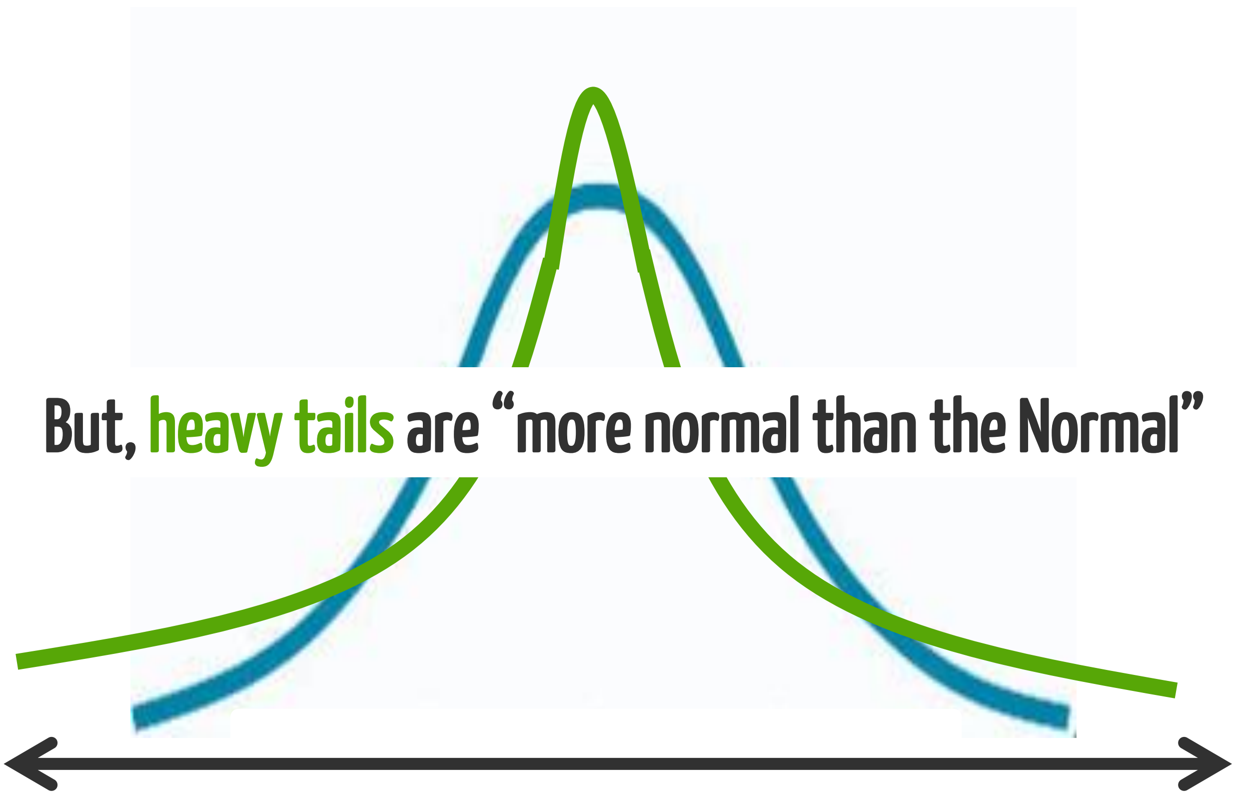
- Clay Shirky in Newsweek (2011)

"The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking [...] as if anything but a bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents ~~us~~ ML researchers from thinking clearly about ~~the world~~ algorithms."

- Clay Shirky in Newsweek (2011)

We’re taught the Gaussian is the “Normal distribution”





But, heavy tails are “more normal than the Normal”

In ML/AI, heavy tails are common in inputs to models & created by core algorithms like SGD.

10:30 - 11:10	**Invited talk:** Liam Hodgkinson	Overparameterization and the Power Law Paradigm
11:10 - 11:35	**Contributed talk:** Vivien Cabannes	Associative Memories with Heavy-Tailed Data
11:35 - 12:00	**Contributed talk:** Dominique Perrault-Joncas	Meta-Analysis of Randomized Experiments
12:00 - 14:00	Lunch break	--
14:00 - 14:40	**Invited talk:** Nisha Chandramoorthy	A dynamical View of Learners, Samplers and Forecasters
14:40 - 15:05	**Contributed talk:** Jeremy Cohen	Adaptive Gradient Methods at the Edge of Stability
15:05 - 15:30	Coffee break	--
15:30 - 16:10	**Invited talk:** Charles Martin	Heavy-Tailed Self -Self-Regularization in DNNs

In ML/AI, heavy tails are common in inputs to models
& created by core algorithms like SGD.

...yet many ML algorithms are designed and analyzed using
intuition and tools based on light-tailed assumptions.

Heavy-tailed phenomena are typically treated as something

MYSTERIOUS, Surprising, & Controversial

Our intuition is flawed because intro probability classes treat heavy-tails as curiosities

Simple, appealing statistical approaches for estimating them have BIG problems

An historic example:

Networking “discovers” heavy tails (early 2000s)

On Power-Law Relationships of the Internet To[...]

Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
michalis@cs.ucr.edu

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
pfal@cs.toronto.edu

Chri[...]
Carne[...]
Dept. [...]
chri[...]

Self-Similarity in World Wide Web Traffic
Evidence and Possible Causes*

M[...] Crovella and Azer Bestavros
[...]uter Science Department
Boston University
Boston, MA 02215
[...]lla,best}@cs.bu.edu

The Effect of Heavy-Tailed Job Size Distributions on Computer System Design.

Mor Harchol-Balter*
Laboratory for Computer Science
ha[...]

Power-Law Distribution of the World Wide Web

Barabási and Albert (*1*) propose an improved version of the Erdös-Rényi (ER) theory of random networks to account for the [...] number of systems, [...]

from other sites, and found that the distribution of links followed a power law (Fig. 1A). Next, we queried the InterNIC database (using the WHOIS search tool at www.[...]orksolutions.com) for the date on which [...]gistered. Whereas [...]at older sites have [...] and gather links at

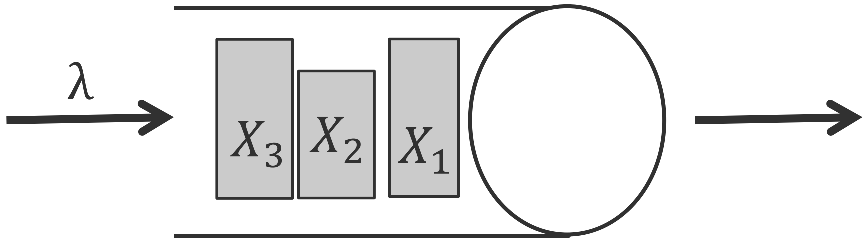
data, we can illustrate the same procedure for the network of movie actors that we discussed (*1*). When the connectivity of the individual actors is plotted as a function of the release year of their first movie (Fig. 1A), the results are very similar to those shown in fig. 1B of Adamic and Huberman's comment. The only difference is that the movie industry had its boom not 4 years ago, as did the WWW, but rather at the beginning of the century; thus, the apparently structureless regime persists much longer. When the connectivity of the actors that debuted in the same year is averaged, however, the average con-

Heavy-Tailed Probability Distributions in the World Wide Web

Mark E. Crovella, Murad S. Taqqu and Azer Bestavros 1 2 3

The existence of heavy-tails required rethinking network design & communication protocols

Example: Scheduling



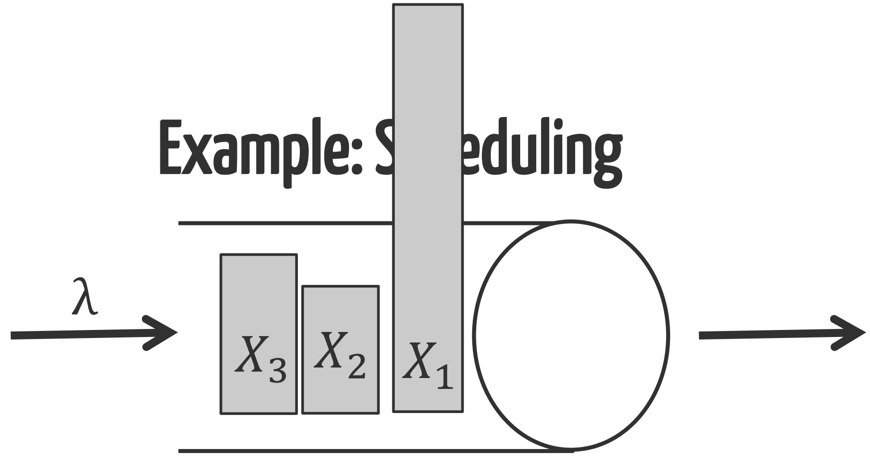
Load: $\rho = \lambda E[X] < 1$

What order should jobs be served in to minimize $\Pr(Delay > t)$ for large t?

Light-tailed → **FIFO is optimal**

Heavy-tailed → **SRPT is optimal**

Example: Scheduling



Load: $\rho = \lambda E[X] < 1$

What order should jobs be served in to minimize $\Pr(Delay > t)$ for large t?

