Homework II
Theoretical Statistics/Machine Learning

Sagar Ghosh
Department of Statistics and Data Sciences
The University of Texas at Austin
Email: sg63684@my.utexas.edu

February 6, 2026

1 Jackknife: Variance and Bias.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued random variables. (Assume through- out that all random variables below have finite second moments whenever a variance is invoked.) Let θ be a parameter of the distribution of the X_i s that we wish to estimate. For each $k \in \mathbb{N}$, let $f_k : \mathbb{R}^k \to \mathbb{R}$ be a (measurable) statistic that is symmetric (invariant under permutations of its arguments). Define the full-sample estimator

$$Z_n = f_n(X_1, X_2, \ldots, X_n)$$

The quality of the estimator Z_n is measured by by its bias and its variance (to be defined below shortly). Since we do not know the distribution of the X_is, we need to estimate the bias and variance. The key difficulty is that we have only one dataset and so the original estimator Z_n and the bias and variance estimators are to be constructed from the same data. The jackknife is a technique to (partially) solve this difficulty.

Define the leave-one-out jackknife estimates

$$Z^{(i)} := f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \quad i = 1, 2, , \ldots, n.$$

and define their average

$$\bar{Z} := \frac{1}{n} \sum_{i=1}^{n} Z^{(i)}.$$

Below we will define jackknife estimates of the variance and bias and show some of their proper- ties.

(a) **Variance Bound.** We define the (simplified) jackknife variance functional

$$\hat{V} := \sum_{i=1}^{n} (Z^{(i)} - \bar{Z})^2.$$

(a_1) On one hand, we have

$$\hat{V} = \sum_{i=1}^{n} (Z^{(i)} - \bar{Z})^2 = \sum_{i=1}^{n} (Z^{(i)})^2 - 2\bar{Z} \sum_{i=i}^{n} Z^{(i)} + n\bar{Z}^2 = \sum_{i=1}^{n} (Z^{(i)})^2 - n\bar{Z}^2.$$

On the other hand, we expand the other expression:

$$\begin{aligned}
\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} (Z^{(i)} - Z^{(j)})^2 &= \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{n} (Z^{(i)})^2 + \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{n} (Z^{(i)})^2 - 2\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} Z^{(i)} Z^{(j)} \\
&= \frac{1}{2} \sum_{i=1}^{n} (Z^{(i)})^2 + \frac{1}{2} \sum_{j=1}^{n} (Z^{(j)})^2 - \frac{1}{n} \left(\sum_{i=1}^{n} Z^{(i)} \right) \left(\sum_{j=1}^{n} Z^{(j)} \right) \\
&= \sum_{i=1}^{n} (Z^{(i)})^2 - n\bar{Z}^2.
\end{aligned}$$

Hence these two expressions are the same, so the identity holds: $\hat{V} = \sum_{i=1}^{n}(Z^{(i)} - \bar{Z})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(Z^{(i)} - Z^{(j)})^2$.

(a_2) Let $W = f_{n-1}(X_1, X_2, \ldots, X_{n-1})$ be the same statistic applied to an i.i.d. sample size of $n-1$. Then $W = Z^{(n)}$. So, $\mathrm{Var}(W) = \mathrm{Var}(Z^{(n)})$. Applying tensorization of variance to $Z^{(n)} = f_{n-1}(X_1, X_2, \ldots, X_{n-1})$ we get

$$\mathrm{Var}(Z^{(n)}) \leq \sum_{k=1}^{n-1} \mathbb{E}\left[\mathrm{Var}(Z^{(n)}|X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})\right]$$

Now fix $k \in \{1, 2, \ldots, n-1\}$ and condition on $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})$. Then we can write $Z^{(n)} = g(X_k)$ and $Z^{(k)} = g(X_n)$ for the same random function g. Because once we fix $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})$, the only variable in $Z^{(n)}$ is X_k and the one in $Z^{(k)}$ is X_n. Now X_n and X_k are i.i.d. and is independent of the conditioned variables $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})$. Therefore,

$$\begin{aligned}
\mathrm{Var}(Z^{(n)}|(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})) &= \mathrm{Var}(g(X_k)|(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})) \\
&= \frac{1}{2}\mathbb{E}\left[(g(X_n) - g(x_k))^2|(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})\right] \\
&= \frac{1}{2}\mathbb{E}\left[(Z^{(n)} - Z^{(k)})^2|(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{n-1})\right]
\end{aligned}$$

Therefore,

$$\mathrm{Var}(Z^{(n)}) \leq \frac{1}{2}\sum_{k=1}^{n-1} \mathbb{E}\left[(Z^{(n)} - Z^{(k)})^2\right]. \tag{1}$$

