Exponential tail bounds and Large Deviation Principle for Heavy-Tailed U-Statistics

Milad Bakhshizadeh

Stanford University

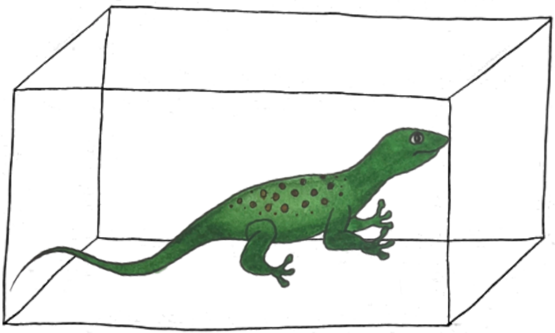
May 24, 2023

Uncertainty

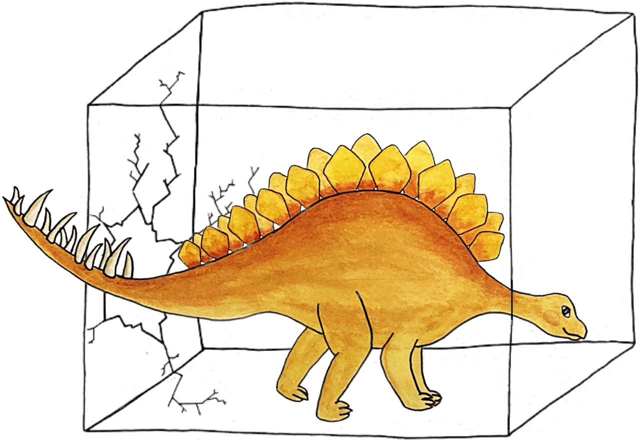
Uncertainty

Uncertainty

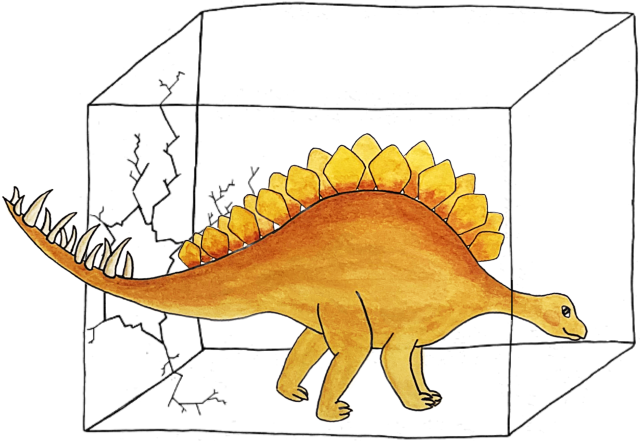
Uncertainty



Uncertainty



Uncertainty : How to make the Dinosaur a lizard?!

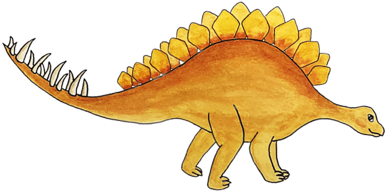


What is this talk about?

1. Heavy-tail makes it challenging to bound uncertainty
2. We can control heavy-tail by truncation
3. Truncation gives optimal bound

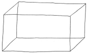
Part 1: problem setup

- Heavy-tail distributions: the dinosaur

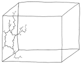


Heavy-tails don't have finite MGF

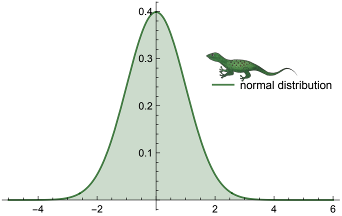
- SubGaussian distributions
  - $\mathbb{P}\left(|X| > t\right) \leq \exp\left(-ct^2\right)$
  - $\|X\|_p = O(\sqrt{p})$
  - $\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c\lambda^2\right)$



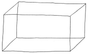
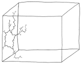
- Heavy-tailed distributions

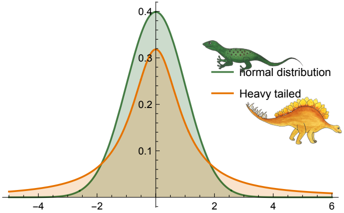


  - $\mathbb{E}\left[\exp\left(\lambda X\right)\right] = \infty, \ \forall \lambda > 0$
  - e.g. Weibull with $k < 1$, $\mathcal{N}(0,1)^\alpha$ $\alpha > 2$, Log-Normal, ...

