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

Milad Bakhshizadeh

Columbia University, Department of Statistics

May 3, 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
  - $\mathbb{P}(|X| > t) \leq \exp(-ct^2)$
  - $\|X\|_p = O(\sqrt{p})$
  - $\mathbb{E}\left[ \exp(\lambda X) \right] \leq \exp(c\lambda^2)$
- Heavy-tailed distributions
  - $\mathbb{E}\left[ \exp(\lambda X) \right] = \infty$, $\forall \lambda > 0$
  - 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
    - \( y = |X\beta| + \mathcal{N}(0, \sigma^2 I_n) \)
    - \( \hat{\beta} = \arg \min_b \| y^2 - (Xb)^2 \|^2 \)
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
    - $y = |X\beta| + \mathcal{N}(0, \sigma^2 I_n)$
    - $\hat{\beta} = \arg\min_b \| y - (Xb)^2 \|^2$
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

$y = |X\beta| + \mathcal{N}(0, \sigma^2 I_n)$
$
\hat{\beta} = \arg\min \| y^2 - (X\beta)^2 \|^2
$
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
    - $y = |X\beta| + \mathcal{N}(0, \sigma^2 I_n)$
    - $\hat{\beta} = \arg \min_b \|y^2 - (Xb)^2\|^2$
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
    - $y = |X\beta| + \mathcal{N}(0, \sigma^2 I_n)$
    - $\hat{\beta} = \arg\min_b \|y^2 - (Xb)^2\|_2^2$
Rise of the heavy-tails: A new era dawns

- Data is heavy-tailed
- Multiplication makes tail heavier
  - $XY, X^n$
  - $N(0,1), N(0,1)^2, N(0,1)^3$
- Real applications
  - Neural nets
  - Phase retrieval
    - $y = |X\beta| + N(0, \sigma^2 I_n)$
    - $\hat{\beta} = \arg \min_b \|y^2 - (Xb)^2\|^2$
U-statistics, a low risk unbiased estimator

\[ U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}), \quad h(\cdot) \text{ symmetric} \]

- \( X_i \) iid
- \( U_n \to \mathbb{E}[h] \)
  - \( h = X_1 \quad \to \hat{\mu} = \bar{X}_n \)
  - \( h = \frac{(X_1 - X_2)^2}{2} \quad \to \hat{\sigma}^2 \)

- How fast \( \mathbb{P}\left( |U_n - \mathbb{E}[h]| > \epsilon \right) \to 0? \)
  - Sample size \( n \)
  - High-dimensional statistics

Milad Bakhshizadeh
Columbia University, Department of Statistics
Exponential tail bounds and Large Deviation Principle for Heavy-Tailed U-Statistics
U-statistics, a low risk unbiased estimator

\[ U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}), \quad h(\cdot) \text{ symmetric} \]

- \[ X_i \text{ iid} \]
- \[ U_n \to \mathbb{E} [h] \]
- \[ h = X_1 \quad \Rightarrow \hat{\mu} = \bar{X}_n \]
- \[ h = \frac{(X_1 - X_2)^2}{2} \quad \Rightarrow \hat{\sigma}^2 \]

- How fast \( \mathbb{P} \left( |U_n - \mathbb{E}[h]| > \epsilon \right) \to 0? \)
- Sample size \( n \)
- High-dimensional statistics
U-statistics, a low risk unbiased estimator

\[
U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} h(X_{i_1}, \ldots, X_{i_m}), \quad h(\cdot) \text{ symmetric}
\]

- \( X_i \) iid
- \( U_n \to \mathbb{E}[h] \)
- \( h = X_1 \to \hat{\mu} = \bar{X}_n \)
- \( h = \frac{(X_1 - X_2)^2}{2} \to \sigma^2 \)

- How fast \( \mathbb{P}\left( |U_n - \mathbb{E}[h]| > \epsilon \right) \to 0? \)
- Sample size \( n \)
- High-dimensional statistics
Concentration inequality, a tool for uncertainty control

- $\mathbb{P}(|U_n| > \epsilon) \leq \exp(-L(n, \epsilon))$
  - $\mathbb{E}[h] = 0$ (no generality loss)
  - Simple
  - Tight (asymptotically)

Recall

- $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i1}, ..., X_{im})$
  - $\mathbb{E}\left[\exp(\lambda h(X_1, ..., X_m))\right] = \infty, \quad \forall \lambda > 0$
Concentration inequality, a tool for uncertainty control

