Variance Bounds

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-02-02.

1 Motivation

A basic phenomenon in high-dimensional statistics and machine learning is that many random variables of interest, under fairly weak assumptions, are *essentially deterministic*: with high probability they lie close to a typical value, such as the expectation or the median.

This phenomenon is already visible in the most classical setting. If $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 < \infty$, then the sample average

$$\bar{X}_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

satisfies $\bar{X}_n \to \mu$ in probability by the weak law of large numbers. Equivalently, for each $\varepsilon > 0$,

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \to 0,$$

so for large n the average is very likely to be close to its mean. Moreover, the central limit theorem gives a much sharper description of the *distribution* of the fluctuations:

$$\sqrt{n}\,(\bar{X}_n - \mu) \;\Rightarrow\; \mathcal{N}(0, \sigma^2),$$

which suggests that the typical scale of fluctuations of $\bar{X}_n$ is $\sigma/\sqrt{n}$.

A key message of concentration theory is that this behavior is not restricted to linear functions like $f(x_1, \ldots, x_n) = \frac{1}{n}\sum_i x_i$. Under fairly weak assumptions, many *nonlinear* functions

$$Z \;=\; f(X_1, \ldots, X_n)$$

of many independent (or weakly dependent) inputs also exhibit small fluctuations, especially when no single coordinate X_i has too much influence on the value of f. This is often summarized by the following principle (Talagrand, 1996):

> "A random variable that depends (in a 'smooth' way) on the influence of many independent variables (but not too much on any of them) is essentially constant."
>
> — Michel Talagrand (1996)

To make this principle precise, we have to decide what we mean by: (i) f depends *smoothly* on the influence of many random variables, and (ii) $f(X_1, \ldots, X_n)$ is *essentially* constant. Different choices lead to different tools and different inequalities.

Concretely, if Z is a real-valued random variable with mean $\mathbb{E}Z$ and (any) median $\mathbb{M}Z$, a *concentration inequality* is an estimate of the form

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant t\} \leqslant \cdots \qquad \text{or} \qquad \mathbb{P}\{|Z - \mathbb{M}Z| \geqslant t\} \leqslant \cdots,$$

valid for $t > 0$, where the right-hand side decays as t grows. A *tail bound* is one-sided (e.g. $\mathbb{P}\{Z - \mathbb{E}Z \geqslant t\} \leqslant \cdots$), and one may always combine upper and lower tail bounds into a two-sided concentration statement via the union bound.

In this lecture, we will make a modest (but foundational) start by focusing on the variance

$$\mathrm{Var}(Z) = \mathbb{E}\big(Z - \mathbb{E}Z\big)^2$$

as a proxy for the size of fluctuations, together with Chebyshev's inequality

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant t\} \leqslant \frac{\mathrm{Var}(Z)}{t^2}.$$

Thus, once we can bound $\mathrm{Var}(Z)$, we automatically obtain a (possibly crude) concentration bound.

The main theme for today will be:

Bound $\mathrm{Var}\big(f(X_1, \ldots, X_n)\big)$ *in terms of how much f can change when one input changes.*

For averages, variance bounds already capture the correct *scale* (e.g. $\sigma/\sqrt{n}$). However, variance alone usually does not yield sharp tail behavior. Starting next week, we will develop methods that control not only the size of fluctuations but also their *distribution*; for example, proving Gaussian-type or exponential-type tail bounds under appropriate assumptions. These results are closer in spirit to what the CLT provides for sums, but they apply far more broadly to nonlinear functions.

2 Variance: basic facts

Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathbb{E}$ denotes expectation. A real-valued random variable Z is in L_2 if $\mathbb{E}[Z^2] < \infty$. When $Z \in L_2$, its variance is

$$\mathrm{Var}(Z) := \mathbb{E}\big[(Z - \mathbb{E}Z)^2\big].$$

Expanding the square and using $\mathbb{E}[Z - \mathbb{E}Z] = 0$ gives the equivalent identity

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}Z)^2.$$

The most basic way to convert a variance bound into a concentration statement is Chebyshev's inequality: for every $t > 0$,

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant t\} \leqslant \frac{\mathrm{Var}(Z)}{t^2}. \tag{1}$$

