Machine Learning Theory

Least Squares and the Efron-Stein Inequality

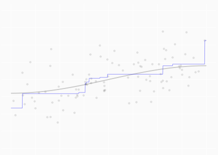
David A. Hirshberg

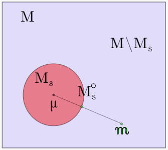
March 28, 2025

Emory University

Where We Left Things

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - m(X_i) \}^2 \quad \text{for a convex set } \mathcal{M}$$





Claim. When $Y_i = \mu(X_i) + \varepsilon_i$ for $\mu \in \mathcal{M}$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$,

$$\|\hat{\mu} - \mu\| < s \quad \text{w.p. } 1 - \delta \text{ if } \quad \frac{s^2}{2} \overset{(a)}{\geq} \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle + s\sigma \sqrt{\frac{2M_n}{\delta n}} \quad \text{for } M_n = 1 + 2\log(2n)$$

What We Actually Proved.

$$\|\hat{\mu} - \mu\| < s \quad \text{whenever} \quad \frac{s^2}{2} \overset{(b)}{\geq} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle$$

Loose End. w.p. $1 - \delta$, $(a) \implies (b)$. That is, …

$$\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle \leq \mathrm{E} \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle + s\sigma \sqrt{\frac{2M_n}{\delta n}} \quad \text{w.p.} \quad 1 - \delta.$$

Our Maximum is Approximately Constant

What we want to show.

$$Z = \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle \quad \text{satisfies} \quad Z \le \operatorname{E} Z + s\sigma\sqrt{\frac{2M_n}{\delta n}} \quad \text{w.p.} \quad 1-\delta \quad \text{for} \quad M_n = 1 + 2\log(2n).$$

We'll show something a bit stronger.

$$|Z - \operatorname{E} Z| \le s\sigma\sqrt{\frac{2M_n}{\delta n}} \quad \text{w.p.} \quad 1-\delta.$$

This is implied by Chebyshev's inequality. A special case of Markov's inequality.

$$\begin{aligned} &P\left\{|Z - \operatorname{E} Z| \le \frac{\operatorname{sd}(Z)}{\sqrt{\delta}}\right\} \\ &= P\left\{|Z - \operatorname{E} Z|^2 \le \frac{\operatorname{Var}(Z)}{\delta}\right\} \\ &\le \frac{\operatorname{E}|Z - \operatorname{E} Z|^2}{\frac{\operatorname{Var}(Z)}{\delta}} = \frac{\operatorname{Var}(Z)}{\frac{\operatorname{Var}(Z)}{\delta}} = \delta. \end{aligned}$$

All we need to do is bound the variance. We need to show that ...

$$\frac{\operatorname{sd}(Z)}{\sqrt{\delta}} \le s\sigma\sqrt{\frac{2M_n}{\delta n}} \quad \text{i.e.} \quad \operatorname{Var}(Z) \le s^2\sigma^2\frac{2M_n}{\delta n}.$$

Variance and Independent Copies

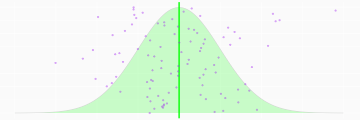
$$\mathrm{Var}[Z] = \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \max_{m \in \mathcal{M}_s^\circ} \sum_{i=1}^n u_i\{m(X_i) - \mu(X_i)\}.$$

- Z is a pretty complicated function of our noise vector ε. To bound its variance, …
- …we'll need to think about it a bit differently than you're probably used to.

$$\begin{aligned}\mathrm{Var}[Z] &= \mathrm{E}\left[\{Z - \mathrm{E}[Z]\}^2\right] \\ &= \frac{1}{2}\mathrm{E}\left[\left\{Z - \tilde{Z}\right\}^2\right]\end{aligned}$$

where Z and $\tilde{Z}$ are independent and identically distributed.