- $\mathbb{P}(\lvert U_n \rvert > \epsilon) \leq \exp(-L(n, \epsilon))$
- $\mathbb{E}[h] = 0$ (no generality loss)
- Simple
- Tight (asymptotically)

Recall

- $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i_1}, \ldots, X_{i_m})$
- $\mathbb{E}\left[\exp(\lambda h(X_1, \ldots, X_m))\right] = \infty, \quad \forall \lambda > 0$
Part 2

The Solution: Truncation
Bound tail and body, separately

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

**Theorem (1)**

\[
\mathbb{P}(U_n > t) \lesssim \exp\left(-\frac{kt^2}{2\text{Var}(h)}\right) + \left(1 + \binom{n}{m}\right)\exp(-l(kt))
\]
There are two different regions of deviation

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

- **Regions:**
  - Small $t$, Gaussian decay
  - Large $t$, like $\exp(-I(kt))$

- **Change point:** $kt^2 \simeq I(kt)$
Part 3

- Tail truncation is optimal (in several cases)
For large same size, the bound is tight

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

**Theorem (2)**

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

- **Assumptions**
  - \( kt^2 \gg l(kt) \)
  - \( l(t) \geq c \alpha \sqrt{t} \) → sub-Weibull
  - \( (h > t) \simeq (\exists i \mid X_i \mid > f(t)) \)
    - \( h = |X_1 - X_2|, (X_1 - X_2)^2, \max \{|X_1|, |X_2|, \ldots, |X_m|\} \)
For large same size, the bound is tight

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

**Theorem (2)**

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

- **Assumptions**
  - \( kt^2 \gg l(kt) \)
  - \( l(t) \geq c \sqrt{t} \) \( \rightarrow \) sub-Weibull
  - \( (h > t) \simeq (\exists i \ |X_i| > f(t)) \)
    - \( h = |X_1 - X_2|, \ (X_1 - X_2)^2, \ \max(|X_1|, |X_2|, \ldots, |X_m|) \)
It yields Large Deviation Principle (LDP)

- Recall
  - \( \lim_{n \to \infty} \frac{-\log P(U_n > t)}{l(kt)} = 1 \)
  - \( l(t) \ll t \)

- LDP
  - \( l(kt) = c \sqrt{kt} \quad \implies \quad \lim_{n \to \infty} \frac{-\log P(U_n > t)}{\sqrt{n}} = c \sqrt{\frac{t}{m}} \)
  - \( U_n \) satisfies LDP, with speed \( \sqrt{n} \), and rate function \( c \sqrt{\frac{t}{m}} \)
It yields Large Deviation Principle (LDP)

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

- LDP
  - \( I(kt) = c \sqrt[\alpha]{kt} \implies \lim_{n \to \infty} \frac{-\log P(U_n > t)}{\sqrt[\alpha]{n} \sqrt{m}} = c \sqrt[\alpha]{\frac{t}{m}} \)
  - \( U_n \) satisfies LDP, with speed \( \sqrt[\alpha]{n} \), and rate function \( c \sqrt[\alpha]{\frac{t}{m}} \)
The extreme event:

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

- Recall:
  - $U_n = \frac{1}{\binom{n}{m}} \sum h(X_{i1}, \ldots, X_{im})$
  - $(\exists i \mid X_i \mid > f(t)) \simeq (h > t)$
  - $-\log \mathbb{P}(U_n > t) \frac{l(kt)}{l(kt)} \rightarrow 1$

- $\mathcal{E} = (\exists i \text{ s.t. } |X_i| > f(kt))$ (the event)
  - $\mathbb{P}(\mathcal{E}) \simeq \exp\left(-l(kt)\right)$
  - $\mathcal{E} \implies (U_n > t), \quad U_n > \frac{1}{\binom{n}{m}}(n-1) kt = \frac{m}{n} kt \simeq t$
Future Work: This was a piece of the puzzle

- Extend Heavy-tailed Analysis toolbox
  - $F_n = F(X_1, \ldots, X_n) \xrightarrow{n \to \infty} \mathbb{E} [F]$

- 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!

\[ \Pr(U_n > t) \leq \exp \left( -\frac{kt^2}{2\text{Var}(h)} \right) + \left( 1 + \binom{n}{m} \right) \exp \left( -I(kt) \right) \]

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

\[ \lim_{n \to \infty} \frac{- \log \Pr(U_n > t)}{I(kt)} = 1 \]

Cartoon characters credit: Mina Latifi