Equivalently, for $t > 0$,

$$\mathbb{P}\Big\{|Z - \mathbb{E}Z| \geqslant t\sqrt{\mathrm{Var}(Z)}\Big\} \leqslant \min\{1, t^{-2}\}. \tag{2}$$

Chebyshev's inequality is extremely general (it only requires a finite second moment), but it only yields polynomial tail decay of order t^{-2}. Still, it already explains why "small variance $\Rightarrow$ concentration around the mean": taking $t = 10\sqrt{\mathrm{Var}(Z)}$ in (1) gives

$$\mathbb{P}\Big\{|Z - \mathbb{E}Z| \leqslant 10\sqrt{\mathrm{Var}(Z)}\Big\} \;\geqslant\; 1 - \frac{1}{10^2} \;=\; 0.99.$$

So whenever $\mathrm{Var}(Z)$ is small, the random variable Z is very likely to lie in a narrow window around $\mathbb{E}Z$.

The rest of this lecture is therefore devoted to methods for bounding $\mathrm{Var}(Z)$ when $Z = f(X_1, \ldots, X_n)$ depends on many independent inputs.

2.1 Two useful identities for variance

The definition $\mathrm{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}Z)^2]$ is not always the most convenient way to work. Here are two alternative viewpoints that will be especially useful once we start perturbing or resampling individual inputs.

1. *Variance as the best constant approximation.* For any $a \in \mathbb{R}$, expand

$$\begin{aligned}\mathbb{E}[(Z-a)^2] &= \mathbb{E}\big[(Z - \mathbb{E}Z + \mathbb{E}Z - a)^2\big] \\ &= \mathbb{E}[(Z-\mathbb{E}Z)^2] + 2(\mathbb{E}Z - a)\mathbb{E}[Z - \mathbb{E}Z] + (\mathbb{E}Z - a)^2 \\ &= \mathrm{Var}(Z) + (\mathbb{E}Z - a)^2.\end{aligned}$$

   In particular,

$$\mathrm{Var}(Z) = \min_{a\in\mathbb{R}} \mathbb{E}[(Z-a)^2], \qquad \text{with the unique minimizer } a^\star = \mathbb{E}Z. \tag{3}$$

   Interpretation: $\mathbb{E}Z$ is the best *constant* predictor of Z under squared loss.

2. *Variance as an average squared distance between two i.i.d. copies.* Let Z' be an independent copy of Z (i.e. $Z' \overset{d}{=} Z$ and $Z' \perp\!\!\!\perp Z$). Then

$$\begin{aligned}\mathbb{E}[(Z-Z')^2] &= \mathbb{E}[Z^2] + \mathbb{E}[(Z')^2] - 2\mathbb{E}[ZZ'] \\ &= 2\mathbb{E}[Z^2] - 2(\mathbb{E}Z)(\mathbb{E}Z') \\ &= 2\big(\mathbb{E}[Z^2] - (\mathbb{E}Z)^2\big) = 2\mathrm{Var}(Z),\end{aligned}$$

   so

$$\mathrm{Var}(Z) = \frac{1}{2}\,\mathbb{E}[(Z-Z')^2]. \tag{4}$$

   Interpretation: if $\mathrm{Var}(Z)$ is small, then two fresh samples of Z tend to be close in mean-square.

2.2 A basic variance bound: bounded range

A very simple (but surprisingly useful) principle is: if a random variable cannot move much, then its variance cannot be large.

If Z is (almost surely) bounded in an interval $[a, b]$, then

$$\mathrm{Var}(Z) \;\leqslant\; \frac{(b-a)^2}{4}. \tag{5}$$

One quick way to see this is to compare to the midpoint $m = (a+b)/2$: since $|Z-m| \leqslant (b-a)/2$ a.s., we have

$$\operatorname{Var}(Z) \;=\; \min_{a\in\mathbb{R}} \mathbb{E}[(Z-a)^2] \;\leqslant\; \mathbb{E}[(Z-m)^2] \;\leqslant\; \frac{(b-a)^2}{4}.$$

The bound (5) is tight when Z takes the values a and b each with probability $1/2$.