- It's the mean squared deviation of Z from its expectation.
- And *half* of the mean squared deviation of Z from *an independent copy* of Z.



Let's use all this to tackle a simplified version of our problem. We'll lose the $\max$.

$$\text{Calculate} \quad \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \sum_{i=1}^{n} u_i.$$

We can do this calculation the 'usual way' with these independent copies.

$$\begin{aligned}
\mathrm{Var}\left[f(\varepsilon)\right] &= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\tilde{\varepsilon}_i\right\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}(\varepsilon_i - \tilde{\varepsilon}_i)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)(\varepsilon_j - \tilde{\varepsilon}_j)\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)^2\right]
\end{aligned}$$

We can use our independent copies to write this more abstractly, keeping everything 'inside' our summing function f.

$$\begin{aligned}
\varepsilon_i - \tilde{\varepsilon}_i &= (\varepsilon_1 + \ldots + \varepsilon_i + \tilde{\varepsilon}_{i-1} + \ldots + \tilde{\varepsilon}_n) - (\varepsilon_1 + \ldots + \varepsilon_{i-1} + \tilde{\varepsilon}_i + \ldots + \tilde{\varepsilon}_n) \\
&= f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{where} \quad \varepsilon^{[i]} = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_i & \tilde{\varepsilon}_{i+1} & \tilde{\varepsilon}_{i+2} & \ldots & \tilde{\varepsilon}_n \end{pmatrix}
\end{aligned}$$

$$\text{Calculate} \quad \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \sum_{i=1}^{n} u_i.$$

We can do this calculation the 'usual way' with these independent copies.

$$\mathrm{Var}\left[f(\varepsilon)\right] = \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\tilde{\varepsilon}_i\right\}^2\right] \qquad = \frac{1}{2}\,\mathrm{E}\left[\{f(\varepsilon) - f(\tilde{\varepsilon})\}^2\right]$$

$$= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}(\varepsilon_i - \tilde{\varepsilon}_i)\right\}^2\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)(\varepsilon_j - \tilde{\varepsilon}_j)\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)^2\right]$$

We can use our independent copies to write this more abstractly, keeping everything 'inside' our summing function f.

$$\varepsilon_i - \tilde{\varepsilon}_i = (\varepsilon_1 + \ldots + \varepsilon_i + \tilde{\varepsilon}_{i-1} + \ldots + \tilde{\varepsilon}_n) - (\varepsilon_1 + \ldots + \varepsilon_{i-1} + \tilde{\varepsilon}_i + \ldots + \tilde{\varepsilon}_n)$$

$$= f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{where} \quad \varepsilon^{[i]} = \begin{pmatrix}\varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_i & \tilde{\varepsilon}_{i+1} & \tilde{\varepsilon}_{i+2} & \ldots & \tilde{\varepsilon}_n\end{pmatrix}$$

$$\text{Calculate} \quad \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \sum_{i=1}^{n} u_i.$$

We can do this calculation the ‘usual way’ with these independent copies.

$$\begin{aligned}
\mathrm{Var}\left[f(\varepsilon)\right] &= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\tilde{\varepsilon}_i\right\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}(\varepsilon_i - \tilde{\varepsilon}_i)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)(\varepsilon_j - \tilde{\varepsilon}_j)\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{E}\left[\{f(\varepsilon) - f(\tilde{\varepsilon})\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n} f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right]
\end{aligned}$$

We can use our independent copies to write this more abstractly,
keeping everything ‘inside’ our summing function f.

$$\begin{aligned}
\varepsilon_i - \tilde{\varepsilon}_i &= (\varepsilon_1 + \ldots + \varepsilon_i + \tilde{\varepsilon}_{i-1} + \ldots + \tilde{\varepsilon}_n) - (\varepsilon_1 + \ldots + \varepsilon_{i-1} + \tilde{\varepsilon}_i + \ldots + \tilde{\varepsilon}_n) \\
&= f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{where} \quad \varepsilon^{[i]} = \begin{pmatrix}\varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_i & \tilde{\varepsilon}_{i+1} & \tilde{\varepsilon}_{i+2} & \ldots & \tilde{\varepsilon}_n\end{pmatrix}
\end{aligned}$$

$$\text{Calculate} \quad \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \sum_{i=1}^{n} u_i.$$