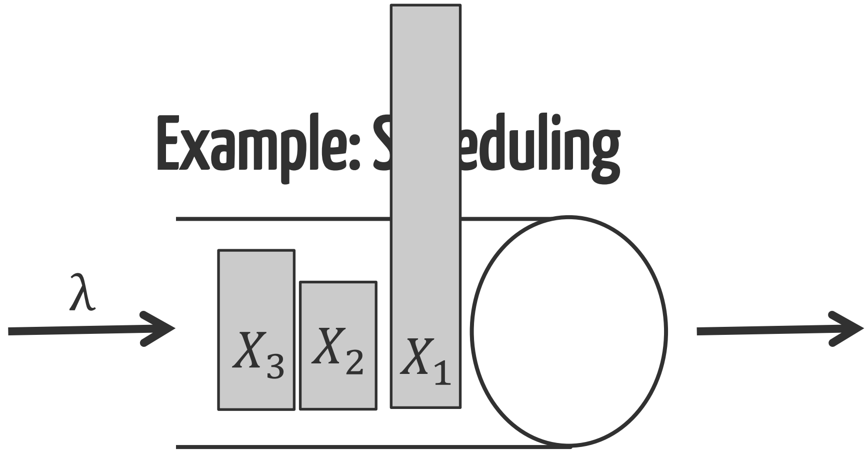
Light-tailed

Heavy-tailed

FIFO is optimal

SRPT is optimal

Example: Scheduling



Load: $\rho = \lambda E[X] < 1$

What order should jobs be served in to minimize $\Pr(Delay > t)$ for large t?

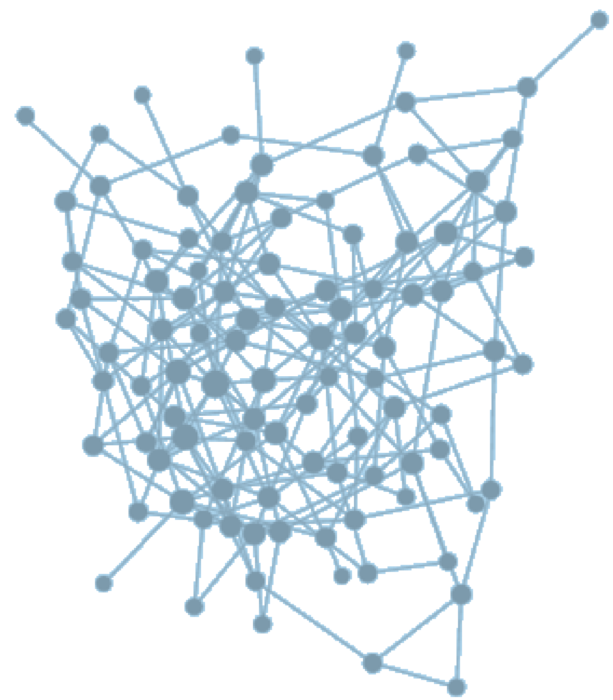
Home > Operations Research > Vol. 60, No. 5 >

Is Tail-Optimal Scheduling Possible?

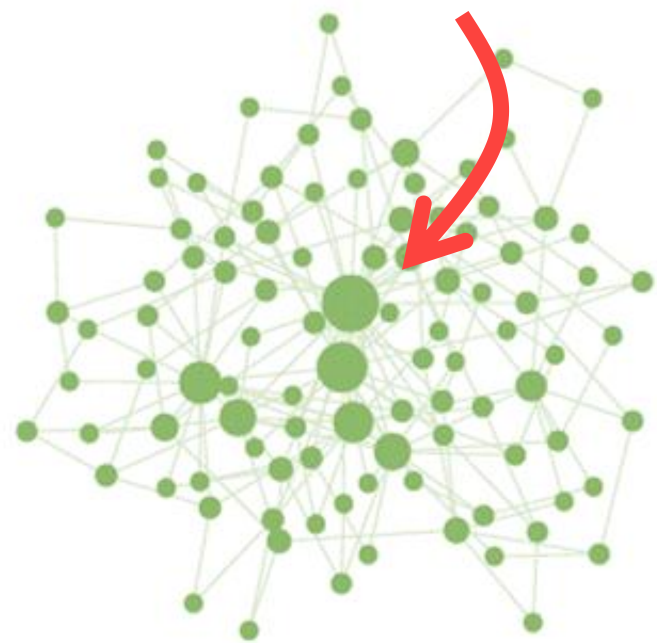
Adam Wierman, Bert Zwart

Published Online: 9 Oct 2012 | https://doi.org/10.1287/opre.1120.1086

Example: The fragility of the internet

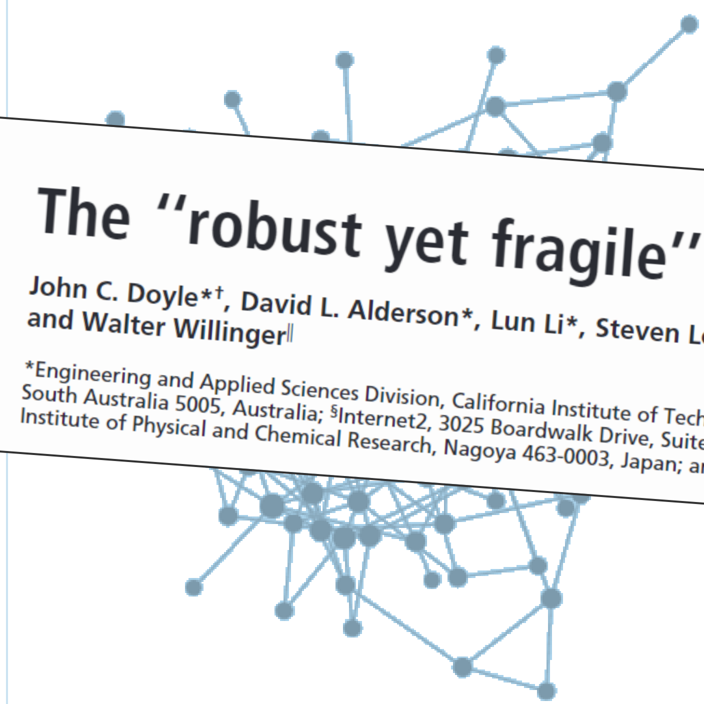


Light-tailed Degrees



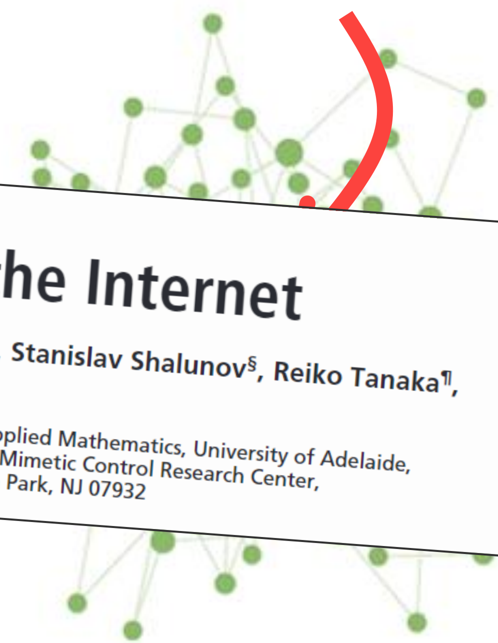
Heavy-tailed Degrees

Example: The fragility of the internet



Light-tailed Degrees

Hubs are vulnerabilities?



Heavy-tailed Degrees

The "robust yet fragile" nature of the Internet

John C. Doyle[*†], David L. Alderson[*], Lun Li[*], Steven Low[*], Matthew Roughan[‡], Stanislav Shalunov[§], Reiko Tanaka[¶], and Walter Willinger[‖]

[*]Engineering and Applied Sciences Division, California Institute of Technology, Pasadena, CA 91125; [‡]Applied Mathematics, University of Adelaide, South Australia 5005, Australia; [§]Internet2, 3025 Boardwalk Drive, Suite 200, Ann Arbor, MI 48108; [¶]Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research, Nagoya 463-0003, Japan; and [‖]AT&T Labs–Research, Florham Park, NJ 07932

But all was not as it seemed...

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
michalis@cs.ucr.edu

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
pfal@cs.toronto.edu

Christos Faloutsos *
Carnegie Mellon Univ.
Dept. of Comp. Science
christos@cs.cmu.edu

On the Bias of Traceroute Sampling

or, Power-law Degree Distributions in Regular Graphs

Dimitris Achlioptas
Microsoft Research
Microsoft Corporation
Redmond, WA 98052
optas@microsoft.com

Aaron Clauset
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
aaron@cs.unm.edu

David Kempe
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
dkempe@usc.edu

Cristopher Moore
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
moore@cs.unm.edu

IEEE/ACM TRANSACTIONS ON NET... VOL. 13, NO. 6, DECEMBER 2005 1205

Understanding Internet Topology: Principles, Models, and Validation

David Alderson, *Member, IEEE*, Lun Li, *Student Member, IEEE*, Walter Willinger, *Fellow, IEEE*, and John C. Doyle, *Member, IEEE*

There are similar stories in power networks, social networks, biology, astronomy, chemistry,...