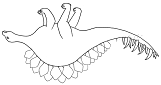


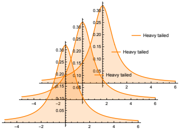
Heavy-tails don't have finite MGF

- SubGaussian distributions
  - $\mathbb{P}\left(|X| > t\right) \leq \exp\left(-ct^2\right)$
  - $\|X\|_p = O(\sqrt{p})$
  - $\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(c\lambda^2\right)$ 
- Heavy-tailed distributions
  - $\mathbb{E}\left[\exp\left(\lambda X\right)\right] = \infty, \ \forall \lambda > 0$ 
  - e.g. Weibull with $k < 1$, $\mathcal{N}(0,1)^{\alpha}$ $\alpha > 2$, Log-Normal, ...



Rise of the heavy-tails: A new era dawns





- Data is heavy-tailed
- Multiplication makes tail heavier
  - XY, X^n
  - $\mathcal{N}(0,1)$, $\mathcal{N}(0,1)^2$, $\mathcal{N}(0,1)^3$
- Real applications
  - Neural nets
  - Phase retrieval

Rise of the heavy-tails: A new era dawns

- Data is heavy-tailed
- Multiplication makes tail heavier
  - XY, X^n
  - $\mathcal{N}(0,1)$, $\mathcal{N}(0,1)^2$, $\mathcal{N}(0,1)^3$
- Real applications
  - Neural nets
  - Phase retrieval

Preprints

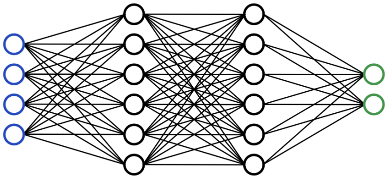
LAST UPDATE	AUTHORS	TITLE
Apr. 2023	M. Bakhshizadeh	Algebra of Sub-Weibull Random Variables (download)

Rise of the heavy-tails: A new era dawns

- Data is heavy-tailed
- Multiplication makes tail heavier
  - XY, X^n
  - $\mathcal{N}(0,1)$, $\mathcal{N}(0,1)^2$, $\mathcal{N}(0,1)^3$
- Real applications
  - Neural nets
  - Phase retrieval

Preprints

LAST UPDATE	AUTHORS	TITLE
Apr. 2023	**M. Bakhshizadeh**	Algebra of Sub-Weibull Random Variables (download)

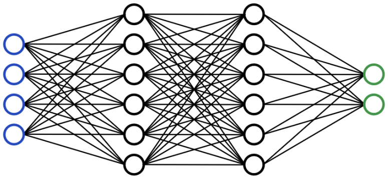


Rise of the heavy-tails: A new era dawns

- Data is heavy-tailed
- Multiplication makes tail heavier
  - XY, X^n
  - $\mathcal{N}(0,1)$, $\mathcal{N}(0,1)^2$, $\mathcal{N}(0,1)^3$
- Real applications
  - Neural nets
  - Phase retrieval

Preprints

LAST UPDATE	AUTHORS	TITLE
Apr. 2023	**M. Bakhshizadeh**	Algebra of Sub-Weibull Random Variables (download)



U-statistics, a low risk unbiased estimator

- $U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}), \qquad h(\cdot)$ symmetric
  - X_i iid
  - $U_n \to \mathbb{E}[h]$
  - $h = X_1 \qquad \to \hat{\mu} = \bar{X}_n$
  - $h = \frac{(X_1 - X_2)^2}{2} \to \hat{\sigma}^2$
- How fast $\mathbb{P}\left(\left|U_n - \mathbb{E}[h]\right| > \epsilon\right) \to 0$?
  - Sample size n
  - High-dimensional statistics

U-statistics, a low risk unbiased estimator

- $U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}), \qquad h(\cdot)$ symmetric
  - X_i iid
  - $U_n \to \mathbb{E}[h]$
  - $h = X_1 \qquad \to \hat{\mu} = \bar{X}_n$
  - $h = \frac{(X_1 - X_2)^2}{2} \to \hat{\sigma}^2$
- How fast $\mathbb{P}\left(\left|U_n - \mathbb{E}[h]\right| > \epsilon\right) \to 0$?
  - Sample size n
  - High-dimensional statistics