We can do this calculation the ‘usual way’ with these independent copies.

$$\begin{aligned}
\mathrm{Var}\left[f(\varepsilon)\right] &= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\tilde{\varepsilon}_i\right\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}(\varepsilon_i - \tilde{\varepsilon}_i)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)(\varepsilon_j - \tilde{\varepsilon}_j)\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{E}\left[\left\{f(\varepsilon) - f(\tilde{\varepsilon})\right\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n} f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}\left\{f\left(\varepsilon^{[j]}\right)\right.\right.
\end{aligned}$$

We can use our independent copies to write this more abstractly,
keeping everything ‘inside’ our summing function f.

$$\begin{aligned}
\varepsilon_i - \tilde{\varepsilon}_i &= (\varepsilon_1 + \ldots + \varepsilon_i + \tilde{\varepsilon}_{i-1} + \ldots + \tilde{\varepsilon}_n) - (\varepsilon_1 + \ldots + \varepsilon_{i-1} + \tilde{\varepsilon}_i + \ldots + \tilde{\varepsilon}_n) \\
&= f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{where} \quad \varepsilon^{[i]} = \begin{pmatrix}\varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_i & \tilde{\varepsilon}_{i+1} & \tilde{\varepsilon}_{i+2} & \ldots & \tilde{\varepsilon}_n\end{pmatrix}
\end{aligned}$$

$$\text{Calculate} \quad \mathrm{Var}[f(\varepsilon)] \quad \text{for} \quad f(u) = \sum_{i=1}^{n} u_i.$$

We can do this calculation the ‘usual way’ with these independent copies.

$$\begin{aligned}
\mathrm{Var}\left[f(\varepsilon)\right] &= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}\varepsilon_i - \sum_{i=1}^{n}\tilde{\varepsilon}_i\right\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n}(\varepsilon_i - \tilde{\varepsilon}_i)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)(\varepsilon_j - \tilde{\varepsilon}_j)\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[(\varepsilon_i - \tilde{\varepsilon}_i)^2\right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\,\mathrm{E}\left[\{f(\varepsilon) - f(\tilde{\varepsilon})\}^2\right] \\
&= \frac{1}{2}\,\mathrm{E}\left[\left\{\sum_{i=1}^{n} f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}\left\{f\left(\varepsilon^{[j]}\right)\right.\right. \\
&= \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\left[\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right]
\end{aligned}$$

We can use our independent copies to write this more abstractly,
keeping everything ‘inside’ our summing function f.

$$\begin{aligned}
\varepsilon_i - \tilde{\varepsilon}_i &= (\varepsilon_1 + \ldots + \varepsilon_i + \tilde{\varepsilon}_{i-1} + \ldots + \tilde{\varepsilon}_n) - (\varepsilon_1 + \ldots + \varepsilon_{i-1} + \tilde{\varepsilon}_i + \ldots + \tilde{\varepsilon}_n) \\
&= f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{where} \quad \varepsilon^{[i]} = \begin{pmatrix}\varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_i & \tilde{\varepsilon}_{i+1} & \tilde{\varepsilon}_{i+2} & \ldots & \tilde{\varepsilon}_n\end{pmatrix}
\end{aligned}$$

The Variance of Sums: $\mathrm{Var}[f(\varepsilon)]$ for $f(u) = \sum_{i=1}^n u_i$

$$\mathrm{Var}[f(\varepsilon)] = \frac{1}{2}\sum_{i=1}^n \mathrm{E}\left[\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right] \qquad \text{for} \quad \varepsilon_j^{[i]} = \begin{cases} \varepsilon_j & j \le i \\ \tilde{\varepsilon}_j & j > i \end{cases}$$

$$= \frac{1}{2}\sum_{i=1}^n \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}^2\right] \qquad \text{for} \quad \varepsilon_j^{(i)} = \begin{cases} \tilde{\varepsilon}_i & j = i \\ \varepsilon_j & j \ne i \end{cases}$$

We can derive the (simpler) second formula from the one we've just worked out. Here's the argument.