3 Examples: variance bounds for functions of many inputs

Before we develop general tools, it is helpful to keep a few canonical examples in mind. The first example (independent sums) is the benchmark case where everything can be computed exactly. The other examples are nonlinear: we simply state the variance bounds to highlight what is possible, and we will return later to the methods that make such results systematic.

1. The simplest linear example (independent sums). Let $(X_1, \ldots, X_n)$ be independent real random variables in L_2, and define

$$Z \;:=\; \sum_{i=1}^{n} X_i, \qquad \text{so that} \qquad \mathbb{E}Z \;=\; \sum_{i=1}^{n} \mathbb{E}X_i. \tag{6}$$

A basic computation shows that independence makes the variance *additive*:

$$\operatorname{Var}(Z) \;=\; \sum_{i=1}^{n} \operatorname{Var}(X_i). \tag{7}$$

(Indeed, $\operatorname{Var}(\sum_i X_i) \;=\; \sum_i \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j)$, and $\operatorname{Cov}(X_i, X_j) \;=\; 0$ for independent pairs.)

In the i.i.d. setting, if $\mathbb{E}X_1 = \mu$ and $\operatorname{Var}(X_1) = \sigma^2 < \infty$, then

$$\mathbb{E}Z = n\mu, \qquad \operatorname{Var}(Z) = n\sigma^2, \qquad \operatorname{Var}(\bar{X}_n) = \operatorname{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{\sigma^2}{n}.$$

This already captures the $\sigma/\sqrt{n}$ fluctuation scale highlighted by the CLT.

A useful comparison: variance versus range for bounded summands. Now suppose each X_i is almost surely supported in an interval $[a_i, b_i]$, and define

$$c_i \;:=\; b_i - a_i, \qquad c \;:=\; (c_1, \ldots, c_n) \in \mathbb{R}^n.$$

Applying $\operatorname{Var}(X_i) \leqslant c_i^2/4$ term-by-term and using (7),

$$\operatorname{Var}(Z) \leqslant \frac{1}{4}\|c\|_2^2, \qquad \text{so} \qquad \sqrt{\operatorname{Var}(Z)} \leqslant \frac{1}{2}\|c\|_2. \tag{8}$$

For comparison, the worst-case range of the sum satisfies

$$\operatorname{range}(Z) \;:=\; \sup Z - \inf Z \;\leqslant\; \sum_{i=1}^{n} c_i = \|c\|_1. \tag{9}$$

When $\|c\|_2 \ll \|c\|_1$ (e.g. many comparable summands), the standard deviation bound (8) can be much smaller than the worst-case range bound (9). As an example, for all ones vector, we have $\|c\|_2 = \sqrt{n}$, while $\|c\|_1 = n$. This is a first quantitative glimpse of Talagrand's principle: when the sum depends smoothly on many inputs, and no single input dominates, we have concentration.

2. A nonlinear example in combinatorial optimization (bin packing). Given numbers $x_1, \ldots, x_n \in [0,1]$, let $f(x_1, \ldots, x_n)$ be the minimum number of bins needed to pack the items so that the total weight in each bin is at most 1. For independent $X_i \in [0,1]$, define $Z = f(X_1, \ldots, X_n)$. One can show (using "bounded differences" ideas developed later) that

$$\operatorname{Var}(Z) \;\leqslant\; \frac{n}{4}.$$

This is already nontrivial because Z is defined through a combinatorial optimization problem.

3. A nonlinear example in random matrix theory (largest eigenvalue of a random symmetric matrix). Let A be an $n \times n$ symmetric random matrix with independent entries $\{X_{ij} : 1 \leqslant i \leqslant j \leqslant n\}$ and $|X_{ij}| \leqslant 1$ almost surely. Let $Z = \lambda_{\max}(A)$. Despite the fact that $\lambda_{\max}$ is defined by an optimization problem, one can show (using the Efron–Stein–Steele resampling method developed later) that

$$\operatorname{Var}(\lambda_{\max}(A)) \;\leqslant\; 16,$$

a dimension-free bound.