U-statistics, a low risk unbiased estimator

- $U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}), \qquad h(\cdot)$ symmetric
  - X_i iid
  - $U_n \to \mathbb{E}[h]$
  - $h = X_1 \qquad \to \hat{\mu} = \bar{X}_n$
  - $h = \frac{(X_1 - X_2)^2}{2} \to \hat{\sigma}^2$
- How fast $\mathbb{P}\left(\left|U_n - \mathbb{E}[h]\right| > \epsilon\right) \to 0$?
  - Sample size n
  - High-dimensional statistics

Concentration inequality, a tool for uncertainty control

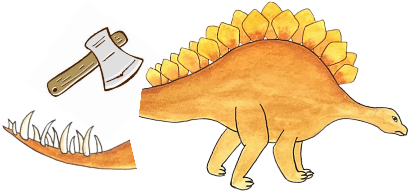
- $\mathbb{P}\left(|U_n| > \epsilon\right) \leq \exp\left(-L(n,\epsilon)\right)$
  - $\mathbb{E}\left[h\right] = 0$ (no generality loss)
  - Simple
  - Tight (asymptotically)
- Recall
  - $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i_1}, ..., X_{i_m})$
  - $\mathbb{E}\left[\exp\left(\lambda h(X_1, ..., X_m)\right)\right] = \infty, \quad \forall \lambda > 0$

Concentration inequality, a tool for uncertainty control

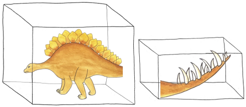
- $\mathbb{P}\left(|U_n| > \epsilon\right) \leq \exp\left(-L(n, \epsilon)\right)$
  - $\mathbb{E}\left[h\right] = 0 \quad$ (no generality loss)
  - Simple
  - Tight (asymptotically)
- Recall
  - $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i_1}, ..., X_{i_m})$
  - $\mathbb{E}\left[\exp\left(\lambda h(X_1, ..., X_m)\right)\right] = \infty, \quad \forall \lambda > 0$

Part 2

- The Solution: Tail Truncation



Bound tail and body, separately



- Define:
  - $\boldsymbol{k} = \lfloor \frac{n}{m} \rfloor$
  - $\boldsymbol{h_L} = h\, \mathbb{1}(h \leq L)$
  - $\mathbb{P}\left(h > t\right) \simeq \exp\left(-\boldsymbol{I(t)}\right), \;\; I(t) \ll t$
- $\mathbb{P}\left(U_n > t\right) \leq \mathbb{P}\left(U_n(h_L) > t\right) + \mathbb{P}\left(\exists i_j \; |h(X_{i_1}, \ldots, X_{i_m}) > L\right)$

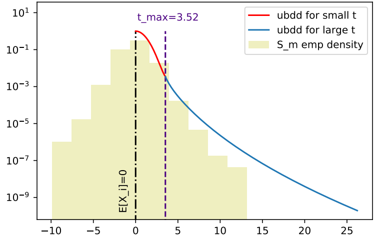
Theorem (1)

$$\mathbb{P}\left(U_n > t\right) \lesssim \exp\left(-\frac{kt^2}{2Var(h)}\right) + \left(1 + \binom{n}{m}\right) \exp\left(-I(kt)\right)$$

There are two different regions of deviation

$$\mathbb{P}\left(U_n > t\right) \lesssim \exp\left(-\frac{kt^2}{2\mathrm{Var}(h)}\right) + \left(1 + \binom{n}{m}\right) \exp\left(-I(kt)\right)$$

- Regions:
  - Small t, Gaussian decay
  - Large t, like $\exp\left(-I(kt)\right)$
- Change point: $kt^2 \simeq I(kt)$



Derivation

$$\mathbb{P}\left(U_n > t\right) \le \mathbb{P}\left(U_n(h_L) > t\right) + \mathbb{P}\left(\exists i_j \; |h(X_{i_1}, \dots, X_{i_m}) > L\right)$$
$$\le \mathrm{e}^{-\lambda t}\mathbb{E}\left[\mathrm{e}^{\lambda U_n(h_L)}\right] + \binom{n}{m}\mathrm{e}^{-I(L)}$$

- issues
  - dependent terms in $U_n(h_L)$
  - $\mathbb{E}\left[\mathrm{e}^{\lambda U_n(h_L)}\right] \xrightarrow{L\to\infty} \infty, \quad$ fixed λ
  - Optimize λ, L together
  - Sharpness

First term $\le \exp\left(-\frac{kt^2}{2v(kt, \beta\frac{I(kt)}{kt})}\right) + \exp\left(-\beta I(kt)\max(\frac{1}{2}, c(t, \beta, k))\right)$