- The pair of vectors $\varepsilon^{[i]}, \varepsilon^{[i-1]}$ have the same joint distribution as $\varepsilon, \varepsilon^{(i)}$.
- It follows that any functions of those pairs,

  $$\text{e.g.} \quad f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right) \quad \text{and} \quad f(\varepsilon) - f(\varepsilon^{(i)}),$$

  have the same distribution. And therefore the same expectation.

How do we know our pairs have the same distribution?

- The first vectors, $\varepsilon^{[i]}$ and ε, have the same distribution.
- To get the second vector from the first, we do the same thing.
  We replace the ith component with an independent copy.

The Efron-Stein inequality: $\mathrm{Var}[f(\varepsilon)]$ for *arbitrary* f

$$\mathrm{Var}\left[f(\varepsilon)\right] \le \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}^2\right] \quad \text{for} \quad \varepsilon_j^{(i)} = \begin{cases} \tilde{\varepsilon}_i & j = i \\ \varepsilon_j & j \ne i \end{cases}$$

- Something very cool happens when we write things this way.
  - What we've derived isn't just a new formula for the variance of a sum.
  - It's a variance bound for *any function* of a vector of independent random variables.
- We call this the *Efron-Stein inequality*.
- There's an equivalent 'positive part' version that's sometimes easier to use.

$$\mathrm{Var}\left[f(\varepsilon)\right] \le \sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}_+^2\right] \quad \text{for} \quad \{z\}_+ = \max\{z, 0\}.$$

- This is nice because $f(x) = \{x\}+^2$ is increasing (whereas $f(x) = x^2$ is not).
- And that means we can substitute an upper bound for what's inside it.

$$\mathrm{Var}\left[f(\varepsilon)\right] \le \sum_{i=1}^{n} \mathrm{E}\{F_i\}_+^2 \le \sum_{i=1}^{n} \mathrm{E}\, F_i^2 \quad \text{for} \quad F_i \ge f(\varepsilon) - f\left(\varepsilon^{(i)}\right).$$

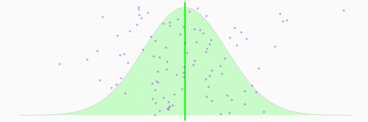
The 'Positive Part' Efron-Stein inequality

$$\begin{aligned}
\operatorname{Var}\left[f(\varepsilon)\right] &\le \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}^2\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}_+^2\right] \quad \text{for} \quad \{z\}_+ = \max\{z, 0\}.
\end{aligned}$$

- What's changed from the first formula to the second?
  - The differences on the right have been replaced with their *positive parts*.
  - We've lost the $\frac{1}{2}$ to compensate.
- Why is this equivalent? Symmetry.
- For any random variable S with a symmetric distribution[1], $\mathrm{E}\,S^2 = 2\,\mathrm{E}\{S\}_+^2$.

Proof.

$$\begin{aligned}
S^2 &= \{S\}_+^2 + \{-S\}_+^2 \\
&= \mathrm{E}\{S\}_+^2 + \mathrm{E}\{-S\}_+^2 \\
&= 2\,\mathrm{E}\{S\}_+^2.
\end{aligned}$$



□

[1] A random variable S has a symmetric distribution if S and $-S$ have the same distribution.