4. A nonlinear example in machine learning (kernel density estimation). Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with density ϕ. Fix a bandwidth $h > 0$ and a kernel $K \geqslant 0$ with $\int K = 1$, and define the KDE

$$\phi_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

For the L_1 error $Z = \|\phi - \phi_n\|_{L_1}$, one can show (again via bounded differences ideas) that

$$\operatorname{Var}(Z) \;\leqslant\; \frac{1}{n}.$$

The examples above all fit the same pattern: control the global fluctuations of $f(X_1, \ldots, X_n)$ by aggregating (in some way) the effects of changing one coordinate at a time. The rest of the lecture develops this idea systematically.

4 Tensorization of variance

The identity

$$\operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \operatorname{Var}(X_i)$$

for independent sums is extremely special, but it has a powerful "shadow" for *general* functions of independent inputs.

Let $X_1, \ldots, X_n$ be independent random variables and let

$$Z \;=\; f(X_1, \ldots, X_n)$$

be square-integrable. For each $i \in \{1, \ldots, n\}$, define the "all-but-i" vector

$$X^{(i)} \;:=\; (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n),$$

and introduce the conditional expectation operator

$$\mathbb{E}^{(i)}[\,\cdot\,] \;:=\; \mathbb{E}[\,\cdot \mid X^{(i)}].$$

We also write the conditional variance given $X^{(i)}$ as

$$\mathrm{Var}^{(i)}(Z) \;:=\; \mathrm{Var}(Z \mid X^{(i)}) \;=\; \mathbb{E}^{(i)}\big[(Z - \mathbb{E}^{(i)}Z)^2\big].$$

Note that $\mathrm{Var}^{(i)}(Z)$ is itself a random variable (it depends on $X^{(i)}$), and

$$\mathbb{E}[\mathrm{Var}^{(i)}(Z)] \;=\; \mathbb{E}\big[(Z - \mathbb{E}^{(i)}Z)^2\big]$$

by the tower property.

Theorem 4.1 (Tensorization of variance)**.** *With the notation above,*

$$\mathrm{Var}(Z) \;\leqslant\; \sum_{i=1}^{n} \mathbb{E}\big[\mathrm{Var}^{(i)}(Z)\big] \;=\; \sum_{i=1}^{n} \mathbb{E}\big[(Z - \mathbb{E}^{(i)}Z)^2\big]. \tag{10}$$

Proof. Let $\mathcal{F}_i := \sigma(X_1, \ldots, X_i)$ be the natural filtration, and write

$$\mathbb{E}_i[\,\cdot\,] \;:=\; \mathbb{E}[\,\cdot \mid \mathcal{F}_i], \qquad \text{with the convention } \mathbb{E}_0[\,\cdot\,] = \mathbb{E}[\,\cdot\,].$$

Define the Doob martingale $M_i := \mathbb{E}_i Z$, and the martingale differences

$$\Delta_i \;:=\; M_i - M_{i-1} \;=\; \mathbb{E}_i Z - \mathbb{E}_{i-1} Z, \qquad i = 1, \ldots, n.$$

Then

$$Z - \mathbb{E}Z \;=\; M_n - M_0 \;=\; \sum_{i=1}^{n} \Delta_i.$$

Martingale differences are orthogonal in L_2: for $j > i$,

$$\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}\big[\Delta_i \, \mathbb{E}(\Delta_j \mid \mathcal{F}_{j-1})\big] = 0,$$

because $\mathbb{E}(\Delta_j \mid \mathcal{F}_{j-1}) = 0$ and Δ_i is $\mathcal{F}_{j-1}$-measurable. Therefore,

$$\mathrm{Var}(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2 = \mathbb{E}\Big(\sum_{i=1}^{n} \Delta_i\Big)^2 = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2]. \tag{11}$$

Now we bring in independence. The key identity is

$$\mathbb{E}_i\big[\mathbb{E}^{(i)} Z\big] \;=\; \mathbb{E}_{i-1} Z. \tag{12}$$

Intuitively: $\mathbb{E}^{(i)}Z$ averages over X_i while holding all other coordinates fixed, so it does not depend on X_i; since X_i is independent of $X^{(i)}$, conditioning on X_i does not change that average, and then a tower-property argument collapses the conditioning down to $(X_1, \ldots, X_{i-1})$.