- $v(L, \eta) \triangleq \mathbb{E}\left[h_L^2 \mathbf{1}(h \le 0) + h_L^2 \exp(\eta h_L)\mathbf{1}(h > 0)\right] \to Var(h)$
- $c(t, \beta, k) \triangleq 1 - \frac{\beta}{2t}\frac{I(kt)}{kt}v(kt, \beta\frac{I(kt)}{kt}) \to 1$

Derivation

$$\mathbb{P}\left(U_n > t\right) \leq \mathbb{P}\left(U_n(h_L) > t\right) + \mathbb{P}\left(\exists i_j \;|h(X_{i_1}, \ldots, X_{i_m}) > L\right)$$

$$\leq \mathrm{e}^{-\lambda t}\mathbb{E}\left[\mathrm{e}^{\lambda U_n(h_L)}\right] + \binom{n}{m}\mathrm{e}^{-I(L)}$$

- issues
  - dependent terms in $U_n(h_L)$
  - $\mathbb{E}\left[\mathrm{e}^{\lambda U_n(h_L)}\right] \xrightarrow{L\to\infty} \infty, \quad$ fixed λ
  - Optimize λ, L together
  - Sharpness

First term $\leq \exp\left(-\frac{kt^2}{2v(kt,\beta\frac{I(kt)}{kt})}\right) + \exp\left(-\beta I(kt)\max(\frac{1}{2}, c(t,\beta,k))\right)$

- $v(L,\eta) \triangleq \mathbb{E}\left[h_L^2\mathbf{1}(h \leq 0) + h_L^2\exp(\eta h_L)\mathbf{1}(h > 0)\right] \to Var(h)$
- $c(t,\beta,k) \triangleq 1 - \frac{\beta}{2t}\frac{I(kt)}{kt}v(kt, \beta\frac{I(kt)}{kt}) \to 1$

Derivation

$$\mathbb{P}(U_n > t) \leq \mathbb{P}\left(U_n(h_L) > t\right) + \mathbb{P}\left(\exists i_j \; |h(X_{i_1}, \ldots, X_{i_m}) > L\right)$$

$$\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda U_n(h_L)}\right] + \binom{n}{m} e^{-I(L)}$$

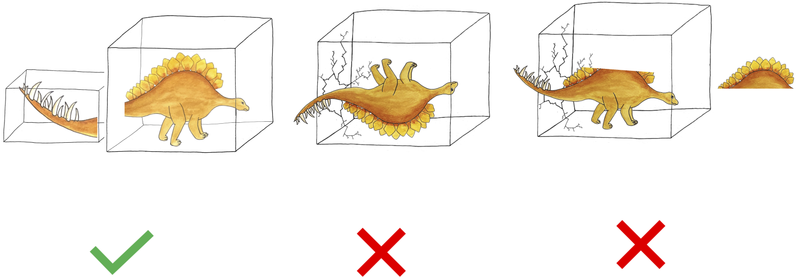
- issues
  - dependent terms in $U_n(h_L)$
  - $\mathbb{E}\left[e^{\lambda U_n(h_L)}\right] \xrightarrow{L\to\infty} \infty, \quad$ fixed λ
  - Optimize λ, L together
  - Sharpness

First term $\leq \exp\left(-\frac{kt^2}{2v(kt, \beta\frac{I(kt)}{kt})}\right) + \exp\left(-\beta I(kt)\max(\frac{1}{2}, c(t, \beta, k))\right)$

- $v(L, \eta) \triangleq \mathbb{E}\left[h_L^2 \mathbf{1}(h \leq 0) + h_L^2 \exp(\eta h_L)\mathbf{1}(h > 0)\right] \to Var(h)$
- $c(t, \beta, k) \triangleq 1 - \frac{\beta}{2t}\frac{I(kt)}{kt} v(kt, \beta\frac{I(kt)}{kt}) \to 1$

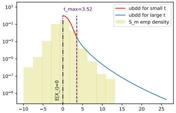
Part 3

- Tail truncation is optimal (in several cases)



For large same size, the bound is tight

$$\mathbb{P}\left(U_n > t\right) \lesssim \exp\left(-\frac{kt^2}{2\mathrm{Var}(h)}\right) + \left(1 + \binom{n}{m}\right)\exp\left(-I(kt)\right)$$