Many of the controversies are still not resolved...

Scale-free networks are rare

Anna D. Broido[1,*] and Aaron Clauset[2,3,4,†]
[1] Department of Applied Mathematics, University of Colorado, Boulder, CO, USA
[2] Department of Computer Science, University of Colorado, Boulder, CO, USA
[3] BioFrontiers Institute, University of Co
[4] Santa Fe Institute, Santa

A central claim in modern network science is that re
meaning that the fraction of nodes with degree k follow
$2 < \alpha < 3$. However, empirical evidence for this beli
real-world networks. We test the universality of sca
statistical tools to a large corpus of nearly 1000 net
technological, and informational sources. We fit the
test its statistical plausibility, and compare it via a
free models, e.g., the log-normal. Across domains,
only 4% exhibiting the strongest-possible evidence
weakest-possible evidence. Furthermore, evidence of
across sources: social networks are at best weakly
biological networks can be called strongly scale fr
scale-free networks and reveal that real-world net
likely require new ideas and mechanisms to expla

Scale-Free Networks Well Done

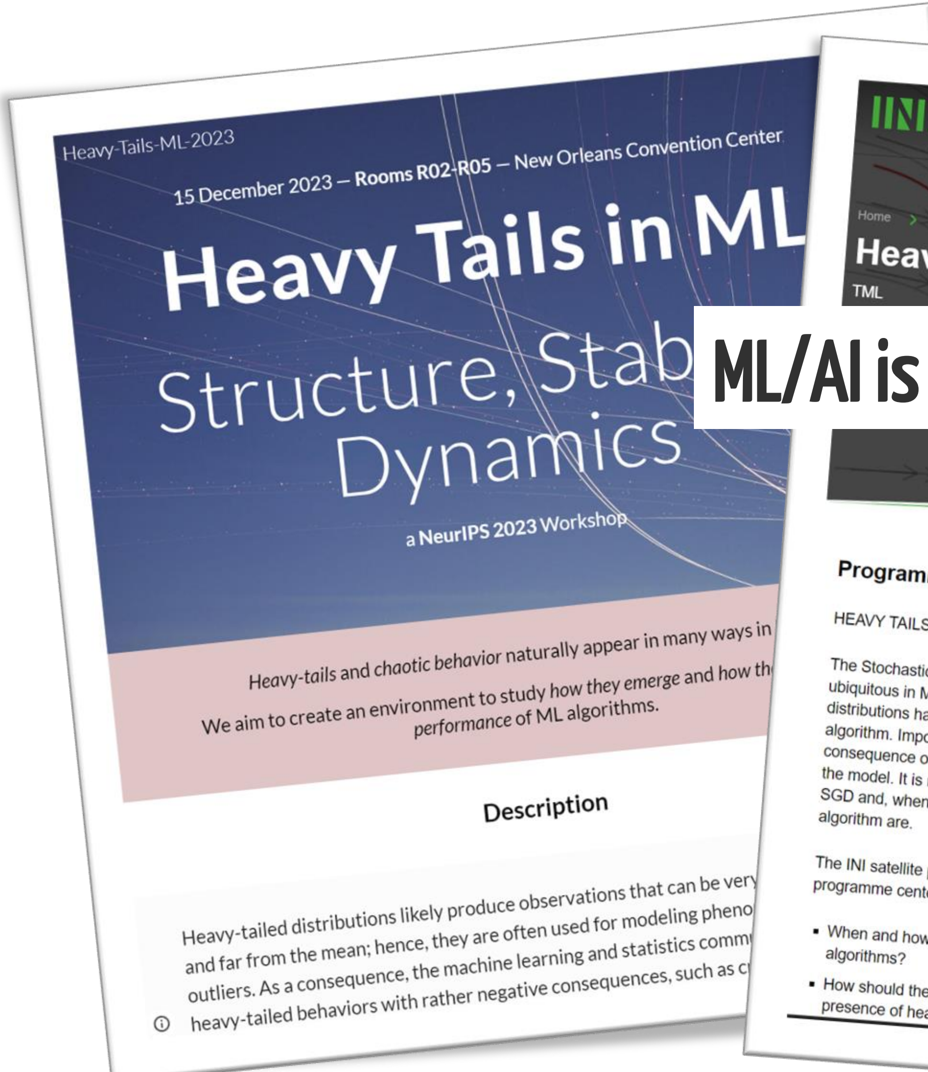
Ivan Voitalov,[1,2] Pim van der Hoorn,[1,2] Remco van der Hofstad,[3] and Dmitri Krioukov[1,2,4,5]
[1] Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA
[2] Network Science Institute, Northeastern University, Boston, Massachusetts 02115, USA
[3] Department of Mathematics and Computer Science, Eindhoven University of Technology, Postbus 513, 5600 MB Eindhoven, Netherlands
[4] Department of Mathematics, Northeastern University, Boston, Massachusetts 02115, USA
[5] Department of Electrical & Computer Engineering, Northeastern University, Boston, Massachusetts 02115, USA

We bring rigor to the vibrant activity of detecting power laws in empirical degree distributions in real-world networks. We first provide a rigorous definition of power-law distributions, equivalent to the definition of regularly varying distributions that are widely used in statistics and other fields. This definition allows the distribution to deviate from a pure power law arbitrarily but without affecting the power-law tail exponent. We then identify three estimators of these exponents that are proven to be statistically consistent—that is, converging to the true value of the exponent for any regularly varying distribution—and that satisfy some additional niceness requirements. In contrast to estimators that are currently popular in network science, the estimators considered here are based on fundamental results in extreme value theory, and so are the proofs of their consistency. Finally, we apply these estimators to a representative collection of synthetic and real-world data. According to their estimates, real-world scale-free networks are definitely not as rare as one would conclude based on the popular but unrealistic assumption that real-world data comes from power laws of pristine purity, void of noise and deviations.

All of this has happened before.
All of this will happen again.

- Battlestar Galactica
(orig. from Peter Pan by J.M. Barrie)

ML/AI is next!


Heavy-Tails-ML-2023

15 December 2023 — Rooms R02-R05 — New Orleans Convention Center

Heavy Tails in ML

Structure, Stab[...] Dynamics

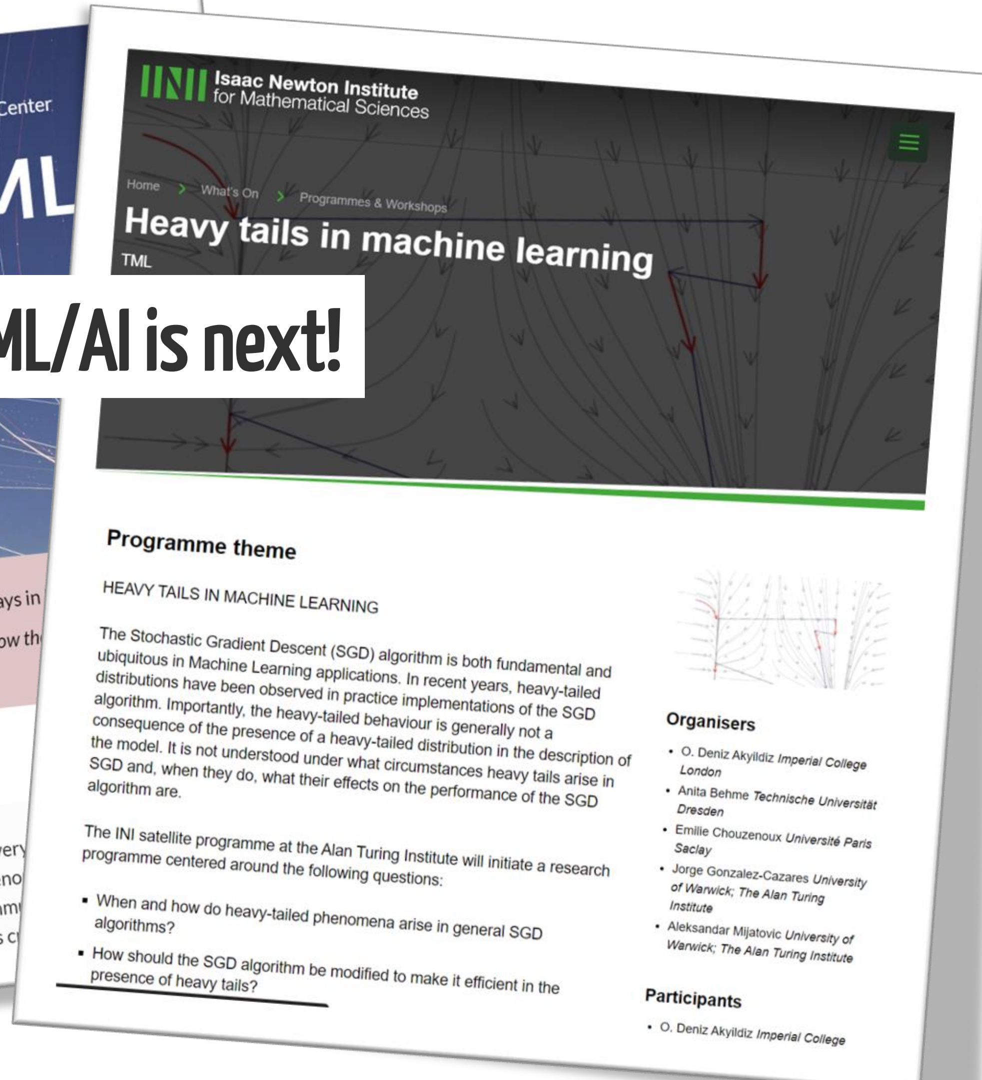
a NeurIPS 2023 Workshop

Heavy-tails and *chaotic behavior* naturally appear in many ways in [...]