Using (12),

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z = \mathbb{E}_i Z - \mathbb{E}_i(\mathbb{E}^{(i)} Z) = \mathbb{E}_i\big[Z - \mathbb{E}^{(i)} Z\big].$$

By conditional Jensen (the square is convex),

$$\Delta_i^2 = \Big(\mathbb{E}_i[Z - \mathbb{E}^{(i)} Z]\Big)^2 \leqslant \mathbb{E}_i\big[(Z - \mathbb{E}^{(i)} Z)^2\big].$$

Taking expectations and using the tower property,

$$\mathbb{E}[\Delta_i^2] \leqslant \mathbb{E}\big[(Z - \mathbb{E}^{(i)} Z)^2\big] = \mathbb{E}\big[\mathrm{Var}^{(i)}(Z)\big].$$

Summing over i and combining with (11) yields (10). ∎

Quick consistency check (independent sums). If $Z = \sum_{j=1}^n X_j$, then

$$\mathbb{E}^{(i)} Z = \sum_{j \neq i} X_j + \mathbb{E} X_i, \qquad \text{so} \qquad Z - \mathbb{E}^{(i)} Z = X_i - \mathbb{E} X_i,$$

and (10) holds with equality.

5 Variants, applications, and examples

Recall the tensorization bound: for $Z = f(X_1, \dots, X_n) \in L_2$ with independent inputs,

$$\operatorname{Var}(Z) \;\leqslant\; \sum_{i=1}^n \mathbb{E}\big[\operatorname{Var}^{(i)}(Z)\big], \qquad \operatorname{Var}^{(i)}(Z) = \operatorname{Var}(Z \mid X^{(i)}). \tag{13}$$

The point of (13) is conceptual: it decomposes the total variance into coordinatewise contributions. In this section we turn this principle into easy-to-use bounds.

5.1 Application 1: bounded differences

A clean way to formalize "f does not change much when one coordinate changes" is to control the *coordinatewise range*.

Fix i and freeze all coordinates except X_i. Define the (random) coordinatewise range of f by

$$\begin{aligned}(D_i f)(X_1, \dots, X_n) := \sup_{x \in \operatorname{supp}(X_i)} & f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) \\ & - \inf_{x \in \operatorname{supp}(X_i)} f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n).\end{aligned}$$

(When distributions are continuous, it is sometimes cleaner to interpret sup / inf as essential sup/inf; the variance bound below is unchanged.)

Corollary 5.1 (Bounded differences: variance bound)**.** *For $Z = f(X_1, \dots, X_n) \in L_2$ with independent inputs,*

$$\operatorname{Var}(Z) \;\leqslant\; \frac{1}{4}\, \mathbb{E}\Big[\sum_{i=1}^n (D_i f)^2\Big]. \tag{14}$$

In particular, if $D_i f \leqslant c_i$ almost surely for deterministic constants $(c_i)_{i=1}^n$, then

$$\operatorname{Var}(Z) \;\leqslant\; \frac{1}{4} \sum_{i=1}^n c_i^2. \tag{15}$$

Proof. Fix i and condition on $X^{(i)}$. Under this conditioning, $Z = f(X_1, \dots, X_n)$ becomes a one-dimensional random variable as X_i varies, and it is supported in an interval of length $D_i f$. By the range bound (5) applied conditionally,

$$\operatorname{Var}^{(i)}(Z) \;\leqslant\; \frac{1}{4}\, (D_i f)^2.$$

Sum over i, take expectations, and apply (13) to obtain (14). The deterministic bound (15) follows immediately. □

Example: bin packing. Given $x_1, \dots, x_n \in [0,1]$, let $f(x_1, \dots, x_n)$ be the minimum number of bins needed to pack the items so that the sum in each bin is at most 1. Let $Z = f(X_1, \dots, X_n)$ for independent $X_i \in [0,1]$.