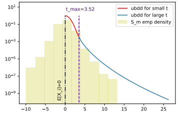
Theorem (2)

$$\lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{I(kt)} = 1$$

- Assumptions
  - $kt^2 \gg I(kt)$
  - $I(t) \geq c\sqrt[\alpha]{t} \quad \rightarrow \quad$ sub-Weibull
  - $(h > t) \simeq (\exists i \; |X_i| > f(t))$
  - $h = |X_1 - X_2|,\ (X_1 - X_2)^2,\ \max\left(|X_1|, |X_2|, \ldots, |X_m|\right)$

For large same size, the bound is tight

$$\mathbb{P}\left(U_n > t\right) \lesssim \exp\left(-\frac{kt^2}{2\mathrm{Var}(h)}\right) + \left(1 + \binom{n}{m}\right) \exp\left(-I(kt)\right)$$



Theorem (2)

$$\lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{I(kt)} = 1$$

- Assumptions
  - $kt^2 \gg I(kt)$
  - $I(t) \geq c\sqrt[\alpha]{t} \quad \rightarrow \quad$ sub-Weibull
  - $(h > t) \simeq (\exists i\ |X_i| > f(t))$
    - $h = |X_1 - X_2|,\ (X_1 - X_2)^2,\ \max\left(|X_1|, |X_2|, \ldots, |X_m|\right)$

It yields Large Deviation Principle (LDP)

- Recall
  - $\lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{I(kt)} = 1$
  - $I(t) \ll t$
- LDP
  - $I(kt) = c\sqrt{kt} \qquad \Longrightarrow \qquad \lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{\sqrt{n}} = c\sqrt{\frac{t}{m}}$
  - U_n satisfies LDP, with speed $\sqrt{n}$, and rate function $c\sqrt{\frac{t}{m}}$

It yields Large Deviation Principle (LDP)

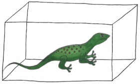
- Recall
  - $\lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{I(kt)} = 1$
  - $I(t) \ll t$
- LDP
  - $I(kt) = c\sqrt[\alpha]{kt} \quad \Longrightarrow \quad \lim_{n\to\infty} \frac{-\log \mathbb{P}(U_n > t)}{\sqrt[\alpha]{n}} = c\sqrt[\alpha]{\frac{t}{m}}$
  - U_n satisfies LDP, with speed $\sqrt[\alpha]{n}$, and rate function $c\sqrt[\alpha]{\frac{t}{m}}$

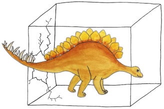
Future Work: This was a piece of the puzzle

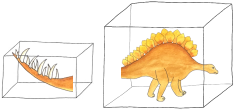
- Extend Heavy-tailed Analysis toolbox
  - Why only U-statistics?
  - $F_n = F(X_1, ..., X_n) \xrightarrow{n \to \infty} \mathbb{E}[F]$
- Applications
  - Finance
  - Differential Privacy
  - Asymptotic Hypothesis Testing
    - Bahadur Efficiency
  - ...

Thanks

Truncation can turn dinos to lizards, so heavy-tails better beware - they're next in line for a makeover!







$$\mathbb{P}\left(U_n > t\right) \lesssim \exp\left(-\frac{kt^2}{2\mathrm{Var}(h)}\right) + \left(1 + \binom{n}{m}\right) \exp\left(-I(kt)\right)$$

$$\mathbb{E}\left[\exp\left(\lambda h(X_1, \ldots, X_m)\right)\right] = \infty,$$

$$\lim_{n\to\infty} \frac{-\log \mathbb{P}\left(U_n > t\right)}{I(kt)} = 1$$

Cartoon characters credit: Mina Latifi

Another application

- Phase retrieval
  - $\boldsymbol{y} = |X\beta| + \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$
  - $\hat{\beta} = \arg\min_{\boldsymbol{b}} \left\| \boldsymbol{y}^2 - (X\boldsymbol{b})^2 \right\|^2$

The extreme event:

One X_i large $\implies \binom{n-1}{m-1}$ kernel terms large

- Recall:
  - $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i_1}, \dots, X_{i_m})$
  - $(\exists i \; |X_i| > f(t)) \simeq (h > t)$
  - $\frac{-\log \mathbb{P}(U_n > t)}{I(kt)} \to 1$
- $\mathcal{E} = \left(\exists i \text{ s.t. } |X_i| > f(kt)\right)$ (the event)
  - $\mathbb{P}(\mathcal{E}) \simeq \exp\left(-I(kt)\right)$
  - $\mathcal{E} \implies (U_n > t)$, $\qquad U_n > \frac{1}{\binom{n}{m}} \binom{n-1}{m-1} kt = \frac{m}{n} kt \simeq t$