We aim to create an environment to study *how they emerge* and how th[...] *performance* of ML algorithms.

Description

Heavy-tailed distributions likely produce observations that can be very [...] and far from the mean; hence, they are often used for modeling pheno[...] outliers. As a consequence, the machine learning and statistics comm[...] heavy-tailed behaviors with rather negative consequences, such as c[...]


Isaac Newton Institute for Mathematical Sciences

Home > What's On > Programmes & Workshops

Heavy tails in machine learning

TML

Programme theme

HEAVY TAILS IN MACHINE LEARNING

The Stochastic Gradient Descent (SGD) algorithm is both fundamental and ubiquitous in Machine Learning applications. In recent years, heavy-tailed distributions have been observed in practice implementations of the SGD algorithm. Importantly, the heavy-tailed behaviour is generally not a consequence of the presence of a heavy-tailed distribution in the description of the model. It is not understood under what circumstances heavy tails arise in SGD and, when they do, what their effects on the performance of the SGD algorithm are.

The INI satellite programme at the Alan Turing Institute will initiate a research programme centered around the following questions:

- When and how do heavy-tailed phenomena arise in general SGD algorithms?
- How should the SGD algorithm be modified to make it efficient in the presence of heavy tails?

Organisers

- O. Deniz Akyildiz *Imperial College London*
- Anita Behme *Technische Universität Dresden*
- Emilie Chouzenoux *Université Paris Saclay*
- Jorge Gonzalez-Cazares *University of Warwick; The Alan Turing Institute*
- Aleksandar Mijatovic *University of Warwick; The Alan Turing Institute*

Participants

- O. Deniz Akyildiz *Imperial College*

Heavy-tailed phenomena are typically treated as something

MYSTERIOUS, Surprising, & Controversial

Our intuition is flawed because intro probability classes focus on light-tailed distributions

Simple, appealing statistical approaches for estimating them have BIG problems

Heavy-tailed phenomena are typically treated as something

~~Mysterious~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

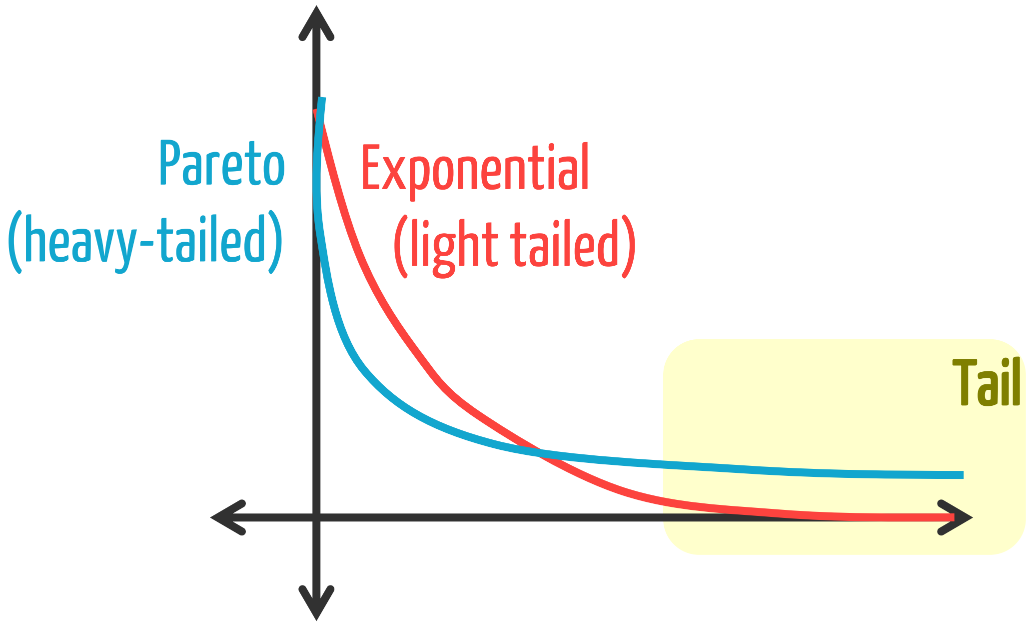
3. Identification

(We won't do this one today)

4. Algorithm Design

Definition: A random variable is heavy-tailed iff $\forall s > 0$,

$$\lim_{x \to \infty} e^{sx} \Pr(X > x) = \infty$$



Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

$$\Pr(X > x) = \bar{F}(x) = \left(\frac{x_{\min}}{x}\right)^{\alpha} \quad \text{for } x \geq x_{\min}$$

p.d.f: $f(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}}$

Notice: $Var[X] = \infty$ if $\alpha < 2$!

Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student's t, Frechet, ...

Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student's t, Frechet, ...

$$X:\ \log X \sim Normal$$

$$Var[X] = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}$$

Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student's t, Frechet, ...

$$\bar{F}(x) = e^{-(x/\lambda)^k}$$

$$\boldsymbol{k < 1: Heavy - tailed}$$

$$k = 1: Exponential$$

$$k = 3.4: Approx\ Normal$$

$$k \to \infty: Deterministic$$

Heavy-tailed distributions have many strange & beautiful properties

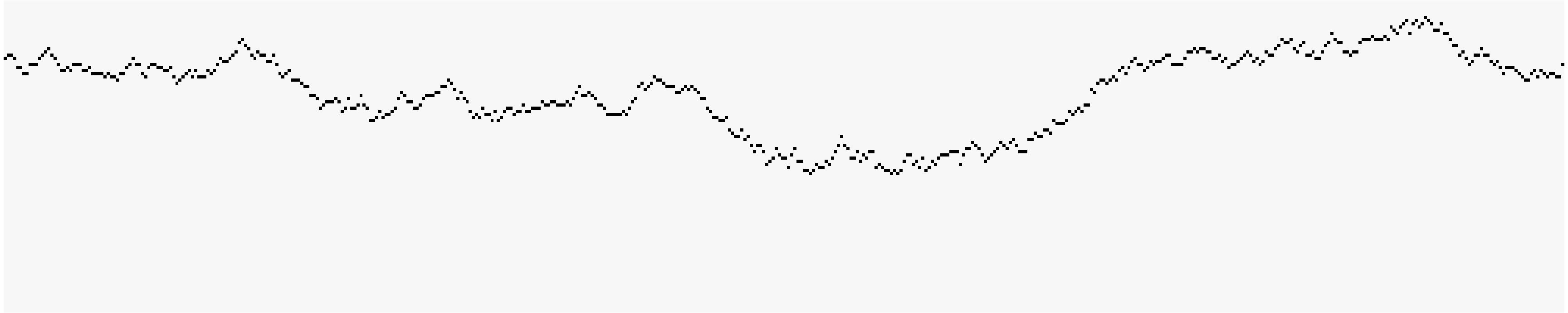
- The "Pareto principle" (e.g. 80% of the wealth owned by 20% of the population)
- Infinite variance or even infinite mean
- Outliers that are much larger than the mean happen "frequently"

....