Changing one item can change the optimal number of bins by at most one: if we change x_i to x_i', we can take an optimal packing for $(x_1, \dots, x_n)$ and, if needed, place item i in a new bin, increasing the number of bins by at most 1. By symmetry (swap x_i and x_i'), we obtain $D_i f \leqslant 1$. Therefore,

$$\mathrm{Var}(Z) \;\leqslant\; \frac{1}{4}\sum_{i=1}^n 1^2 = \frac{n}{4}.$$

Example: kernel density estimation. Let $X_1, \dots, X_n$ be i.i.d. real samples from a distribution with density ϕ. Fix $h > 0$ and a kernel $K \geqslant 0$ with $\int_{\mathbb{R}} K(u)\,du = 1$, and define

$$\phi_n(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \qquad Z = \int_{\mathbb{R}} |\phi(x) - \phi_n(x)|\,dx = \|\phi - \phi_n\|_{L_1}.$$

Let $\phi_n^{(i)}$ be the KDE obtained by replacing X_i with some other value x_i'. By the reverse triangle inequality,

$$\big|\|\phi - \phi_n\|_{L_1} - \|\phi - \phi_n^{(i)}\|_{L_1}\big| \;\leqslant\; \|\phi_n - \phi_n^{(i)}\|_{L_1}.$$

But

$$\phi_n(x) - \phi_n^{(i)}(x) = \frac{1}{nh}\left(K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i'}{h}\right)\right),$$

so using $K \geqslant 0$ and $\int K = 1$,

$$\begin{aligned}\|\phi_n - \phi_n^{(i)}\|_{L_1} &\leqslant \frac{1}{nh}\int_{\mathbb{R}}\left|K\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i'}{h}\right)\right| dx \\ &\leqslant \frac{1}{nh}\left(\int_{\mathbb{R}} K\left(\frac{x - x_i}{h}\right) dx + \int_{\mathbb{R}} K\left(\frac{x - x_i'}{h}\right) dx\right) = \frac{2}{n}.\end{aligned}$$

Thus $D_i f \leqslant 2/n$, and (15) gives

$$\mathrm{Var}(Z) \;\leqslant\; \frac{1}{4}\sum_{i=1}^n \left(\frac{2}{n}\right)^2 \;=\; \frac{1}{n}.$$

5.2 Application 2: resampling coordinates (Efron–Stein–Steele)

Tensorization (13) involves the conditional mean $\mathbb{E}^{(i)}Z$, which is sometimes hard to compute explicitly. A common way to make it concrete is to introduce a resampled version of the ith coordinate.

Let $(X_1', \dots, X_n')$ be an independent copy of $(X_1, \dots, X_n)$, independent of everything else, and define

$$Z^{(i)} \;:=\; f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n), \qquad i = 1, \dots, n.$$

Conditional on $X^{(i)}$, the pair $(Z, Z^{(i)})$ is i.i.d. (it is obtained by plugging in two independent draws X_i and X_i' while keeping all other coordinates fixed). Therefore, applying (4) conditionally yields

$$\mathrm{Var}^{(i)}(Z) = \mathrm{Var}(Z \mid X^{(i)}) = \frac{1}{2}\,\mathbb{E}^{(i)}\big[(Z - Z^{(i)})^2\big].$$

Plugging this into (13) gives the classical resampling form.

Corollary 5.2 (Efron–Stein–Steele inequality)**.** *For $Z = f(X_1, \dots, X_n) \in L_2$ with independent inputs,*

$$\mathrm{Var}(Z) \;\leqslant\; \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\big[(Z - Z^{(i)})^2\big]. \tag{16}$$

A useful one-sided form. If W, W' are i.i.d., then $(W - W')$ is symmetric about 0, so

$$\frac{1}{2}\,\mathbb{E}[(W - W')^2] = \mathbb{E}[(W - W')_+^2] = \mathbb{E}[(W - W')_-^2], \qquad (a)_+ := \max\{a, 0\}.$$

Applying this identity conditionally (given $X^{(i)}$) shows

$$\mathrm{Var}^{(i)}(Z) = \mathbb{E}^{(i)}\big[(Z - Z^{(i)})_+^2\big] = \mathbb{E}^{(i)}\big[(Z - Z^{(i)})_-^2\big],$$

and therefore tensorization yields

$$\mathrm{Var}(Z) \;\leqslant\; \sum_{i=1}^{n} \mathbb{E}\big[(Z - Z^{(i)})_+^2\big] \qquad \text{and likewise} \qquad \mathrm{Var}(Z) \;\leqslant\; \sum_{i=1}^{n} \mathbb{E}\big[(Z - Z^{(i)})_-^2\big]. \tag{17}$$

The one-sided form is especially useful when we can upper bound $(Z - Z^{(i)})_+$ directly.

