THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 7a: Efron Stein inequality

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

Efron Stein inequality

- Consider n independent random variables in some metric space $\mathcal{X}$.
- Consider a function $g : \mathcal{X}^n \to \mathbb{R}$
- Let $Z := g(X_1, \ldots, X_n)$
- We are interested in computing $\text{var}(g(X_1, \ldots, X_n))$
- Define $E_i(Z) = E[Z|X_{1:i-1}, X_{i+1:n}]$

An upper bound

Theorem

$$var(Z) \leq \sum_{i=1}^{n} E\left[Z - E_i[Z]\right]^2$$

Proof.

- For two arbitrary bounded random variables X, Y, we have: $E[XY] = E[E[XY|Y]] = E[YE[X|Y]]$
- Let $V := Z - E[Z]$
- Let $V_i := E[Z|X_{1:i}] - E[Z|X_{1:i-1}]$
- Clearly $V = \sum_i V_i$

□

Proof continued

$$\mathrm{var}(Z) = E\left[\sum_i V_i\right]^2 \tag{1}$$

$$= \sum_i E[V_i^2] + 2\sum_{i<j} E[V_i V_j] = \sum_i E[V_i^2] \tag{2}$$

- Why is the last step true? For $i > j$

$$\begin{aligned} E[V_i V_j] &= E[E[V_i V_j | X_1, \ldots, X_j]] \\ &= E[V_j E[V_i | X_1, \ldots, X_j]] = 0 \end{aligned}$$

Proof cont.

- Note that for three independent random variables X, Y, Z

$$E[g(X,Y,Z)|X] = E[[g(X,Y,Z)|X,Z]|X,Y]$$

$$\begin{aligned}
LHS &= \int_{y,z} g(x,y,z)f(y,z|x)dydz = \int_z (\int_y g(x,y,z)f(y|x,z)dy)f(z|x)dz \\
&= \int_z E[g(X,Y,Z)|X,Z]f(z|x)dz \\
&\overset{independence}{=} \int_z E[g(X,Y,Z)|X,Z]f(z|x,y)dz \\
&= E[E[g(X,Y,Z)|X,Z]|X,Y]
\end{aligned}$$

Proof cont.

-

$$
\begin{aligned}
V_i^2 &= (E[Z|X_{1:i}] - E[Z|X_{1:i-1}])^2 \\
&= (E[Z|X_{1:i}] - E[Z|X_{1:i-1}])^2 \\
&= (E[E[Z|X_{1:n}]|X_{1:i}] - E[E[Z|X_{1:i-1}, X_{i+1:n}]|X_{1:i}])^2 \\
&= (E[E[Z|X_{1:n}] - E[Z|X_{1:i-1}, X_{i+1:n}]|X_{1:i}])^2 \\
&= (E[Z - E_i Z|X_{1:i}])^2 \\
&\leq E[(Z - E_i Z)^2|X_{1:i}] \\
E[V_i^2] &\leq E[(Z - E_i Z)^2]
\end{aligned}
$$

The Efron Stein inequality

Theorem

Let $X'_1, \dots X'_n$ denote an independent copy of $X_1, \dots, X_n$. Let $Z'_i = g(X_{1:i-1}, X'_i, X_{i+1:n})$. We have:

$$var(Z) \leq \frac{1}{2} \sum_i E[(Z - Z'_i)^2].$$

Proof.

- If X, Y are iid, $\text{var}(X) = \frac{E[X - Y]^2}{2}$
- Conditioned on $X_{1:i-1}, X_{i:n}$, Z and Z'_i are independent and so

$$E_i[Z - Z'_i]^2 = \frac{E_i[Z - Z'_i]^2}{2}$$

$$\text{var}(Z) \leq \sum_{i=1}^{n} E\left[Z - E_i[Z]\right]^2 = \frac{E[E_i[Z - Z'_i]^2]}{2}$$

Remarks

- For $g(X_1, \ldots, X_n) = \sum_i X_i$ we have an equality.
- So in some sense, sums of independent random variables are the least concentrated functions
- Consider a function with the Bounded Difference property, i.e.

$$\sup_{x_{1:n}, x_i' \in \mathcal{X}} |g(x_1, \ldots, x_n) - g(x_{1:i-1} x_i' x_{i+1:n})| \leq c_i$$

- We have:

$$\text{var}(g(X)) \leq \frac{1}{2} \sum_i c_i^2$$

Example: longest common subsequence

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be two sequences of coin flips. Z is the length of the longest common subsequence.

$$Z = \max\{k : X_{i_1} = Y_{j_1}, \dots, X_{i_k} = Y_{j_k}\}$$

where $1 \le i_1 < i_2 \dots$ and $1 \le j_1 < j_2 \dots$.