These are driven by 3 "defining" properties

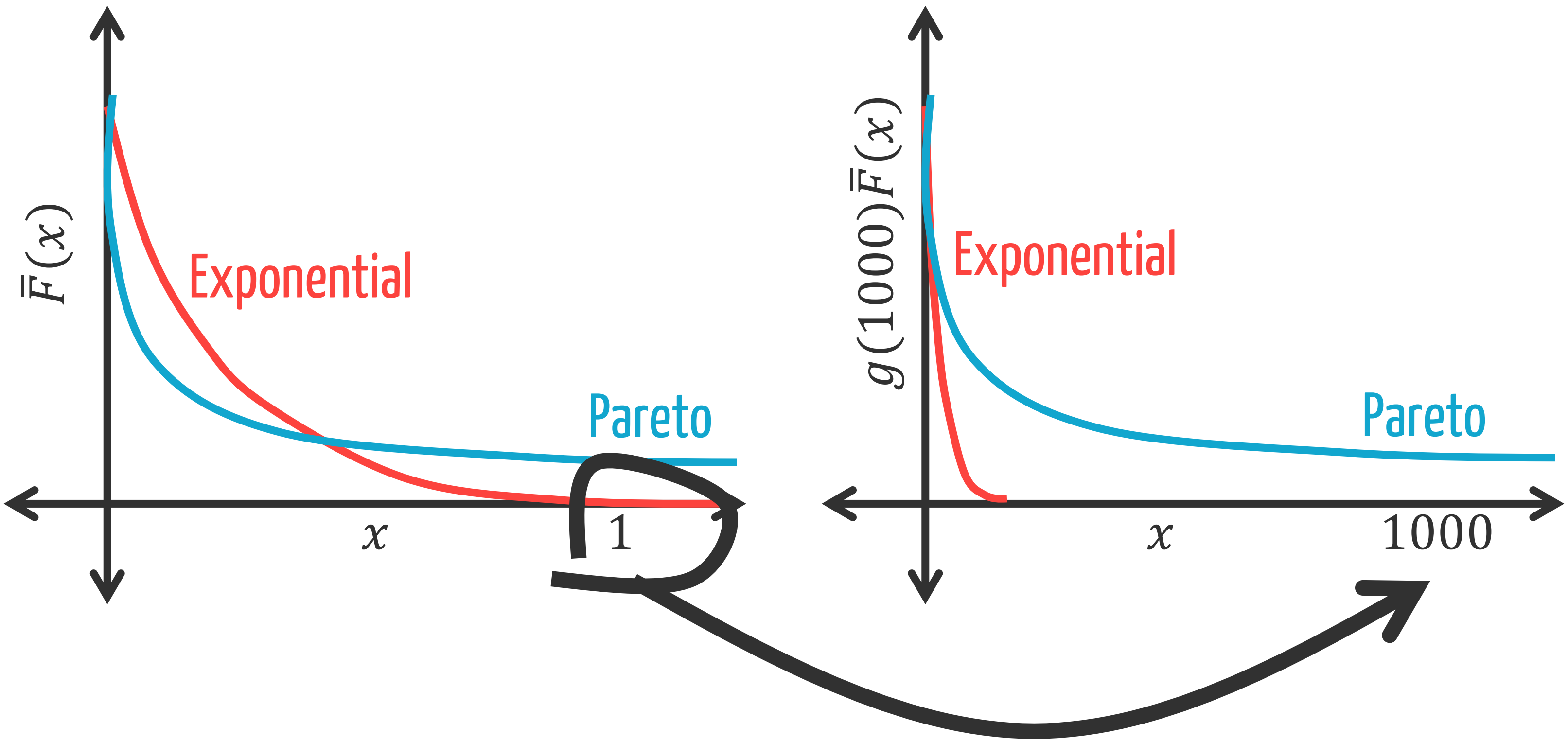
1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up" (see the book!)

Scale invariance



Scale invariance

F is scale invariant if there exists an x_0 and a g such that $\boxed{\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)}$ for all λ, x such that $\lambda x \geq x_0$.



Scale invariance

F is scale invariant if there exists an x_0 and a g such that $\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)$ for all λ, x such that $\lambda x \geq x_0$.

<u>Theorem</u>: A distribution is scale invariant if and only if it is Pareto.

<u>Example</u>: Pareto distributions

$$\bar{F}(\lambda x) = \left(\frac{x_{\min}}{\lambda x}\right)^{\alpha} = \bar{F}(x)\left(\frac{1}{\lambda}\right)^{\alpha}$$

Scale invariance

F is scale invariant if there exists an x_0 and a g such that
$\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)$ for all λ, x such that $\lambda x \geq x_0$.

Asymptotic scale invariance

F is asymptotically scale invariant if there exists a continuous, finite g such that
$\lim_{x \to \infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = g(\lambda)$ for all λ.

Example: Regularly varying distributions

F is regularly varying if $\bar{F}(x) = x^{-\rho} L(x)$, where $L(x)$ is slowly varying,
i.e., $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.

$\Updownarrow$

Theorem: A distribution is asymptotically scale invariant iff it is regularly varying.

Asymptotic scale invariance

F is asymptotically scale invariant if there exists a continuous, finite g such that
$\lim_{x\to\infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = g(\lambda)$ for all λ.

Example: Regularly varying distributions

F is regularly varying if $\bar{F}(x) = x^{-\rho} L(x)$, where $L(x)$ is slowly varying,
i.e., $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.

Regularly varying distributions are extremely easy to work with analytically.
They behave like Pareto distributions with respect to the tail.
→ "Karamata" theorems
→ "Tauberian" theorems

Heavy-tailed distributions have many strange & beautiful properties

- The "Pareto principle" (e.g. 80% of the wealth owned by 20% of the population)
- Infinite variance or even infinite mean
- Outliers that are much larger than the mean happen "frequently"

....

These are driven by 3 "defining" properties

1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up" (see the book!)

A thought experiment

Suppose that during lecture I polled 50 students about their heights and the number of instagram followers they have...

The sum of the heights was ~300 feet.
The sum of the number of instagram followers was 1,025,000

What led to these large values?

A thought experiment

Suppose that during lecture I polled 50 students about their heights and the number of instagram followers they have...

The sum of the heights was ~300 feet.
The sum of the number of instagram followers was 1,025,000

8'3" (2.5m)

614 million followers

A thought experiment

Suppose that during lecture I polled 50 students about their heights and the number of instagram followers they have...

The sum of the heights was ~300 feet.
The sum of the number of instagram followers was 1,025,000

A bunch of people were probably just over 6' tall
(Maybe the basketball teams were in the class.)

"Conspiracy principle"

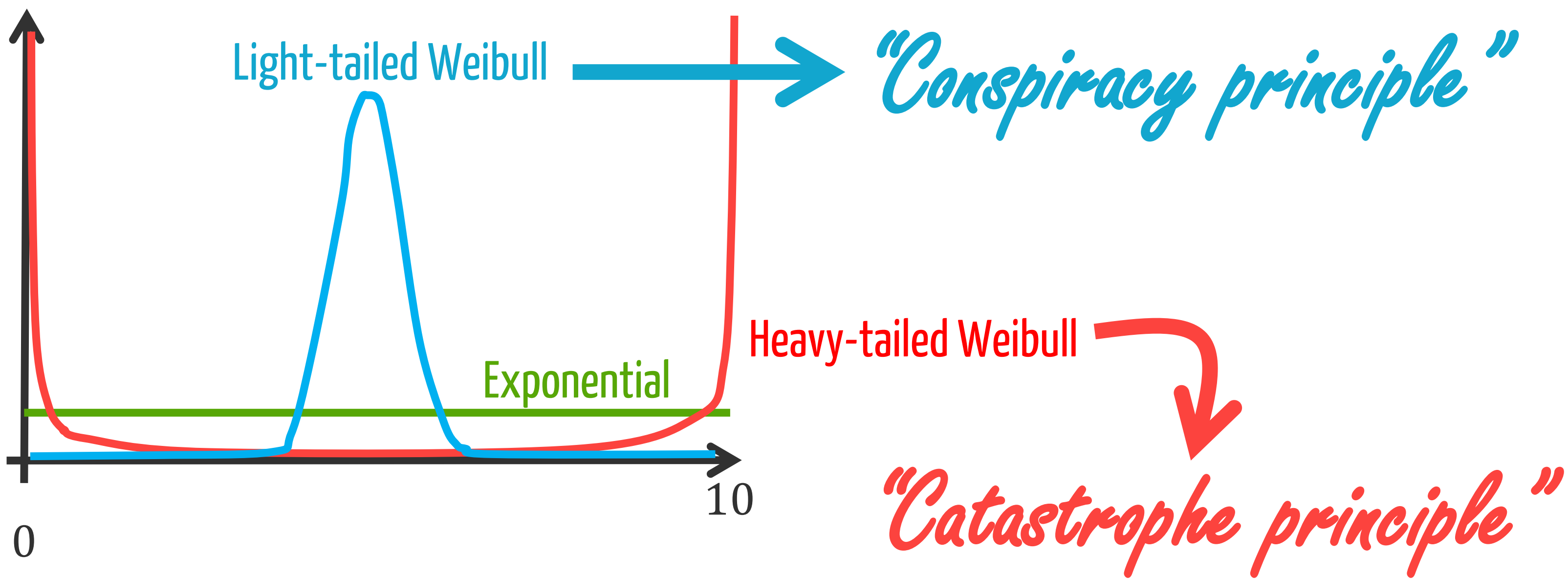
One person was probably a social media celebrity and had ~1 million followers.

"Catastrophe principle"

Example

Consider X_1, X_2 i.i.d. Weibull with mean 1.