The Variance of our Maximum

$$\mathrm{Var}[f(\varepsilon)] \le \sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f\left(\varepsilon^{(i)}\right)\right\}_+^2\right] \quad \text{for} \quad f(x) = \max_{m \in \mathcal{M}_s^\circ} \langle x, m-\mu\rangle$$

What do the terms on the right look like?

$$\begin{aligned} f(\varepsilon) - f\left(\varepsilon^{(i)}\right) &= \max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu\rangle - \max_{m \in \mathcal{M}_s^\circ} \left\langle \varepsilon^{(i)},\ m-\mu\right\rangle \\ &\le \langle \varepsilon,\ \hat{m}-\mu\rangle - \left\langle \varepsilon^{(i)},\ \hat{m}-\mu\right\rangle \quad \text{for} \quad \hat{m} = \operatorname*{argmax}_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m-\mu\rangle \\ &= \left\langle \varepsilon - \varepsilon^{(i)},\ \hat{m}-\mu\right\rangle = \frac{1}{n}\{\hat{m}(X_i) - \mu(X_i)\}(\varepsilon_i - \tilde{\varepsilon}_i). \end{aligned}$$

Plugging in these bounds, we get ...

$$\begin{aligned} \mathrm{Var}[f(\varepsilon)] &\le \frac{1}{n} \times \mathrm{E}\,\frac{1}{n}\sum_{i=1}^{n} \{\underbrace{\hat{m}(X_i) - \mu(X_i)}_{U_i}\}^2 (\underbrace{\varepsilon_i - \tilde{\varepsilon}_i}_{V_i})^2 && = \frac{1}{n} \times \mathrm{E}\langle U, V\rangle_{L_2(\mathrm{P_n})} \\ &= \frac{1}{n} \times \frac{1}{n}\sum_{i=1}^{n} \{\hat{m}(X_i) - \mu(X_i)\}^2\, \mathrm{E} \max_{i \in 1\ldots n} (\varepsilon_i - \tilde{\varepsilon}_i)^2 && = \frac{1}{n} \times \mathrm{E}\|U\|_{L_1(\mathrm{P_n})}\|V\|_{L_\infty(\mathrm{P_n})} \\ &= \frac{1}{n} \times s^2 \times \mathrm{E} \max_{i \in 1\ldots n} (\varepsilon_i - \tilde{\varepsilon}_i)^2 \\ &\le \frac{1}{n} \times s^2 \times 2\sigma^2 M_n \quad \text{for} \quad M_n = 1 + 2\log(2n).^2 \end{aligned}$$

[2] M_n bounds the maximum of the squares of n independent standard normals. Scaling by $2\sigma^2$ gives a bound for normals with variance $\mathrm{Var}[\varepsilon_i - \tilde{\varepsilon}_i] = 2\sigma^2$.

A Proof of the Efron-Stein inequality

A Variance Formula

$$
\begin{aligned}
\operatorname{Var}[f(\varepsilon)] &= \mathrm{E}\, f(\varepsilon)^2 - \{\mathrm{E}\, f(\varepsilon)\}^2 \\
&= \mathrm{E}\, f(\varepsilon)^2 - \mathrm{E}\, f(\varepsilon)\, \mathrm{E}\, f(\tilde{\varepsilon}) \\
&= \mathrm{E}\, f(\varepsilon)\{f(\varepsilon) - \mathrm{E}\, f(\varepsilon)\} \\
&= \mathrm{E}\, f(\varepsilon)\left\{\sum_{i=1}^{n} f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\} \\
&= \sum_{i=1}^{n} \mathrm{E}\, f(\varepsilon)\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\} \quad \text{where} \quad \varepsilon_j^{[i]} = \begin{cases} \varepsilon_j & j \le i \\ \tilde{\varepsilon}_j & j > i \end{cases}
\end{aligned}
$$