Example: largest eigenvalue of a random symmetric matrix. Let A be an $n \times n$ symmetric random matrix whose entries $\{X_{ij} : 1 \leqslant i \leqslant j \leqslant n\}$ are independent and satisfy $|X_{ij}| \leqslant 1$ almost surely (with $A_{ij} = A_{ji} = X_{ij}$). Let

$$Z := \lambda_{\max}(A) = \sup_{\|u\|_2 = 1} u^\top A u.$$

We treat each upper-triangular entry (i, j) as a "coordinate." For each $1 \leqslant i \leqslant j \leqslant n$, let X'_{ij} be an independent copy of X_{ij}, and let $A^{(ij)}$ be the matrix obtained by replacing X_{ij} (and X_{ji}) by X'_{ij} (and X'_{ji}), keeping all other entries fixed. Write

$$Z^{(ij)} := \lambda_{\max}(A^{(ij)}).$$

Let v be a unit eigenvector associated with Z, so $Z = v^\top A v$. Since $Z^{(ij)} = \sup_{\|u\|_2=1} u^\top A^{(ij)} u \geqslant v^\top A^{(ij)} v$, we have

$$(Z - Z^{(ij)})_+ \leqslant \big(v^\top A v - v^\top A^{(ij)} v\big)_+ = \big(v^\top (A - A^{(ij)}) v\big)_+.$$

If $i < j$, then $A - A^{(ij)}$ has only the (i, j) and (j, i) entries nonzero, so

$$v^\top (A - A^{(ij)}) v = 2 v_i v_j (X_{ij} - X'_{ij}).$$

If $i = j$, then $v^\top (A - A^{(ii)}) v = v_i^2 (X_{ii} - X'_{ii})$. In all cases, using $|X_{ij} - X'_{ij}| \leqslant 2$ yields the uniform bound

$$(Z - Z^{(ij)})_+ \;\leqslant\; 4\,|v_i v_j|, \qquad \text{hence} \qquad (Z - Z^{(ij)})_+^2 \;\leqslant\; 16\, v_i^2 v_j^2.$$

Summing over all $1 \leqslant i \leqslant j \leqslant n$,

$$\sum_{1 \leqslant i \leqslant j \leqslant n} (Z - Z^{(ij)})_+^2 \leqslant 16 \sum_{1 \leqslant i \leqslant j \leqslant n} v_i^2 v_j^2 \leqslant 16 \sum_{i=1}^{n} \sum_{j=1}^{n} v_i^2 v_j^2 = 16 \Big(\sum_{i=1}^{n} v_i^2\Big)^2 = 16,$$

since $\|v\|_2 = 1$. Taking expectations and applying the one-sided Efron–Stein bound (17) gives

$$\mathrm{Var}(\lambda_{\max}(A)) = \mathrm{Var}(Z) \;\leqslant\; \mathbb{E}\Big[\sum_{1\leqslant i\leqslant j\leqslant n} (Z - Z^{(ij)})_+^2\Big] \;\leqslant\; 16.$$

Thus the variance of the top eigenvalue is bounded by a universal constant, independent of the matrix dimension. (The same argument applies to $\lambda_{\min}(A)$, and with minor modifications to $\|A\|$.)

6 Look ahead

The results in this lecture show that changes of individual coordinates of $f(X_1, \ldots, X_n)$ control the variance of $f(X_1, \ldots, X_n)$. In other words, we are bounding the variance using some type of *discrete derivative*, which reflects the smoothness of f with respect to its inputs.

One may wonder whether it is possible to obtain variance bounds in terms of ordinary (calculus) derivatives, rather than discrete coordinate changes. In the next lecture, we will explore this angle via convex and Gaussian Poincaré inequalities, building upon our results from this lecture.

Source material

Parts of this lecture are based on references: Boucheron et al. (2013); Raginsky and Sason (2013); Tropp (2023); van Handel (2016), in addition to the author's accumulated experience working on related topics.

References

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press.

Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–247.

Talagrand, M. (1996). A new look at independence. *The Annals of Probability*, pages 1–34.

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

van Handel, R. (2016). Probability in high dimension. Lecture Notes (Princeton University). APC 550.