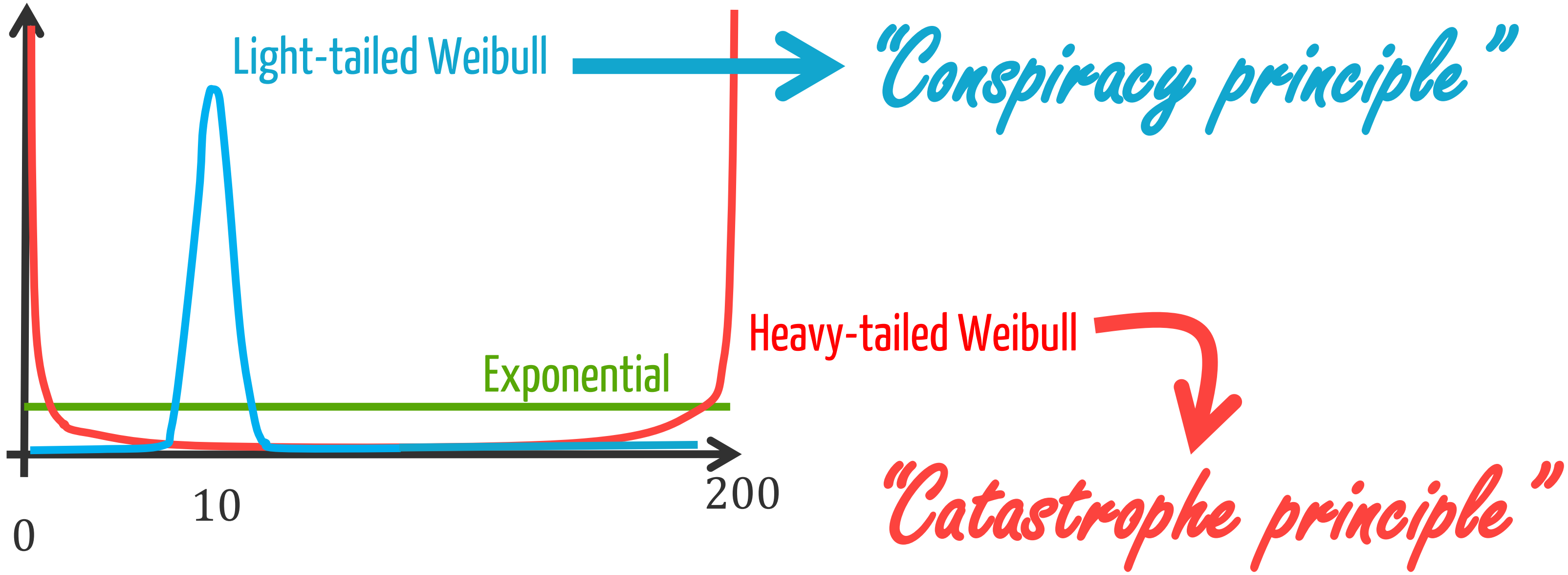
Given the rare event $X_1 + X_2 = 10$, what is the marginal density of X_1?



Example

Consider $X_1, \ldots, X_{20}$ i.i.d. Weibull with mean 1.

Given the rare event $X_1 + \cdots + X_{20} = 200$, what is the marginal density of X_1?



"Catastrophe principle"

$$\Pr(\max(X_1, \dots, X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$

$$\Rightarrow \Pr(\max(X_1, \dots, X_n) > t \mid X_1 + \cdots + X_n > t) \to 1$$

"Conspiracy principle"

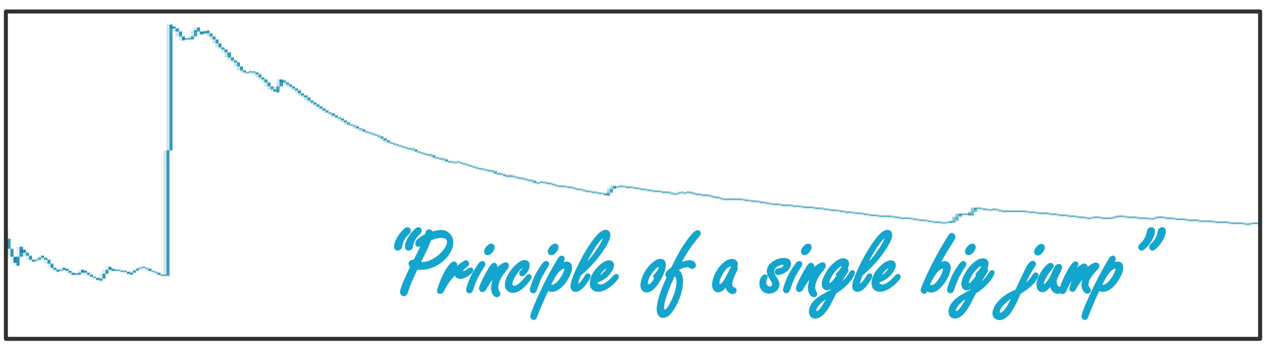
$$\Pr(\max(X_1, \dots, X_n) > t) = o(\Pr(X_1 + \cdots + X_n > t))$$

"Catastrophe principle"

$$\Pr(\max(X_1,\ldots,X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$

$$\Rightarrow \Pr(\max(X_1,\ldots,X_n) > t \mid X_1 + \cdots + X_n > t) \to 1$$

Extremely useful for analyzing random walks, MDPs, …



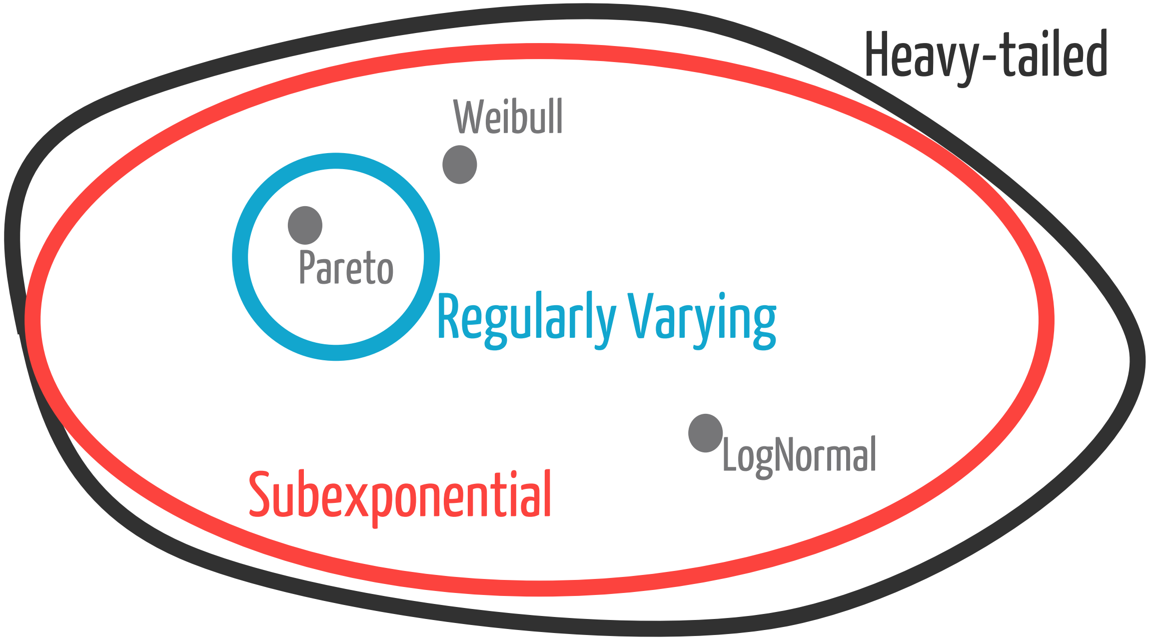
"Catastrophe principle"

$$\Pr(\max(X_1, \dots, X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$
$$\Rightarrow \Pr(\max(X_1, \dots, X_n) > t \mid X_1 + \cdots + X_n > t) \to 1$$

$$\Updownarrow$$

Subexponential distributions

F is subexponential if for i.i.d. X_i, $\Pr(X_1 + \cdots + X_n > t) \sim nPr(X_1 > t)$



Subexponential distributions

F is subexponential if for i.i.d. X_i, $\Pr(X_1 + \cdots + X_n > t) \sim nPr(X_1 > t)$

Heavy-tailed distributions have many strange & beautiful properties

- The "Pareto principle" (e.g. 80% of the wealth owned by 20% of the population)
- Infinite variance or even infinite mean
- Outliers that are much larger than the mean happen "frequently"

....

These are driven by 3 "defining" properties

1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up" (see the book!)

Heavy-tailed phenomena are treated as something

~~Mysterious~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

(We won't do this one today)

4. Algorithm Design

Where do heavy-tails come from in ML/AI applications?

Option 1. They come from the data (and costs) in the applications.

Option 2. They are created by the algorithms we use.

A toy example: SGD*

$$\min_{x \in \mathbb{R}} f(x)$$

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = X_k - \eta(2A_k X_k + B_k)$$

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x\in\mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = \underbrace{X_k - \eta(2A_k}_{C_k}X_k + B_k)$$

*This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E\left[Ax^2 + Bx\right]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + \eta B_k$$

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + \underbrace{\eta B_k}_{D_k}$$

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + D_k$$

This process is both multiplicative and additive.

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

We know a lot about additive processes!