The Swapping Trick

$$\mathrm{Var}\left[f(\varepsilon)\right] = \sum_{i=1}^{n} \mathrm{E}\, f(\varepsilon)\Big\{f\Big(\varepsilon^{[i]}\Big) - f\Big(\varepsilon^{[i-1]}\Big)\Big\} \quad \text{where} \quad \varepsilon_j^{[i]} = \begin{cases} \varepsilon_j & j \le i \\ \tilde{\varepsilon}_j & j > i \end{cases}$$

- Think of the ith term as a function of ε: $g_i(\varepsilon) = f(\varepsilon)\{f(\varepsilon^{[i]}) - f(\varepsilon^{[i-1]})\}$.
- Swapping $\varepsilon_i \to \tilde{\varepsilon}_i$ doesn't change the distribution of ε.
- So it doesn't change the distribution — or expectation — of $g_i(\varepsilon)$.

$$\overset{A_i = g_i(\varepsilon)}{f(\varepsilon)\Big\{f\Big(\varepsilon^{[i]}\Big) - f\Big(\varepsilon^{[i-1]}\Big)\Big\}} \to \overset{B_i = g_i(\varepsilon^{(i)})}{f(\varepsilon^{(i)})\Big\{f\Big(\varepsilon^{[i-1]}\Big) - f\Big(\tilde{\varepsilon}^{[i]}\Big)\Big\}} \quad \text{for} \quad \tilde{\varepsilon}_j^{(i)} = \begin{cases} \tilde{\varepsilon}_i & j = i \\ \varepsilon_j & j \ne i \end{cases}$$

$$= -f(\varepsilon^{(i)})\Big\{f\Big(\varepsilon^{[i]}\Big) - f\Big(\varepsilon^{[i-1]}\Big)\Big\}.$$

Because $A_i = B_i = (A_i + B_i)/2$, it follows that $\mathrm{Var}\left[f(\varepsilon)\right] = \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}[A_i + B_i]$ where

$$A_i + B_i = \Big\{f(\varepsilon) - f(\varepsilon^{(i)})\Big\}\Big\{f\Big(\varepsilon^{[i]}\Big) - f\Big(\varepsilon^{[i-1]}\Big)\Big\}$$

Finishing Up

$$
\begin{aligned}
\mathrm{Var}\left[f(\varepsilon)\right] &= \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f(\varepsilon^{(i)})\right\}\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}\right] \\
&\overset{(a)}{\leq} \frac{1}{2}\sqrt{\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f(\varepsilon^{(i)})\right\}^2\right] \sum_{i=1}^{n} \mathrm{E}\left[\left\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\right\}^2\right]} \\
&\overset{(b)}{=} \frac{1}{2}\sqrt{\left\{\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f(\varepsilon^{(i)})\right\}^2\right]\right\}^2} \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathrm{E}\left[\left\{f(\varepsilon) - f(\varepsilon^{(i)})\right\}^2\right].
\end{aligned}
$$

The rest boils down to

(a) Using the $\langle \cdot, \cdot \rangle_{L_2(\mathrm{P})}$ Cauchy-Schwarz bound on each term in the sum.

(b) Our observation, from a few slides back, that $\{f(\varepsilon) - f(\varepsilon^{(i)})\}^2$ and $\{f\left(\varepsilon^{[i]}\right) - f\left(\varepsilon^{[i-1]}\right)\}^2$ have the same distribution.

References

- Sourav Chatterjee's class Stein's method and applications.
  - The proof of the Efron-Stein inequality is based on lecture 10.
- Boucheron, Lugosi, and Massart's *Concentration inequalities: A nonasymptotic theory of independence*.
  - The bound on the variance of the maximum $\max_{m \in \mathcal{M}_s^\circ} \langle \varepsilon,\ m - \mu \rangle$ is based on Example 3.6 in Chapter 3.
  - The bound M_n on $\mathrm{E} \max_{i \in 1 \ldots n} \varepsilon_i^2$ for $\varepsilon_i \overset{iid}{\sim} N(0, 1)$ is from Lemma 11.3 in Chapter 11.