Using the exchangeability of $(Z^{(1)}, Z^{(2)}, \ldots, Z^{(n)})$, we can write

$$\mathbb{E}\left[\sum_{k=1,k\neq i}^{n}(Z^{(i)} - Z^{(k)})^2\right] = \mathbb{E}\left[\sum_{k=1,k\neq j}^{n}(Z^{(j)} - Z^{(k)})^2\right]$$

for each pair (i, j). Hence, the RHS of Equation 1 can be written as

$$\frac{1}{2}\sum_{k=1}^{n-1} \mathbb{E}\left[(Z^{(n)} - Z^{(k)})^2\right] \leq \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[(Z^{(i)} - Z^{(j)})^2\right] = \mathbb{E}[\mathrm{Var}(\hat{V})].$$

Finally using the RHS of (a_1), we get

$$\mathrm{Var}(W) \leq \mathbb{E}[\hat{V}].$$

(b) **Exact Bias for Quadratic Statistic.** let θ be a parameter of the distribution of X_1 and define the full sample estimator:

$$Z_n = f_n(X_1, X_2, \ldots, X_n).$$

We assume the bias admits a first order expansion:

$$\mathbb{E}Z_n - \theta = \frac{c}{n} + \mathcal{O}(n^{-2})] \quad \text{as} \quad n \to \infty.$$

We define the jackknife bias estimate $\hat{B}_n = (n-1)(\bar{Z} - Z_n)$ and bias corrected estimator $\tilde{Z}_n = Z_n - \hat{B}_n = Z_n - (n-1)(\bar{Z} - Z_n) = nZ_n - (n-1)\bar{Z}$. We note that $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, hence

$$\begin{aligned}
\mathbb{E}[\bar{Z}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f_{n-1}(X_1, X_2, \ldots, X_{n-1})\right] \quad \text{[Since } X_i\text{s are iid]} \\
&= \mathbb{E}\left[f_{n-1}(X_1, X_2, \ldots, X_{n-1})\right] \\
&= \mathbb{E}[Z_{n-1}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\tilde{Z}_n - \theta &= \mathbb{E}[nZ_n] - \mathbb{E}[(n-1)Z_{n-1}] \\
&= n\mathbb{E}Z_n - (n-1)\mathbb{E}Z_{n-1} \\
&= n\left[\frac{c}{n} + \mathcal{O}(n^{-2})\right] - (n-1)\left[\frac{c}{n-1} + \mathcal{O}((n-1)^{-2})\right] \quad \text{[For large } n\text{]} \\
&= \mathcal{O}(n^{-2}).
\end{aligned}$$

2 Order Statistics: Variance vs Spacings

Let $X_1, X_2, \ldots, X_n$ be independent random variables with

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

denote the order statistics. Define the spacing

$$\Delta_k = X_{(k+1)} - X_{(k)}, k = 1, 2, \ldots, n-1.$$

Assume $\mathbb{E}[X_{(k)}^2] < \infty$ for the relevant k so that the variances below are finite. For each $i \in \{1, 2, \ldots, n\}$, let X_i' be an independent copy of X_i (independent of everything), and let $\{X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(n)}\}$ be the resampled vector where

$$X_j^{(i)} = \begin{cases} X_j, \text{if } j \neq i \\ X_i', \text{if } j = i. \end{cases}$$

Let $X_k^{(i)}$ denote the $k-$ the order statistic for the resampled sample. We recall the one-sided ESS inequality:

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - Z^{(i)})_+^2], \quad (a)_+ := \max\{a, 0\},$$

as well as the analogous bound with $[Z^{(i)} - Z]_+^2$ (apply the inequality to $-Z$). Throughout we use $\mathbf{1}\{E\}$ for the indicator function.

(a) **The Maximum**

(a_1) Note that $X_{(n)} = \max\{X_1, \ldots, X_n\}$, and $X_{(n)}^{(i)} = \max\{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\}$. We also note that if $X_{(n)} = X_i$, then $X_{(n)}^{(i)} \geq X_{(n-1)}$ [since $X_i' \leq X_j \leq X_{(n-1)}$ for all $j \neq i$]. Therefore,

$$[X_{(n)} - X_{(n)}^{(i)}]_+ = \begin{cases} 0, & \text{if } \max\{X_1, \ldots, X_n\} = \max\{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\} = X_j, \text{ for } j \neq i \\ (X_i - X_{(n-1)}), & \text{if } X_i = \max\{X_1, ..., X_n\} \\ 0, & \text{if } X_{(n)} = X_i < X_{(n)}^{(i)} = \max\{X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\}. \end{cases}$$

Therefore, $(X_{(n)} - X_{(n)}^{(i)})_+ \leq (X_{(n)} - X_{(n-1)})\mathbf{1}\{X_i = X_{(n)}\}$. Using the tensorization of variance, we get

$$\begin{aligned} \mathrm{Var}(X_{(n)}) &\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(n)} - (X_{(n)}^{(i)}))_+^2] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(n)} - X_{(n-1)})^2 \mathbf{1}\{X_i = X_{(n)}\}] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(n)} - X_{(n-1)})^2] \mathbb{P}[X_i = X_{(n)}] \quad \text{[By Cauchy-Schwarz inequality]} \\ &= \mathbb{E}[\Delta_{n-1}^2] \sum_{i=1}^{n} \frac{1}{n} \\ &= \mathbb{E}[\Delta_{n-1}^2] \end{aligned}$$

(a_2) We compute the distributions of $X_{(n)}$ and $X_{(n)} - X_{(n-1)}$. For $t \in [0,1]$

$$\begin{aligned}
\mathbb{P}[X_{(n)} \leq t] &= \mathbb{P}[X_1 \leq t, X_2 \leq t, ..., X_n \leq t] \\
&= \prod_{i=1}^{n} \mathbb{P}[X_i \leq t] \quad \text{[because of independence]} \\
&= \prod_{i=1}^{n} t \\
&= t^n.
\end{aligned}$$

Hence, $\mathbb{P}[X_{(n)} = t] = nt^{n-1}$. So, $X_{(n)} \sim \text{Beta}(n,1)$. Note that, the event $\{\Delta_{n-1} = X_{(n)} - X_{(n-1)}\, t\}$ implies none of the samples X_is can fall in the range $[u,v]$ for some $u, v \in [0,1]$ with $v - u = t$. Therefore,

$$\begin{aligned}
\mathbb{P}[\Delta_{n-1} = X_{(n)} - X_{(n-1)} \leq t] &= 1 - \mathbb{P}[\Delta_n > t] \\
&= 1 - \mathbb{P}[\{X_i < u\} \cup \{X_i > v\}; \text{for all } i] \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}[\{X_i < u\} \cup \{X_i > v\}] \\
&= 1 - \prod_{i=1}^{n} [(u) + (1 - v)] \\
&= 1 - \prod_{i=1}^{n} [1 - t] \\
&= 1 - (1-t)^n.
\end{aligned}$$

Hence, $\mathbb{P}[\Delta_{n-1} = t] = n(1-t)^{n-1}$. So, $\Delta_n \sim \text{Beta}(1,n)$. Hence, $\text{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}$. And $\mathbb{E}[\Delta_{n-1}^2] = \frac{2}{(n+1)(n+2)}$. Since $\frac{n}{n+1} < 2$, $\text{Var}(X_{(n)}) < \mathbb{E}[\Delta_{n-1}^2]$.

(a_3) Now assume $X_i \overset{\text{i.i.d.}}{\sim} \exp(1)$ with pdf $e^{(}- x)\mathbf{1}\{x > 0\}$. The cdf of $X_{(1)}$ is

$$\begin{aligned}
\mathbb{P}[X_{(1)} \leq t] &= 1 - \mathbb{P}[X_{(1)} \geq t] \\
&= \prod_{i=1}^{n} \mathbb{P}[X_i \geq t] \\
&= e^{-nt}.
\end{aligned}$$

Hence $X_{(1)} \sim \exp(n)$. We will show by induction that $\Delta_k = X_{(k+1)} - X_{(k)} \sim \exp(n-k)$. We will use the memory-less property of the exponential distribution, i.e., $Y = X - t | X > t \sim \exp(1)$. Note that $\Delta_1 = X_{(1+1)} - X_{(1)} = \min\{X_i - X_{(1)}\}$, i.e., if we shift the distributions of $X_{(k)}$s for $k \geq 2$ by $X_{(1)}$, using the memory-less property we can conclude that $\Delta_1 \sim \exp(n-1)$ since $\{X_i - X_{(1)}\}_i$ has $n-1$ exponential random variables in its collection. By the induction hypothesis if we assume that $\Delta_k = X_{(k+1)} - X_{(k)} \sim \exp(n-k)$ for $n-2 \geq k \geq 2$, then using the same trick [shifting the rest $n-k$ variables by $X_{(k)}$], we can conclude that

$\Delta_{k+1} = X_{(k+2)} - X_{(k+1)} \sim \exp(n-(k+1))$. We also note that the Δ_ks are mutually independent, because once $X_{(k)}$ occurs, the entire process restarts at $X_{(k)}$ using the memory-less property. We now compute the $\text{Var}(X_{(n)})$. Note that $X_{(n)} = \sum_{k=1}^{n-1} \Delta_k + X_{(1)}$ and they all are mutually independent. Hence,

$$\begin{aligned}
\text{Var}(X_{(n)}) &= \sum_{k=1}^{n-1} \text{Var}(\Delta_k) + \text{Var}(X_{(1)}) \\
&= \sum_{k=1}^{n-1} \frac{1}{(n-k)^2} + \frac{1}{n^2} \\
&= \sum_{k=1}^{n} \frac{1}{k^2},
\end{aligned}$$

the second equality follows because if $Y \sim \exp(\lambda)$, $\text{Var}(Y) = \frac{1}{\lambda^2}$, and $\mathbb{E}[Y^2] = \frac{2}{\lambda^2}$. Hence,

$$\mathbb{E}[(X_{(n)} - X_{(n-1)})^2] = \mathbb{E}[\Delta_{n-1}^2] = 2,$$

since $\Delta_{n-1} \exp(n-(n-1)) = \exp(1)$. As $n \to \infty$, $\text{Var}(X_{(n)}) = \sum_{k=1}^{n} \frac{1}{k^2} \to \frac{\pi^2}{6}$. But the RHS is fixed at 2. So, the ratio of RHS to LHS is $\frac{12}{\pi^2} \approx 1.216$.

(b) **The general Order Statistics** we generalize the previous result for any order statistic. Fix $k \in \{2, ..., n-1\}$.

(b_1) Note that if $X_i \leq X_{(k)}$, then replacing X_i by X_i' can only increase the $k-$th minimum in $X^{(i)}$ to at most $X_{(k+1)}$, i.e., $(X_{(k)}^{(i)} - X_{(k)})_+ \leq (X_{(k+1)} - X_{(k)})\mathbf{1}\{X_i \leq X_{(k)}\}$. Now using the tensorization of variance,

$$\begin{aligned}
\text{Var}(X_{(k)}) &\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(k)} - X_{(k)}^{(i)})_+^2] \\
&\leq \sum_{k=1}^{n} \mathbb{E}[(X_{(k+1)} - X_{(k)})^2 \mathbf{1}\{X_i \leq X_{(k)}\}] \\
&= \mathbb{E}[(X_{(k+1)} - X_{(k)})^2] \sum_{i=1}^{n} \mathbf{1}\{X_i \leq X_{(k)}\} \\
&\leq k\mathbb{E}[(X_{(k+1)} - X_{(k)})^2] \\
&= k\mathbb{E}[\Delta_k^2]
\end{aligned}$$

Similarly, we note that if X_i is among the top $(n-k+1)$ maximum values of $\{X_1, ..., X_n\}$, then replacing X_i by X_i' can decrease $X_{(k)}^{(i)}$ to at most $X_{(k-1)}$, i.e., $X_{(k)}^{(i)} \geq X_{(k-1)}$. Hence, $(X_{(k)} - X_{(k)}^{(i)})_+ \leq (X_{(k)} -$

$X_{(k-1)})\mathbf{1}\{X_i \geq X_{(k)}\}$. Using the tensorization of variance we get

$$\begin{aligned}
\mathrm{Var}(X_{(k)}) &\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(k)} - X_{(k)}^{(i)})_+^2] \\
&\leq \sum_{i=1}^{n} \mathbb{E}[(X_{(k)} - X_{(k-1)})^2 \mathbf{1}\{X_i \geq X_{(k)}\}] \\
&= \mathbb{E}[(X_{(k)} - X_{(k-1)})^2] \sum_{i=1}^{n} \mathbf{1}\{X_i \geq X_{(k)}\} \\
&\leq (n-k+1)\mathbb{E}[(X_{(k)} - X_{(k-1)})^2] \\
&= (n-k+1)\mathbb{E}[\Delta_{k-1}^2].
\end{aligned}$$

(b_2) Since for $1 \leq k \leq [n/2]$,

$$\min\{k\mathbb{E}[\Delta_k^2], (n-k+1)\mathbb{E}[\Delta_{k-1}^2]\} = k\mathbb{E}[\Delta_k^2]$$

and for $[n/2] \leq k \leq n$,

$$\min\{k\mathbb{E}[\Delta_k^2], (n-k+1)\mathbb{E}[\Delta_{k-1}^2]\} = (n-k+1)\mathbb{E}[\Delta_{k-1}^2].$$

Combining these two we get

$$\mathrm{Var}(X_{(k)}) \leq \begin{cases} k\mathbb{E}[\Delta_k^2], & \text{for } 1 \leq k \leq [n/2] \\ (n-k+1)\mathbb{E}[\Delta_{k-1}^2]\}, & \text{for } [n/2] \leq k \leq n. \end{cases}$$

3 Rademacher Process: Bounding the Variance.

Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ be independent Rademacher variables with $\mathbb{P}[\epsilon_i = 1] = \mathbb{P}[\epsilon_i = -1] = \frac{1}{2}$. Let $T \subseteq \mathbb{R}^n$ be a finite set, we write $t = \{t_1, t_2, \ldots, t_n\}, t \in T$. Define $Z = \sup_{t\in T} \sum_{i=1}^n t_i\epsilon_i$. We also introduce the parameters

$$\sigma^2 = \sup_{t\in T} \sum_{i=1}^n t_i^2 \quad ; \quad \sigma_\infty^2 = \sum_{i=1}^n \sup_{t\in T} t_i^2.$$

(a) **A weak variance bound.** We start simple and show a coordinatewise variance bound for Z.

(a_1) For $t \in \mathbb{R}^n$,

$$\mathrm{Var}\left(\sum_{i=1}^n t_i\epsilon_i\right) = \sum_{i=1}^n t_i^2 \mathrm{Var}(\epsilon_i) = \sum_{i=1}^n t_i^2 \cdot 1 = \sum_{i=1}^n t_i^2.$$

Hence,

$$\sup_{t\in T} \mathrm{Var}\left(\sum_{i=1}^n t_i\epsilon_i\right) = \sup_{t\in T} \sum_{i=1}^n t_i^2 = \sigma^2.$$

(a_2) Since $\epsilon_\in\{-1, 1\}$, the range of Z pertaining to the $i-th$ component of the Rademacher variables is $2\sup_{t\in T} |t_i|$, since $\epsilon_i t_i \in [-t_i, t_i]$ and we are taking the sup over all $t \in T$. Hence by variance bound, we can write,

$$\begin{aligned}
\mathrm{Var}(Z) &\leq \frac{1}{4} \sum_{i=1}^n \mathrm{Range}(\epsilon_i t_i)^2 \\
&\leq \frac{1}{4} \sum_{i=1}^n \left(2 \sup_{t\in T} |t_i|\right)^2 \\
&\leq \frac{1}{4} \sum_{i=1}^n 4 \sup_{t\in T} t_i^2 \quad \text{[Pointwise supremum of a convex family is again convex]} \\
&= \sigma_\infty^2.
\end{aligned}$$

(b) **A sharper variance bound**

(b_1) Let $\epsilon_1', \epsilon_2', \ldots, \epsilon_n'$ be independent copies of $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$. We define

$$\epsilon^{(i)} = (\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_i', \epsilon_{i+1}, \ldots, \epsilon_n), \quad Z^{(i)} = \sup_{t\in T} \sum_{j=1}^n \epsilon_j^{(i)} t_j.$$

Assume the finite set T is given as $T = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$. Now, using a fixed ordering of T, we can pick up the smallest index $k = k(\epsilon) \in \{1, 2, ..., n\}$ such that for $\mathbf{t}^* = \mathbf{t}_k^*(\epsilon)$ the sup of Z is attained, i.e. $Z = \sum_{j=1}^n \epsilon_j t_j^*$,

where $\mathbf{t}^* = \{t_1^*, \ldots, t_n^*\}$. Hence,

$$\begin{aligned} Z^{(i)} &= \sup_{t\in T} \sum_{j=1}^{n} \epsilon_j^{(i)} t_j \\ &= \sup_{t\in T} \left[\sum_{j=1}^{n} \epsilon_j t_j + \epsilon_i' t_i - \epsilon_i t_i \right] \\ &\sup_{t\in T} \sum_{j=1}^{n} \epsilon_j t_j + \sup_{t\in T} (\epsilon_i' - \epsilon_i) t_i \\ &= Z + \sup_{t\in T} (\epsilon_i' - \epsilon_i) t_i. \end{aligned}$$

Therefore,

$$(Z - Z^{(i)})_+ = |\sup_{t\in T} (\epsilon_i - \epsilon_i') t_i| \leq |\epsilon_i - \epsilon_i'||t_i^*|.$$

Using the independence of ϵ_i and ϵ_i', we also compute $\mathbb{E}[(\epsilon_i - \epsilon_i')^2] = \mathbb{E}[\epsilon_i^2] + \mathbb{E}[(\epsilon_i')^2] - 2\mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_i'] = 1+1+0 = 2$.

Finally using the tensorization of variance we get

$$\begin{aligned} \mathrm{Var}(Z) &\leq \sum_{j=1}^{n} \mathbb{E}[(Z - Z^{(j)})_+^2] \\ &\leq \sum_{j=1}^{n} \mathbb{E}[(\epsilon_j - \epsilon_j')^2 |t_j^*|^2] \\ &= 2 \sum_{j=1}^{n} |t_j^*|^2 \\ &= 2 \sup_{t\in T} \sum_{j=1}^{n} t_j^2 \\ &= 2\sigma^2. \end{aligned}$$

(b_2) We take $T = \{e_1, e_2, ..., e_n\}$, the standard Euclidean basis in $\mathbb{R}^n$. Then

$$\sigma^2 = \sup_{t\in T} \sum_{j=1}^{n} t_j^2 = \sup_{t\in T} \|t\| = \sup_{t\in T} 1 = 1.$$

And

$$\sigma_\infty^2 = \sum_{j=1}^{n} \sup_{t\in T} t_j^2 = \sum_{j=1}^{n} \sup_{t\in T} 1 = n.$$

This clearly satisfies $\frac{\sigma^2}{\sigma_\infty^2} = n$. Hence the bound in ($a_2$) can only be improved by a factor of n as in (b_1).

4 Polynomial vs Exponential Moments Methods Bounds.

This exercise compares two ways of optimizing Markov's inequality: using polynomial moments or using exponential moments (Chernoff bounds). We will show that polynomial-moment opti-mization is always at least as good numerically, but is often much harder to use.

Y is a non-negative random variable and fix $t > 0$. Define $M(t) := \inf_{q \in \mathbb{Z}_+} \frac{\mathbb{E}[Y^q]}{t^q}$ and $C(t) := \inf_{\lambda > 0} \mathbb{E}[e^{\lambda(Y-t)}] = \inf_{\lambda > 0} e^{-\lambda t}\mathbb{E}[e^{\lambda Y}]$. Note that $M(t)$ and $C(t)$ both arise from applying markov's inequality to $Y \to Y^q$ and $Y \to e^{\lambda Y}$ respectively. Hence $\mathbb{P}[Y \geq t] \leq M(t)$ and $\mathbb{P}[Y \geq t] \leq C(t)$.

(a) **Polynomial Moments are at least as good.** Fix $\lambda > 0$ and note that the sequence $\mathbb{E}\left[\sum_{q=0}^{r} \frac{(\lambda Y)^q}{q!}\right]$ is increasing in r and it is always bounded by $\mathbb{E}[e^{\lambda Y}]$, hence by monotone convergence theorem, we can write $\mathbb{E}[e^{\lambda Y}] = \mathbb{E}\left[\sum_{q=0}^{\infty} \frac{(\lambda Y)^q}{q!}\right] = \sum_{q=0}^{\infty} \frac{\lambda^q}{q!}\mathbb{E}[Y^q]$. Therefore,

$$\begin{aligned} e^{-\lambda t}\mathbb{E}[e^{\lambda Y}] &= \sum_{q=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^q}{q!} \frac{\mathbb{E}[Y^q]}{t^q} \\ &\geq \left\{ \inf_{p \in \mathbb{Z}_+} \frac{\mathbb{E}[Y^p]}{t^p} \right\} \sum_{q=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^q}{q!} \\ &= M(t) \sum_{q=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^q}{q!} \\ &= M(t), \end{aligned}$$

since $\sum_{q=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^q}{q!} = 1$ [Poisson pmf].

(b) **Chernoff is still useful.** Even though polynomial moment bounds are theoretically at least as strong, Chernoff bounds are still useful in practice for several reasons:

   (a) **MGFs factor perfectly for independent sums.** If $Y = \sum_i X_i$ with independent random variables X_i, then $\mathbb{E}\left[e^{\lambda Y}\right] = \prod_i \mathbb{E}\left[e^{\lambda X_i}\right]$, which makes Chernoff bounds easy to compute and optimize.

   (b) **Stronger exponential tails.** Chernoff bounds directly yield sharp exponential decay in t, often with optimal constants.

   (c) **High-order moments are difficult to control.** Computing or bounding $\mathbb{E}[Y^q]$ for large q is typically much harder than bounding the moment generating function, whereas Chernoff bound can bound any polynomial moments.

   (d) **Optimization over a single parameter.** Optimizing over a single real parameter λ is usually simpler and more stable than choosing an optimal integer moment order q.

5 Sub gaussian Charecterization

This exercise proves several equivalent (up to constants) ways of expressing that a centered ran- dom variable has Gaussian-type tails. The three most common viewpoints are: (i) a quadratic CGF bound, (ii) a Gaussian tail bound, (iii) Gaussian-like moment growth. Let Z be a real-valued random variable with $\mathbb{E}[Z] = 0$. Define the centered CGF $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}] \in (-\infty, \infty], \lambda \in \mathbb{R}$.

(a) **Three Sub gaussian Properties.** For $v \geq 0$, consider the statements:

- (i) **The CGF Bound.** For all $\lambda \in \mathbb{R}$, $\psi_Z(\lambda) = \frac{v\lambda^2}{2}$.
- (ii) **The Tail Bound.** For all $t \geq 0$, $\mathbb{P}[|Z| \geq t] \leq 2\exp(-t^2/(2v))$.
- (iii) **The Moment Growth.** There exists an absolute constant $C > 0$ such that for all $p \geq 1$, $\mathbb{E}[|Z|^p]^{1/p} \leq C\sqrt{vp}$.

Our goal is to prove that (i), (ii), (iii) are equivalent up to universal constants by establishing the implication chain $(i) \to (ii) \to (iii) \to (i')$, where (i') is the same as (i) but with v replaced by cv, for a universal constant $c > 0$.

(a_1) $[(i) \implies (ii)]$ For $\lambda > 0$, using the function $\phi(u) = e^{\lambda u}$ on top of Markov's inequality, we get

$$\mathbb{P}[Z \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] = \exp[-\lambda t + \log \mathbb{E}[e^{\lambda Z}]] = \exp[-\lambda t + \psi_Z(\lambda)].$$

By noting $\psi_Z(\lambda) \leq v\lambda^2/2$, we have $\mathbb{P}[Z \geq t] \leq \exp[-\lambda t + v\lambda^2/2]$. The LHS of the last inequality does not depend on λ, hence $\mathbb{P}[Z \geq t] \leq \inf_{\lambda>0} \exp[-\lambda t + v\lambda^2/2]$. Since the expression $\exp[-\lambda t + v\lambda^2/2]$ is minimized at $\lambda = t/v$, by putting the value of λ we get

$$\mathbb{P}[Z \geq t] \leq \inf_{\lambda>0} \exp[-\lambda t + v\lambda^2/2] = \exp[-t^2/(2v)].$$

For the other part, using $-Z$ instead of Z, and with the same procedure, we get $\mathbb{P}[Z \leq -t] = \mathbb{P}[-Z \geq t] \leq \exp[-t^2/(2v)]$. Combining these two we get

$$\mathbb{P}[|Z| \geq t] \leq 2\exp[-t^2/(2v)].$$

(a_2) $[(ii) \implies (iii)]$ Let X be a bounded non-negative random variable with $0 \leq X \leq r$. Let $\Phi(r) = \int_0^r \mathbb{P}[X =$

$t]dt$ bet the CDF of X. Then

$$\begin{aligned}\mathbb{E}[X^p] &= \int_0^r x^p \mathbb{P}[X = x]dx \\ &= [x^p \Phi(x)]_0^r - \int_0^r px^{p-1}\Phi(x)dx \\ &= \int_0^r px^{p-1}\Phi(r)dx - \int_0^r px^{p-1}\Phi(x)dx \\ &= \int_0^r px^{p-1}[\Phi(r) - \Phi(x)]dx.\end{aligned}$$

Now by letting $r \to \infty$ and using monotone convergence theorem, we get [Note that $\Phi(r) \to 1$ as $r \to \infty$ and $\Phi(r) - \Phi(x) = \mathbb{P}[X \geq x]$],

$$\mathbb{E}[X^p] = \int_0^\infty px^{p-1}\mathbb{P}[X \geq x]dx.$$

Hence for any random variable Z, we have $\mathbb{E}[|Z|^p] = p\int_0^\infty z^{p-1}\mathbb{P}[|Z| \geq z]dz$. Using the tail bound from (ii) we get

$$\begin{aligned}\mathbb{E}[|Z|^p] &= p\int_0^\infty t^{p-1}\mathbb{P}(|Z| \geq t)\, dt \\ &\leq p\int_0^\infty t^{p-1}e^{-t^2/(2v)}\, dt.\end{aligned}$$

Make the change of variables $z = t^2/(2v)$, so that $t = \sqrt{2vz}$ and $dt = \frac{\sqrt{2v}}{2\sqrt{z}}\, dz$. Then

$$\begin{aligned}\mathbb{E}[|Z|^p] &\leq p\int_0^\infty (2v)^{(p-1)/2} z^{(p-1)/2} e^{-z}\frac{\sqrt{2v}}{2\sqrt{z}}\, dz \\ &= \frac{p}{2}(2v)^{p/2}\int_0^\infty z^{\frac{p}{2}-1}e^{-z}\, dz \\ &= \frac{p}{2}(2v)^{p/2}\Gamma\left(\frac{p}{2}\right).\end{aligned}$$

Using the standard bound on the Gamma function,

$$\Gamma(p) \leq C\, p^{p-\frac{1}{2}}e^{-p}, \qquad p \geq 1,$$

we obtain

$$\begin{aligned}\mathbb{E}[|Z|^p] &\leq C\, p(2v)^{p/2}\left(\frac{p}{2}\right)^{\frac{p}{2}-\frac{1}{2}} \\ &\leq C(vp)^{p/2}.\end{aligned}$$

Finally taking the pth root on both sides, we get $\mathbb{E}[|Z|^p]^{1/p} \leq C\sqrt{vp}$

(a_3) $[(iii) \implies (i')]$ We assume that $\mathbb{E}[|Z|^p]^{1/p} \leq C\sqrt{vp}$.

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &= \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[Z^k]}{k!} \\
&\leq \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[|Z|^k]}{k!} \\
&\leq \sum_{k=0}^{\infty} \frac{\lambda^k C^k (vk)^{k/2}}{k!} \\
&\leq \sum_{k=0}^{\infty} \frac{(\lambda C e \sqrt{v})^k}{k^{k/2}} \\
&\leq \sum_{k=0}^{\infty} \{\lambda C e \sqrt{v/2}\}^k \\
&\leq \exp\left((c_1 v \lambda^2/2)\right).
\end{aligned}$$

The last inequality is always true, for small λ, $\lambda C e\sqrt{v/2} < 1$, so the geometric series converges and it can be obviously bounded by $\exp\left((c_1 v\lambda^2/2)\right)$. Otherwise, for large λ, the series $\sum_{k=0}^{\infty}\{\lambda C e\sqrt{v/2}\}^k$ diverges but not more than as an exponential rate.

(b) **Variance Proxy.** We have $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$. Hence for small λ, $\psi_Z'(\lambda) = \mathbb{E}[Ze^{\lambda Z}]/\mathbb{E}[e^{\lambda Z}]$. So, $\psi_Z'(0) = \mathbb{E}[Z]$. Similarly, $\psi_Z^{"}(\lambda) = \mathbb{E}[Z^2 e^{\lambda Z}]/\mathbb{E}[e^{\lambda Z}] - \{\mathbb{E}[Ze^{\lambda Z}]/\mathbb{E}[e^{\lambda Z}]\}^2$. So, $\psi_Z^{"}(0) = \mathbb{E}[Z^2] - \{\mathbb{E}[Z]\}^2 = \mathrm{Var}(Z)$. Now using $\psi_Z(\lambda) \leq (v\lambda^2)/2$, we get

$$\mathrm{Var}(Z) = \psi_Z^{"}(\lambda)|_{\lambda=0} \leq \frac{d^2}{d\lambda^2}\frac{v\lambda^2}{2}|_{\lambda=0} = v.$$

Therefore $\mathrm{Var}(Z) \leq v$.

(c) **Exponential Square-Integrability.** We assume the sub-gaussian tail bound

$$\mathbb{P}[|Z| \geq t] \leq \exp(-t^2/2v), \quad t > 0.$$

Now, we know that for a non-negative random variable X, $\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}[X \geq t]dt$. Using $X = \exp(Z^2/cv)$, we

get

$$\begin{aligned}
\mathbb{E}[\exp(Z^2/cv)] &= \int_0^\infty \mathbb{P}[\exp(Z^2/cv) \geq t]dt \\
&= \int_0^\infty \mathbb{P}[Z^2 \geq cv\log(t)]dt \\
&= 1 + \int_1^\infty \mathbb{P}\left[|Z| \geq \sqrt{cv\log(t)}\right] dt \quad [\text{For } t \leq 1,\ \exp(Z^2/cv) \geq t] \\
&\leq 1 + \int_1^\infty 2\exp\left[-c\log(t)/2\right] dt \\
&= 1 + \int_1^\infty 2t^{-c/2} dt \\
&= 1 + \frac{4}{c-2} \quad [\text{for } c > 2] \\
&= 2 \quad [\text{for } c = 6].
\end{aligned}$$

For the converse, we can write,

$$\begin{aligned}
\mathbb{P}[|Z| \geq t] &= \mathbb{P}\left[\frac{Z^2}{cv} \geq \frac{t^2}{cv}\right] \\
&= \mathbb{P}\left[\exp(Z^2/cv) \geq \exp(t^2/cv)\right] \\
&\leq \exp(-t^2/cv)\mathbb{E}[Z^2/cv] \quad [\text{By Markov's Inequality}] \\
&\leq 2\exp(-t^2/cv).
\end{aligned}$$

Hence, we have shown that $\mathbb{P}[|Z| \geq t] \leq 2\exp(-t^2/cv)$, now using the implication $(ii) \to (iii) \to (i')$ from part (a), we can show that Z has a sub-Gaussian tail.