We've all been taught that the Gaussian is "normal" because of the Central limit theorem

But the Central Limit Theorem
we're taught in intro probability is not complete!

A quick review

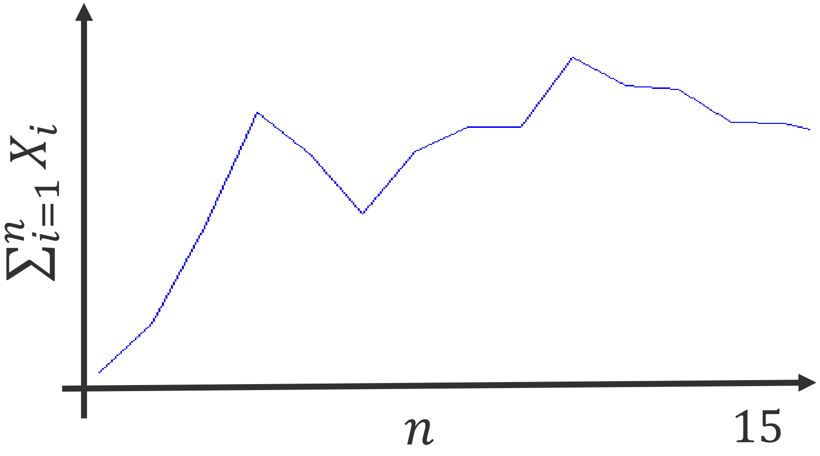
Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Law of Large Numbers (LLN): $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_i]\ a.s.$ when $E[X_i] < \infty$

A quick review

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

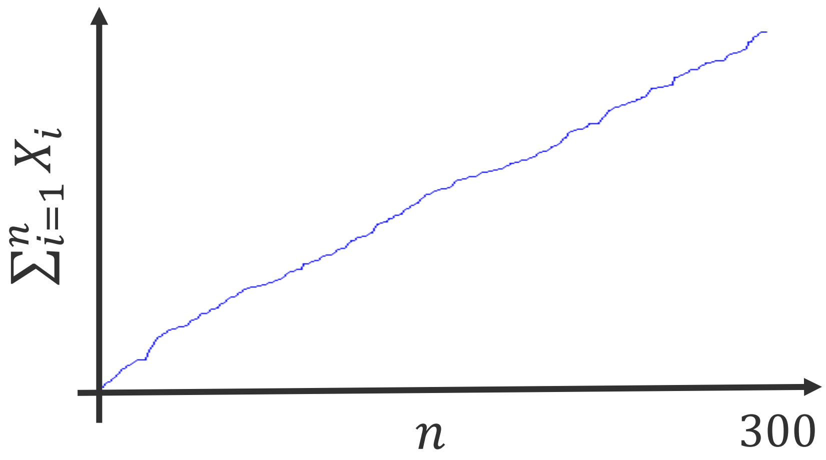
Law of Large Numbers (LLN): $\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$



A quick review

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Law of Large Numbers (LLN): $\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$



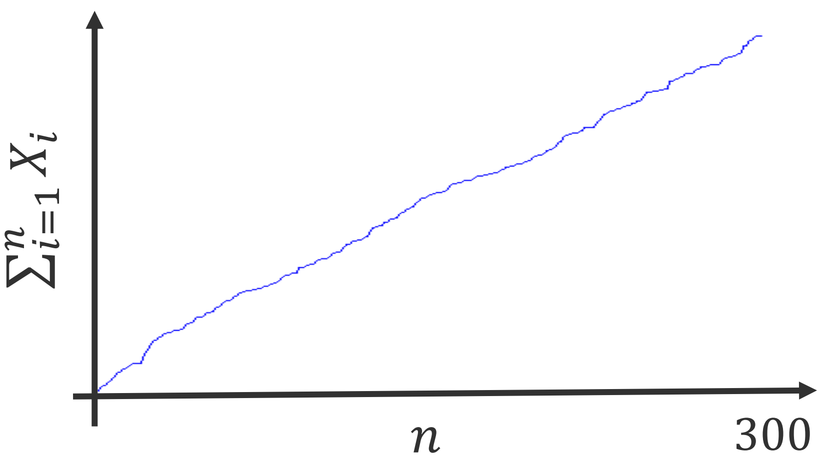
A quick review

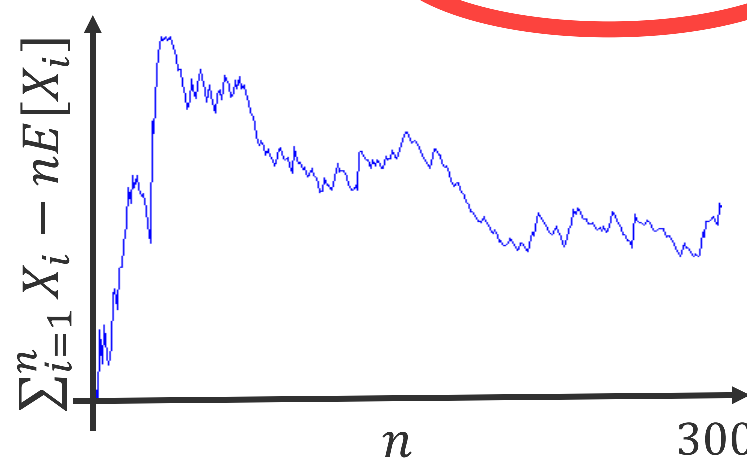
Consider i.i.d. X_i. How does $\sum_{i=}^{n}$ **What if $Var[X_i] = \infty$?**

Law of Large Numbers (LLN): $\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$

Central Limit Theorem (CLT): $\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$

where $Z \sim Normal(0, \sigma^2)$ with $\text{Var}[X_i] = \sigma^2 < \infty$.





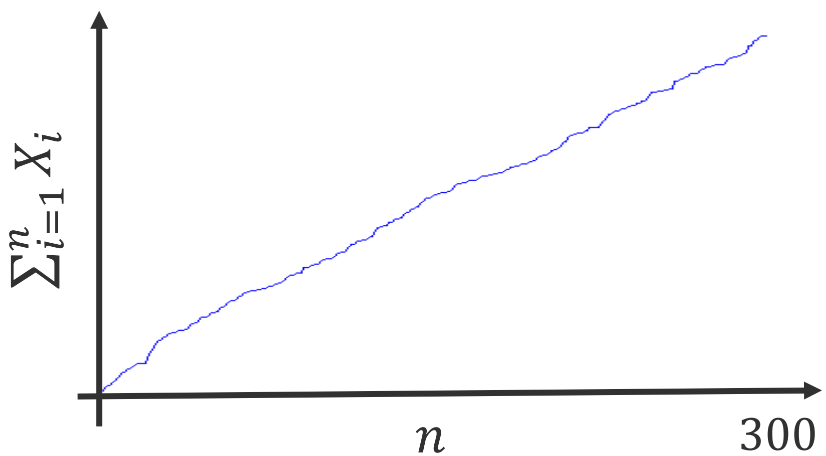
A quick review

Consider i.i.d. X_i. How does $\sum_{i}^{n}$ **What if $Var[X_i] = \infty$?**

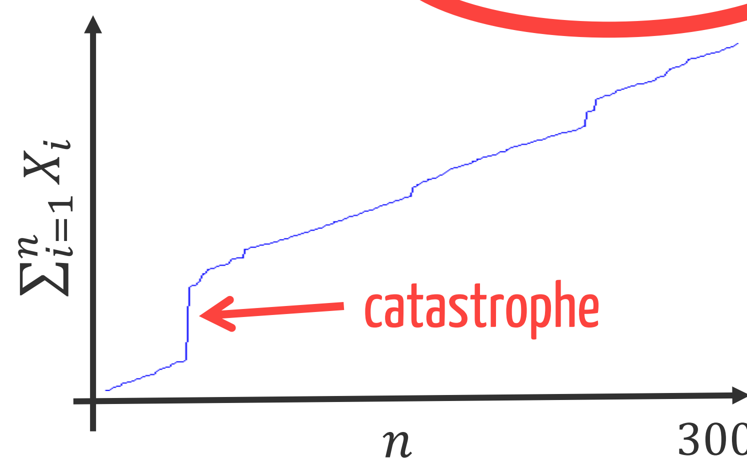
Law of Large Numbers (LLN): $\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$

Central Limit Theorem (CLT): $\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$

where $Z \sim Normal(0, \sigma^2)$ with $Var[X_i] = \sigma^2 < \infty$.



$\boldsymbol{Var[X_i] < \infty}$



$\boldsymbol{Var[X_i] = \infty}$

A quick review

Consider i.i.d. X_i. How does $\sum_{i=1}^{n}$

What if $Var[X_i] = \infty$?

Law of Large Numbers (LLN): $\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$

Central Limit Theorem (CLT): $\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$

where $Z \sim Normal(0, \sigma^2)$ with $Var[X_i] = \sigma^2 < \infty$.