- It is well known that $E[Z]/n \to \mu$ where $\mu \in [0.757, 0.837]$.
- If you change one bit of X, it can change Z by at most one, so,

$$\text{var}(Z) \le n/2$$

- So Z concentrates around its mean.

Uniform deviation

For $X_1, \ldots, X_n$ iid random variables, let $\hat{P}_n(A) = \frac{1}{n} 1(X_i \in A)$ and $P_n(A) = P(X_i \in A)$. We are interested in te quantity $Z := \sup_A |\hat{P}_n(A) - P_n(A)|$

- If we change one X_i, Z changes by $1/n$ at most.
- So $\mathrm{var}(Z) \leq \frac{1}{2n}$ by the Efron Stein inequality.
- Can we do better?

Uniform deviation

For $X_1, \ldots, X_n$ iid random variables, let

$$Z = sup_{f \in \mathcal{F}} \sum_j f(X_j).$$

For simplicity, assume $Ef[X_i] = 0$. We will show that the E/S inequality gives a much tighter upper bound that the one we just derived.

- $\text{var}(Z) \leq \frac{1}{2} \sum_i E[(Z - Z_i')^2]$
- Say f^* achieves the supremum for Z and f_* achieves the supremum for Z_i

$$\begin{aligned} f_*(X_i) - f_*(X_i') \leq Z - Z_i &\leq f^*(X_i) - f^*(X_i') \\ (Z - Z_i)^2 &\leq \max((f_*(X_i) - f_*(X_i'))^2, (f^*(X_i) - f^*(X_i'))^2) \\ &\leq \sup_{f \in \mathcal{F}} (f(X_i) - f(X_i'))^2 \end{aligned}$$

Uniform deviation

$$
\begin{aligned}
\mathrm{var}(Z) &\leq \frac{1}{2}\sum_i E\left[\sup_{f\in\mathcal{F}}(f(X_i) - f(X_i'))^2\right] \\
&\overset{(i)}{\leq} \sum_i E\left[\sup_{f\in\mathcal{F}}(f(X_i)^2 + f(X_i')^2)\right] \\
&\leq 2\sum_i E\sup_{f\in\mathcal{F}} f(X_i)^2
\end{aligned}
$$

- (i) uses $|2ab| \leq a^2 + b^2$
- If $f(X_i) \in [-1, 1]$ we get $\mathrm{var}(Z) \leq 2n$
- But if the maximum variance of $f(X_i)$ is small we have a significant improvement.

Self bounding functions

Definition

A non-negative function $g : \mathcal{X}^n] \to \mathcal{R}$ has the self bounding property if there exist functions $g_i : \mathcal{X}^{n-1} \to \mathcal{R}$ such that for all $x_1, \ldots, x_n \in \mathcal{X}$ and $i \in [n]$,

- $0 \leq g(x_1, \ldots, x_n) - g_i(x_{1:i-1}, x_{i+1:n}) \leq 1$
- $\sum_i \left(g(x_1, \ldots, x_n) - g_i(x_{1:i-1}, x_{i+1:n})\right) \leq g(x_1, \ldots, x_n)$

- Clearly, $\sum_i (g(x_{1:n}) - g_i(x_{1:i-1}, x_{i+1:n}))^2 \leq g(x_1, \ldots, x_n) =: Z$
- Now Theorem 1 gives:

$$\mathrm{var}(Z) \leq \sum_i E[(Z - E_i[Z])^2] \leq \sum_i E[(Z - g_i(x_{1:i-1}, x_{i+1:n}))^2] \leq E[g(x_{1:n})]$$

- So $\mathrm{var}(Z) \leq E[Z]$