The Generalized Central Limit Theorem (GCLT):

$$\sum_{i=1}^{n} X_i = nE[X_i] + n^{1/\alpha}Z + o(n^{1/\alpha})$$

Finite variance → Light-tailed (Normal)

Infinite variance → Heavy-tailed (power law)

Additive processes can lead to heavy-tails, depending on the input.

What about **multiplicative** processes?

Multiplicative processes almost always lead to heavy tails

An example:

$$Y_1, Y_2 \sim Exponential(\mu)$$

$$\Pr(Y_1 \cdot Y_2 > x) \geq \Pr\left(Y_1 > \sqrt{x}\right)^2$$
$$= e^{-2\mu\sqrt{x}}$$

$\Rightarrow Y_1 \cdot Y_2$ is heavy-tailed!

Multiplicative processes almost always lead to heavy tails

$$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$$

$$\log P_n = \log Y_1 + \log Y_2 + \cdots + \underbrace{\log Y_n}_{X_n}$$

Central Limit Theorem

$$\log P_n = n\, E[X_i] + \sqrt{n} Z + o\left(\sqrt{n}\right), \text{ where } Z \sim Normal(0, \sigma^2)$$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\left(\frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu}\right)^{1/\sqrt{n}} \rightarrow \boldsymbol{H \sim LogNormal(0, \sigma^2)}$$

where $\mu = e^{E[\log Y_i]}$

and $\mathrm{Var}[\log Y_i] = \sigma^2 < \infty$.

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + D_k$$

This process is both multiplicative and additive.

* This is a very simplistic version. Later talks/posters in this workshop will give more detailed treatments!

A toy example: SGD*

$$\min_{x \in \mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + D_k$$

This process is both multiplicative and additive.

Under minor technical conditions, $X_k \to F$ such that

$$\lim_{x \to \infty} \frac{\log \bar{F}(x)}{\log x} = s^* \text{ where } s^* = \sup(s \geq 0 | E[X_k^s] \leq 1)$$

regularly varying → SGD leads to heavy tails, even when A and B are light tailed!

A toy example: SGD*

$$\min_{x\in\mathbb{R}} E[Ax^2 + Bx]$$

Assuming $E[A] > 0$ and gradients are available, SGD with learning rate η follows

$$X_{k+1} = C_k X_k + D_k$$

This process is both multiplicative and additive.

Under minor technical conditions, $X_k \to F$ such that

$\boxed{\lim_{x\to\infty} \frac{\log \bar{F}(x)}{\log x} = s^*}$ where $s^* = \sup(s \geq 0 | E[X_k^s] \leq 1)$

regularly varying → SGD leads to heavy tails, even when A and B are light tailed, & this leads to better generalization too!

Heavy-tailed phenomena are treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

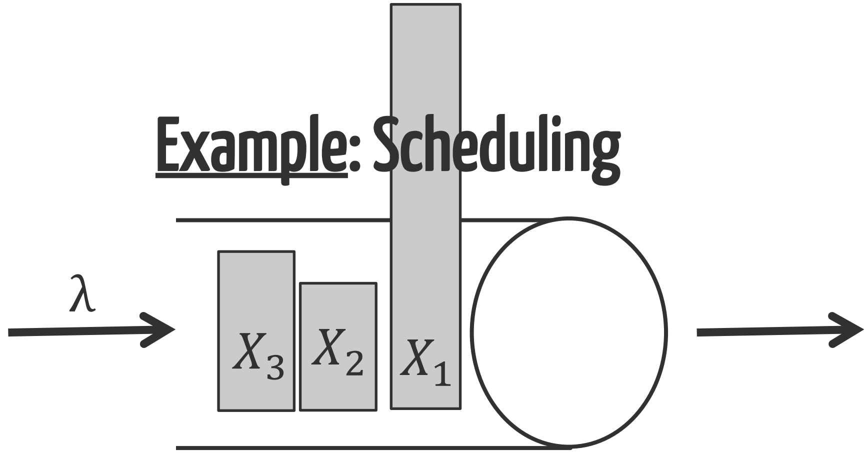
(We won't do this one today)

4. Algorithm Design

How does one “design” algorithms in the face of heavy tails?

Key: Minimize the impact of catastrophes

Example: Scheduling



Load: $\rho = \lambda E[X]$

What order should jobs be served in to minimize $\Pr(Delay > t)$ for large t?

Light-tailed

Heavy-tailed

FIFO

Large jobs can delay lots of smaller jobs (by a lot)

SRPT

The impact of large jobs on small jobs is minimal

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability:

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability:

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Why not just use the sample mean, $\bar{\mu}_n = \sum X_i / n$?

Light-tailed → **Optimal**

$$\varepsilon(n, \delta) = \sigma \sqrt{\frac{\log 1/\delta}{n}}$$

Heavy-tailed → **Bad**

$$\varepsilon(n, \delta) = \sigma \sqrt{\frac{1}{n\delta}}$$

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

[Tukey & McLaughlin 1963], [Bickel 1965], [Stigler 1973]

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \dots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

1. Divide data into two equal parts.
2. Use first part to determine truncation points $\beta = Y^*_{(1-\varepsilon)n}$ and $\alpha = Y^*_{\varepsilon n}$
3. Trim outliers using truncation points $Y_i = [X_i]_\alpha^\beta$
4. Estimate using sample mean of Y_i

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

$$\varepsilon(n, \delta) = 9\sigma\sqrt{\frac{\log 8/\delta}{n}}$$

+ it's robust to corruption!
[Lugosi and Mendelson, 2019]

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

Idea 2: Median of means

[Nemirovsky and Yudin 1983], [Jerrum,Valiant, and Vazirani 1986], [Alon, Matias, and Szegedy 2002]

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \dots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

Idea 2: Median of means

1. Divide data into k equal groups.
2. Compute the sample average of each group.
3. Compute the median of the sample averages.

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

Idea 2: Median of means

$$\varepsilon(n, \delta) = 9\sigma\sqrt{\frac{\log 8/\delta}{n}}$$

+ it works even when the variance is infinite!

[Bubeck, Cesa-Bianchi, and Lugosi, 2013]

Example: Estimating the mean

Goal: Given i.i.d. samples $X_1, \ldots, X_n$ with mean μ, develop estimates of the mean that are good with high probability,

$$\Pr(|\hat{\mu}_n - \mu| > \varepsilon(n, \delta)) \leq \delta$$

Key: Minimize the impact of catastrophes

Idea 1: Trim the outliers

Idea 2: Median of means

Note: Both methods depend on knowing δ (when setting truncation/group sizes). This is unavoidable.

How does one "design" algorithms in the face of heavy tails?

Key: Minimize the impact of catastrophes

Prioritization

Clipping/Truncation

Median of means

These ideas have been applied to SGD, RL, and bandits in recent years, but there are still many open problems in these and other areas!

Heavy-tailed phenomena are typically treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

(see the book!)

Heavy-tailed phenomena are typically treated as something

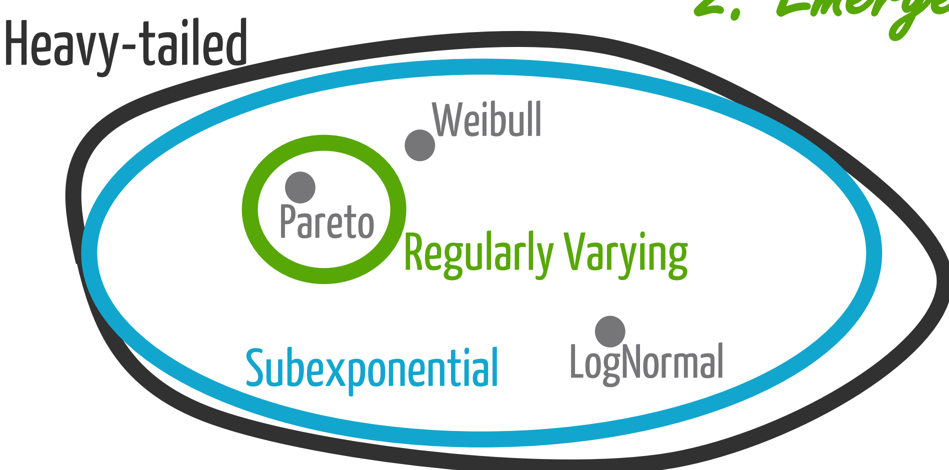
~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

(see the book!)



Heavy-tailed distributions have many beautiful & strange properties

1) Scale Invariance → Regularly varying distributions

2) The "catastrophe principle" → Subexponential distributions

Heavy-tailed phenomena are typically treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification
(see the book!)

We've all been taught that <u>the Normal is "normal"</u>
because of the Central Limit Theorem, **BUT**
Heavy-tails are more "normal" than the Normal!

Heavy-tailed phenomena are typically treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. *Properties*

2. *Emergence*

3. *Identification*
(see the book!)

4. *Algorithm Design*

Tail events can't be avoided, so algorithms must
Minimize the impact of catastrophes!

An Introduction to Heavy Tails for ML Researchers

Conspiracies, Catastrophes, and the Principle of a Single Big Jump

For details, references, etc., see: