



# Variance, Covariance and Concentration

From the variance of a sum to Chebyshev, MGFs and Chernoff bounds

---

Probability Theory • Lecture 5

Applied Data Analysis — Module I

September 2026

# Roadmap & learning objectives

---

Lecture 4 computed a mean and a variance one distribution at a time. Today: the rules that make them useful, and what they guarantee.

- Finish what Lecture 3 started: the **rules of variance, covariance, the variance of a sum, and correlation** — every rule with its proof.
- See variance and correlation **in pictures**.
- Prove **Markov's** and **Chebyshev's** inequalities — what a mean, or a variance, guarantees about tails — and work through what they give and what they cost.
- Define the **moment generating function**; prove  $M^{(k)}(0) = \mathbb{E}[X^k]$ , uniqueness, and the sum rule: **convolutions become products**.
- Prove the **Chernoff bound** and state **Hoeffding's inequality**.
- ML tie-in: exponential tails replace Chebyshev's  $1/(n\epsilon^2)$  and shrink the test set you need.

## Where we stand after Lectures 3–4

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_x g(x)p_X(x); & \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}X)^2; \\ X \perp\!\!\!\perp Y &\Rightarrow \mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y; & \mathbb{P}(X \geq a) &\leq \mathbb{E}[X]/a.\end{aligned}$$

- Each distribution in Lecture 4 needed its own trick: indicators for the binomial, tail sums for the geometric, factorial moments for the Poisson.
- The PMF of a sum required a convolution — a sum inside a sum.

### Today's question

What does knowing  $\mathbb{E}[X]$  or  $\text{Var}(X)$  actually buy you — how does a sum behave, how sure can you be about a tail — and is there one object that stores every moment at once and turns “add independent variables” into “multiply”?



1.

# Variance, covariance and correlation

# Variance — what it measures

## Definition

Let  $\mu = \mathbb{E}[X]$ . The **variance** is  $\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$  and the **standard deviation** is  $\sigma_X = \sqrt{\text{Var}(X)}$ .

- Read it as a *mean squared distance from the centre*; the square root brings the units back.
- $\text{Var}(X) \geq 0$  always, and  $\text{Var}(X) = 0$  exactly when  $X$  is a constant.

## Theorem (The mean is the best constant)

For every constant  $c$ :  $\mathbb{E}[(X - c)^2] = \text{Var}(X) + (\mu - c)^2$ . Hence  $c = \mu$  minimizes the mean squared error.

## Proof.

Write  $X - c = (X - \mu) + (\mu - c)$  and expand the square:  $\mathbb{E}[(X - \mu)^2] + 2(\mu - c) \mathbb{E}[X - \mu] + (\mu - c)^2$ ; the middle term vanishes because  $\mathbb{E}[X - \mu] = 0$ . □

This is why the sample mean is the least-squares constant fit.

# Two rules you will use every day

## Theorem

(i)  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ .      (ii)  $\text{Var}(aX + b) = a^2 \text{Var}(X)$  for constants  $a, b$ .

## Proof.

(i)  $\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu \mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2$  by linearity.

(ii)  $aX + b$  has mean  $a\mu + b$ , so  $(aX + b) - (a\mu + b) = a(X - \mu)$  and

$\text{Var}(aX + b) = \mathbb{E}[a^2(X - \mu)^2] = a^2 \text{Var}(X)$ . □

- **Rule (i), a fair die:**  $\mathbb{E}[X] = 3.5$ ,  $\mathbb{E}[X^2] = \frac{91}{6}$ , so  $\text{Var}(X) = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \approx 2.92$ .
- **Rule (ii):** shifting ( $b$ ) changes nothing; scaling by  $a$  multiplies  $\sigma$  by  $|a|$  — temperatures in  $^{\circ}\text{F}$  have 1.8 times the standard deviation of the same data in  $^{\circ}\text{C}$ .

## The classic mistake

$\text{Var}(X + Y) \neq \text{Var}(X) + \text{Var}(Y)$  in general, and  $\text{Var}(2X) = 4 \text{Var}(X) \neq \text{Var}(X) + \text{Var}(X)$ : variance is not linear.

# Covariance — the definition

## Definition

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

## Theorem (Computational form)

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]; \text{ in particular } \text{Cov}(X, X) = \text{Var}(X).$$

## Proof.

$$\mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] = \mathbb{E}[XY] - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y = \mathbb{E}[XY] - \mu_X \mu_Y. \quad \square$$

- **Sign is the story.**  $X$  above its mean tends to come with  $Y$  above its mean  $\Rightarrow$  the product is mostly positive,  $\text{Cov} > 0$ ; opposite tendencies give  $\text{Cov} < 0$ .
- **Units are the problem.**  $\text{Cov}$  is in  $(\text{units of } X) \times (\text{units of } Y)$ , so its size means nothing on its own; correlation fixes this in three slides.
- **Two coins.**  $X$  = number of heads,  $Y = \mathbf{1}\{\text{first coin is heads}\}$ :  
 $\mathbb{E}[XY] = \mathbb{E}[X \mid Y = 1] \mathbb{P}(Y = 1) = 1.5 \cdot \frac{1}{2} = \frac{3}{4}$ , so  $\text{Cov}(X, Y) = \frac{3}{4} - 1 \cdot \frac{1}{2} = \frac{1}{4}$ .

# Properties of covariance

## Theorem

(1)  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ . (2)  $\text{Cov}(X, c) = 0$  for every constant  $c$ . (3) **Bilinearity:**  $\text{Cov}(aX + bY, Z) = a \text{Cov}(X, Z) + b \text{Cov}(Y, Z)$ . (4) If  $X \perp\!\!\!\perp Y$  then  $\text{Cov}(X, Y) = 0$ .

## Proof.

(1) the product is symmetric. (2)  $c - \mathbb{E}[c] = 0$ . (3)  $(aX + bY) - \mathbb{E}[aX + bY] = a(X - \mu_X) + b(Y - \mu_Y)$ ; multiply by  $(Z - \mu_Z)$ , take expectations, use linearity. (4) independence gives  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$  (Lecture 3), so the computational form is 0.  $\square$

## The converse of (4) is false

Let  $X$  be uniform on  $\{-1, 0, 1\}$  and  $Y = X^2$ . Then  $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0 = \mathbb{E}[X]\mathbb{E}[Y]$ , so  $\text{Cov}(X, Y) = 0$  — yet  $Y$  is a *function* of  $X$ . Covariance sees only *linear* dependence. Bilinearity (3) is the workhorse: it expands Cov of sums exactly as we expand products of polynomials.

# Variance of a sum

## Theorem

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y), \quad \text{Var} \left( \sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

If the  $X_i$  are pairwise uncorrelated — in particular if independent — then  $\text{Var} \left( \sum_i X_i \right) = \sum_i \text{Var}(X_i)$ .

## Proof.

With  $S = \sum_i X_i$ ,  $\text{Var}(S) = \text{Cov}(S, S) = \sum_i \sum_j \text{Cov}(X_i, X_j)$  by bilinearity in each slot. The diagonal terms  $i = j$  are the variances; the rest are the covariances. For two variables the off-diagonal part is  $\text{Cov}(X, Y) + \text{Cov}(Y, X) = 2 \text{Cov}(X, Y)$ . □

- Expectation is additive *always*; variance is additive only *without covariance*.
- Positive covariance inflates the spread of a sum, negative deflates it: diversification.

## Worked example — the sample mean

Let  $X_1, \dots, X_n$  be i.i.d. with mean  $\mu$  and variance  $\sigma^2$ , and  $\bar{X} = \frac{1}{n} \sum_i X_i$ .

### Theorem

$$\mathbb{E}[\bar{X}] = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \text{ so } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

### Proof.

Linearity:  $\mathbb{E}[\bar{X}] = \frac{1}{n} \sum_i \mu = \mu$ . For the variance use rule (ii) with  $a = \frac{1}{n}$ , then independence:

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var} \left( \sum_i X_i \right) = \frac{1}{n^2} \sum_i \sigma^2 = \frac{\sigma^2}{n}.$$



- **The square-root law.** To halve the noise you need *four* times the data; averaging 100 measurements divides the error by 10, not 100. This one line is why A/B tests are expensive.
- **The same law for a sum.**  $S_n = \sum_i X_i$  has mean  $n\mu$  and standard deviation  $\sigma\sqrt{n}$ : the total grows like  $n$ , its noise like  $\sqrt{n}$ , so the relative noise  $\rightarrow 0$  — the law of large numbers in embryo.
- **Chebyshev with this variance:**  $\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \leq \sigma^2 / (n\varepsilon^2)$  — the bound we improve on later today.

## Worked example — variances through indicators

Let  $X = \sum_{i=1}^n I_i$  with  $I_i = \mathbf{1}\{A_i\}$  and  $p_i = \mathbb{P}(A_i)$ . Then  $\text{Var}(I_i) = p_i(1 - p_i)$  and  $\text{Cov}(I_i, I_j) = \mathbb{P}(A_i \cap A_j) - p_i p_j$ .

### Independent trials: Binomial( $n, p$ )

All covariances vanish, so

$\text{Var}(X) = \sum_i p(1 - p) = np(1 - p)$  — Lecture 4's formula, now with a one-line proof.

### Dependent trials: matching

$n$  letters into  $n$  envelopes at random;  $X$  = number of correct matches,  $p_i = \frac{1}{n}$ ,  $\mathbb{P}(A_i \cap A_j) = \frac{1}{n(n-1)}$ . So  $\text{Cov}(I_i, I_j) = \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}$  and

$$\text{Var}(X) = n \cdot \frac{1}{n} \left(1 - \frac{1}{n}\right) + n(n-1) \cdot \frac{1}{n^2(n-1)} = 1.$$

- Mean and variance of the number of matches are both exactly 1 for every  $n \geq 2$  — a Poisson signature, and indeed the law tends to  $\text{Pois}(1)$ .
- The recipe is always the same: variances on the diagonal, pairwise covariances off it. No PMF required.

# Correlation

## Definition

For  $\sigma_X, \sigma_Y > 0$  the **correlation coefficient** is  $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \text{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right)$ .

## Theorem

**(1)**  $-1 \leq \rho \leq 1$ . **(2)**  $\rho = \pm 1$  if and only if  $Y = aX + b$  with probability one, with  $\text{sign}(a) = \rho$ . **(3)**  $\rho(aX + b, cY + d) = \text{sign}(ac) \rho(X, Y)$ : correlation ignores units.

## Proof.

Let  $U = (X - \mu_X)/\sigma_X$  and  $V = (Y - \mu_Y)/\sigma_Y$ ; both have variance 1, and  $\text{Cov}(U, V) = \rho$ . By the variance-of-a-sum formula,  $0 \leq \text{Var}(U \pm V) = 2(1 \pm \rho)$ , which gives (1).  $\rho = 1$  means  $\text{Var}(U - V) = 0$ , so  $U - V$  is constant and  $Y$  is affine in  $X$  with positive slope;  $\rho = -1$  likewise with  $U + V$ . (3) follows from bilinearity and rule (ii). This is Cauchy-Schwarz in probabilistic clothing:  $|\mathbb{E}[UV]| \leq \sqrt{\mathbb{E}[U^2]\mathbb{E}[V^2]}$ . □

# How to read a correlation

- $\rho$  measures **linear** association only:  $Y = X^2$  with symmetric  $X$  has  $\rho = 0$  and perfect dependence. Look at the scatter plot, not just the number.
- $\rho = 0.9$  is not “90% related”: a straight line explains  $\rho^2 = 81\%$  of the variance of  $Y$  — the  $R^2$  of simple regression is exactly  $\rho^2$ .
- **Correlation is not causation**: ice-cream sales and drownings are correlated through summer. A covariance is a symmetric number; there is no arrow in it.
- In ML, strongly correlated features are **redundant** (multicollinearity, double counting in distance-based methods). Look at the correlation matrix before you train.

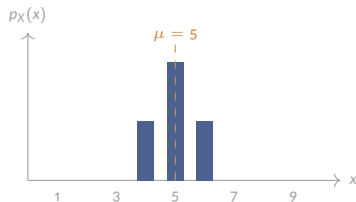
## The covariance matrix

For  $\mathbf{X} = (X_1, \dots, X_d)$ ,  $\Sigma_{ij} = \text{Cov}(X_i, X_j)$  is symmetric and positive semidefinite, because  $\mathbf{a}^\top \Sigma \mathbf{a} = \text{Var}(\mathbf{a}^\top \mathbf{X}) \geq 0$ . Its eigenvectors are the principal components.

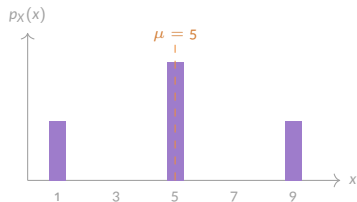
Variance: spread. Covariance: co-movement, in mixed units. Correlation: co-movement, unit-free, in  $[-1, 1]$  — and linear only.

## Seeing variance — same mean, different spread

Both distributions below have mean 5. The variance measures how far the probability mass sits from that mean.



Values 4, 5, 6 with probabilities  $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ :  $\text{Var} = 0.5$ ,  
 $\sigma \approx 0.71$ .

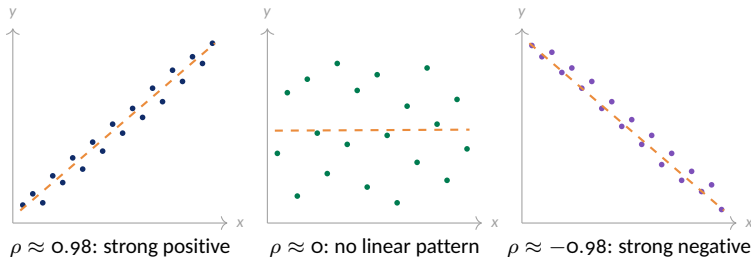


Values 1, 5, 9 with probabilities  $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ :  $\text{Var} = 8$ ,  
 $\sigma \approx 2.83$ .

- Same centre, same probabilities — only the distance of the side values from the mean changed, by a factor of 4; the variance changed by  $4^2 = 16$ .
- $\sigma$  is the natural unit of “how far from typical”: a value  $2\sigma$  away is ordinary on the right, impossible on the left. Chebyshev’s inequality, three slides on, makes this a guarantee.

# Seeing correlation — three clouds of points

Twenty pairs  $(x_i, y_i)$  in each panel. Correlation asks one question: does the cloud lie along a straight line?



- Left and right: knowing  $x$  almost pins down  $y$ ; the dashed line explains  $\rho^2 \approx 96\%$  of the variance of  $y$ . The sign is only the direction of the slope.
- Middle: the best line is flat,  $\rho \approx 0$  — but a U-shaped cloud would give  $\rho = 0$  too. Correlation sees lines and only lines.

2.

## Two inequalities: what a mean and a variance guarantee

# Markov's inequality — a bound from the mean alone

## Theorem (Markov)

If  $X \geq 0$  and  $a > 0$ , then  $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ .

## Proof.

$$(1) \text{ pointwise: } \mathbf{1}\{X \geq a\} \leq \frac{X}{a} \quad \text{because} \quad \begin{cases} X \geq a : & 1 \leq X/a, \\ X < a : & 0 \leq X/a \quad (X \geq 0); \end{cases}$$

$$(2) \text{ take } \mathbb{E}: \quad \mathbb{P}(X \geq a) = \mathbb{E}[\mathbf{1}\{X \geq a\}] \leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{1}{a} \mathbb{E}[X].$$



- Three ingredients:  $X \geq 0$ , a finite mean, monotonicity of  $\mathbb{E}$ . *No distribution was assumed.* With  $a = k \mathbb{E}[X]$ : at most a fraction  $1/k$  of the probability sits above  $k$  times the mean.
- **Tight:**  $X = a$  with probability  $\mu/a$ , else 0, gives  $\mathbb{E}[X] = \mu$  and  $\mathbb{P}(X \geq a) = \mu/a$  exactly.

# Markov — a worked example

A web service answers requests in  $T$  milliseconds,  $T \geq 0$ . The only thing we know is  $\mathbb{E}[T] = 200$ .

What can we guarantee about slow requests?

$$\mathbb{P}(T \geq 1000) \leq \frac{200}{1000} = 0.2: \text{ at most one request in five takes a second or more.}$$
$$\mathbb{P}(T \geq 2000) \leq 0.1, \quad \mathbb{P}(T \geq 10\,000) \leq 0.02.$$

- The bound falls only like  $1/a$ : to guarantee “fewer than 1%” you must go to  $a = 20\,000$  ms. Weak — but it holds for *every* latency distribution with that mean, including ones you have never seen.
- Compare with a model. If  $T \sim \text{Exp}(1/200)$  (Lecture 6), the truth is  $\mathbb{P}(T \geq 1000) = e^{-5} \approx 0.007$ : Markov overstates by a factor of 30. That gap is the price of knowing the mean and nothing else.
- Run it backwards as a **sanity check**: a report claiming “mean latency 200 ms” and “30% of requests above 1 s” contradicts itself — one of the two numbers is wrong.

# Chebyshev's inequality — a bound from the variance

## Theorem (Chebyshev)

If  $\mu = \mathbb{E}[X]$  and  $\text{Var}(X) = \sigma^2 < \infty$ , then for every  $k > 0$

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}, \quad \text{equivalently} \quad \mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

## Proof.

$$\{|X - \mu| \geq k\} = \{(X - \mu)^2 \geq k^2\} \quad (\text{both sides} \geq 0, \text{ squaring is monotone})$$

$$\mathbb{P}(|X - \mu| \geq k) = \mathbb{P}((X - \mu)^2 \geq k^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2} \quad (\text{Markov, } Y = (X - \mu)^2)$$

$$k = t\sigma : \quad \mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{\sigma^2}{t^2\sigma^2} = \frac{1}{t^2}.$$

- Markov applied to the squared deviation: two-sided, and decaying like  $1/k^2$  instead of  $1/k$ . Universal numbers: at least 75% of *any* distribution lies within  $2\sigma$  of its mean, at least 89% within  $3\sigma$  (the normal: 95% and 99.7%).



# Chebyshev — a worked example

Exam scores with mean 70 and standard deviation 10; the distribution itself is unknown.

## Guarantees from two numbers

$\mathbb{P}(|X - 70| \geq 20) \leq \frac{100}{400} = 0.25$ : at least 75% of students scored between 50 and 90.

$\mathbb{P}(|X - 70| \geq 30) \leq \frac{100}{900} \approx 0.11$ ; one-sided:  $\mathbb{P}(X \geq 100) \leq \mathbb{P}(|X - 70| \geq 30) \leq 0.11$ .

## The sample mean, revisited

Flip a fair coin  $n$  times and let  $\bar{X}$  be the proportion of heads:  $\mathbb{E}[\bar{X}] = \frac{1}{2}$ ,  $\text{Var}(\bar{X}) = \frac{1}{4n}$ . Then

$$\mathbb{P}(|\bar{X} - \tfrac{1}{2}| \geq 0.1) \leq \frac{1/(4n)}{0.01} = \frac{25}{n}.$$

With  $n = 1000$  flips the proportion is within  $\pm 0.1$  of  $\frac{1}{2}$  with probability at least 0.975, and the guarantee improves without limit as  $n$  grows — the **law of large numbers**, proved in one line. (The true probability is about  $10^{-10}$ : Chebyshev is enormously pessimistic; the MGF section does better.)



3.

# The moment generating function

# Definition — moment generating function

## Definition (MGF)

$M_X(t) = \mathbb{E}[e^{tx}]$ , for those  $t \in \mathbb{R}$  where the expectation is finite. We require it to be finite on an interval  $(-h, h)$  around 0.

- $M_X(0) = 1$  always. It is LOTUS with  $g(x) = e^{tx}$ : for a discrete  $X$ ,  $M_X(t) = \sum_x e^{tx} p_X(x)$ .
- It works for *any* random variable — negative values, continuous ones (Lecture 6) alike.
- The price: it need not exist. Existence near 0 is a genuine assumption, and a strong one (next slides).

## Two quick ones

Bernoulli( $p$ ):  $M(t) = 1 - p + pe^t$ .    Poisson( $\lambda$ ):  $M(t) = e^{\lambda(e^t - 1)}$ .

# Moments from the MGF

## Theorem

If  $M_X$  is finite on  $(-h, h)$ , then  $X$  has moments of all orders,  $M_X$  is infinitely differentiable at 0, and

$$M_X^{(k)}(0) = \mathbb{E}[X^k], \quad \text{equivalently} \quad M_X(t) = \sum_{k \geq 0} \frac{\mathbb{E}[X^k]}{k!} t^k.$$

## Proof.

Expand  $e^{tX} = \sum_k t^k X^k / k!$  and exchange sum and expectation (justified by absolute convergence on  $(-h, h)$ ). Comparing with the Taylor series of  $M_X$  identifies the coefficients.  $\square$

## Poisson, again

$M(t) = e^{\lambda(e^t - 1)}$ ,  $M'(t) = \lambda e^t M(t)$ , so  $M'(0) = \lambda$ ;  $M''(t) = \lambda e^t M(t) + \lambda^2 e^{2t} M(t)$  gives  $\mathbb{E}[X^2] = \lambda + \lambda^2$  and  $\text{Var}(X) = \lambda$ .

# Uniqueness, sums, and affine maps

## Theorem

(1) **Uniqueness:** if  $M_X = M_Y$  on an interval around 0, then  $X \stackrel{d}{=} Y$ . (2) **Sums:**  $X \perp\!\!\!\perp Y \Rightarrow M_{X+Y}(t) = M_X(t)M_Y(t)$ . (3) **Affine:**  $M_{aX+b}(t) = e^{bt}M_X(at)$ .

## Proof.

(2)  $\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]$  by independence. (3)  $\mathbb{E}[e^{t(aX+b)}] = e^{bt}\mathbb{E}[e^{(at)X}]$ . (1) is a genuine theorem of analysis (inversion of the Laplace transform) — we use it, we do not prove it here.  $\square$

- These three lines replace every convolution computation in Lecture 4.
- Standardization  $Z = (X - \mu)/\sigma$  is an affine map, so  $M_Z(t) = e^{-\mu t/\sigma}M_X(t/\sigma)$  — the first step of every CLT proof (Lecture 14).

# A catalogue of MGFs

| Distribution              | $M_X(t)$                       | valid for         | $\mathbb{E}[X]$ |
|---------------------------|--------------------------------|-------------------|-----------------|
| Bernoulli( $p$ )          | $1 - p + pe^t$                 | all $t$           | $p$             |
| Binomial( $n, p$ )        | $(1 - p + pe^t)^n$             | all $t$           | $np$            |
| Geometric( $p$ )          | $\frac{pe^t}{1 - (1 - p)e^t}$  | $e^t < 1/(1 - p)$ | $1/p$           |
| Poisson( $\lambda$ )      | $e^{\lambda(e^t - 1)}$         | all $t$           | $\lambda$       |
| Normal( $\mu, \sigma^2$ ) | $e^{\mu t + \sigma^2 t^2 / 2}$ | all $t$           | $\mu$           |

- Read the table sideways: the binomial MGF is the Bernoulli MGF to the  $n$ -th power — the sum rule of the previous slide, and the binomial theorem in one line.
- The normal row is a preview (Lecture 8), but it is the one to remember:  $\log M_X(t)$  is a *quadratic*, which is what makes the normal the fixed point of the CLT.
- The geometric row shows the typical failure mode: the MGF exists only on part of the line.

## When the MGF does not exist

- For the geometric,  $\mathbb{E}[e^{tx}] = \infty$  once  $e^t \geq 1/(1-p)$ : the tail  $(1-p)^k$  decays geometrically, and  $e^{tk}$  eventually beats it.
- For a **lognormal** or a Pareto with any tail index,  $M_X(t) = \infty$  for every  $t > 0$ . The St. Petersburg payoff of Lecture 3 has no MGF either — it has no mean.
- Finiteness of  $M_X$  near 0 forces **sub-exponential tails**:  $\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t)$  decays exponentially. That is a modelling assumption, not a formality.

### The fix

The **characteristic function**  $\varphi_X(t) = \mathbb{E}[e^{itx}]$  always exists, since  $|e^{itx}| = 1$ . It keeps uniqueness and the product rule, and it is the tool that proves the CLT in Lecture 14. Everything today has a characteristic-function twin.

Rule of thumb: for heavy-tailed returns, claim sizes or waiting times, do not reach for a Chernoff bound.



4.

## Concentration: the Chernoff method

# The Chernoff bound

## Theorem (Chernoff)

For every  $t > 0$ ,  $\mathbb{P}(X \geq a) \leq e^{-ta} M_X(t)$ , and therefore

$$\mathbb{P}(X \geq a) \leq \exp \left( - \sup_{t>0} \{ta - \log M_X(t)\} \right).$$

## Proof.

$x \mapsto e^{tx}$  is increasing for  $t > 0$ , so  $\{X \geq a\} = \{e^{tX} \geq e^{ta}\}$ . Markov's inequality applied to the nonnegative variable  $e^{tX}$  gives  $\mathbb{P}(e^{tX} \geq e^{ta}) \leq e^{-ta} \mathbb{E}[e^{tX}]$ . The bound holds for every  $t > 0$ , so we may take the best one. □

- The exponent is the best trade-off between the growth of  $e^{ta}$  and of  $M_X(t)$  — the rate function of large deviations.
- Markov used one moment and gave  $1/a$ ; Chebyshev used two and gave  $1/a^2$ ; the MGF uses all of them and gives  $e^{-ca}$ .

# From Chernoff to Hoeffding

## Theorem (Hoeffding's lemma)

If  $\mathbb{E}[Z] = 0$  and  $Z \in [c, d]$  almost surely, then  $M_Z(t) \leq e^{t^2(d-c)^2/8}$  for all  $t \in \mathbb{R}$ .

## Theorem (Hoeffding's inequality)

Let  $Z_1, \dots, Z_n$  be independent with  $Z_i \in [c_i, d_i]$  and let  $\bar{Z}$  be their average. Then

$$\mathbb{P}(|\bar{Z} - \mathbb{E}\bar{Z}| \geq \varepsilon) \leq 2 \exp \left( -\frac{2n^2\varepsilon^2}{\sum_i (d_i - c_i)^2} \right), \quad \text{and for } Z_i \in [0, 1]: \quad 2e^{-2n\varepsilon^2}.$$

## Proof sketch.

Centre the variables, apply Chernoff to the sum, factorize the MGF by independence, bound each factor with the lemma, optimize the quadratic in  $t$ , and add the two tails. Bounded losses — 0/1 error, clipped scores — satisfy the hypothesis exactly.  $\square$

# Three bounds on the same tail

A fair coin flipped  $n = 100$  times; how unlikely is 60 or more heads? Exact answer: 0.0284.

| Tool                 | What it uses | Bound on $\mathbb{P}(X \geq 60)$ | Comment                          |
|----------------------|--------------|----------------------------------|----------------------------------|
| Markov               | the mean     | $\leq 0.83$                      | almost no information            |
| Chebyshev            | the variance | $\leq 0.25$                      | one-sided, still crude           |
| Chernoff / Hoeffding | the MGF      | $\leq 0.135$                     | right shape: $e^{-2n\epsilon^2}$ |
| Exact (binomial)     | everything   | 0.0284                           | the target                       |

- The gap between the last two rows is the price of a bound that holds for *every* bounded distribution.
- What matters is not the constant but the dependence:  $1/(n\epsilon^2)$  versus  $e^{-2n\epsilon^2}$ . Polynomial decay needs a hundred times more data than exponential decay for the same guarantee.



5.

# Machine learning connections

# How many test points do you actually need?

A classifier has true error  $R$ ; the test error  $\hat{R}_n$  is an average of  $n$  independent 0/1 losses.

## Chebyshev (Lecture 3)

$\mathbb{P}(|\hat{R}_n - R| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$ . For  $\varepsilon = 0.01$  at confidence 95%:  $n \geq 50,000$ .

## Hoeffding (today)

$\mathbb{P}(|\hat{R}_n - R| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$ . The same guarantee needs  $n \geq \frac{\ln(2/0.05)}{2\varepsilon^2} \approx 18,400$ .

- Nearly a factor of three, from nothing but a better use of the same data — the MGF exploits boundedness, the variance does not.
- Solve the bound for  $\varepsilon$  instead: with  $n = 10,000$  test points you can certify  $\pm 0.014$  at 95%. Report that, not three decimal places of accuracy.
- The exponential form is also what survives a union bound over many models:  $\log |H|$  enters additively, which is why capacity is a *count* (Lecture 13).



6.

# Wrap-up

## Pitfalls & summary

- Expectation is linear; **variance is not**.  $\text{Var}(2X) = 4 \text{Var}(X)$ , and a sum picks up every covariance.
- Correlation measures **linear** association only, and  $\rho = 0$  does not mean independent.
- **Markov** needs  $X \geq 0$  and a mean; **Chebyshev** needs a variance. Both hold for every distribution — and both are crude for exactly that reason.
- The **MGF** may fail to exist; the characteristic function always works. Products of MGFs require **independence** — the one hypothesis linearity of expectation never needed.
- A Chernoff bound is only as good as the  $t$  you choose: always optimize, and never use it on heavy tails.

$$\text{Var} \left( \sum_i X_i \right) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j); \quad \mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}; \quad M_X(t) = \mathbb{E}[e^{tX}];$$

moments from derivatives, sums from products.

# Seminar problems & next lecture

- 1  $X$  is a fair die. Compute  $\text{Var}(X)$  via  $\mathbb{E}[X^2]$ ; then  $\text{Var}(X_1 + X_2)$  for two dice and  $\text{Var}(2X_1)$ . Explain the difference.
- 2  $X$  = number of heads in three tosses,  $Y = \mathbf{1}\{\text{first toss is heads}\}$ . Find  $\text{Cov}(X, Y)$  and  $\rho(X, Y)$ .
- 3 Show  $\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Var}(Y)$ . For i.i.d.  $X, Y$  the sum and the difference are uncorrelated — are they independent?
- 4  $X \geq 0$  with  $\mathbb{E}[X] = 3$ . Bound  $\mathbb{P}(X \geq 12)$ ; construct an  $X$  for which the bound is attained.
- 5 Scores have mean 50 and s.d. 5. Bound  $\mathbb{P}(X \leq 35 \text{ or } X \geq 65)$ . How many i.i.d. observations make  $\mathbb{P}(|\bar{X} - 50| \geq 1) \leq 0.05$  by Chebyshev?
- 6  $X \sim \text{Bin}(n, p)$ ,  $Y \sim \text{Bin}(m, p)$  independent. Prove  $X + Y \sim \text{Bin}(n + m, p)$  with one line of MGFs.
- 7 Compute the MGF of a fair die and use Chernoff to bound  $\mathbb{P}(S_{100} \geq 400)$  for the sum of 100 dice. Compare with Chebyshev.
- 8 A classifier is right with probability 0.8. Using Hoeffding, how many test points make  $\mathbb{P}(|\hat{R}_n - R| \geq 0.02) \leq 0.01$ ?

*Next: continuous random variables — densities, the exponential, and how to sample anything.*



# Further reading

## Core textbooks — variance, covariance, the inequalities, MGFs

- S. Ross, *A First Course in Probability*, 10th ed., Pearson, 2019 — Ch. 7 (covariance, correlation, MGFs), Ch. 8 (Markov and Chebyshev).
- J. Blitzstein, J. Hwang, *Introduction to Probability*, 2nd ed., CRC Press, 2019 — Ch. 6 (moments, MGFs), Ch. 7 (covariance, correlation), Ch. 10 (inequalities). Free at [probabilitybook.net](http://probabilitybook.net).
- D. Bertsekas, J. Tsitsiklis, *Introduction to Probability*, 2nd ed., Athena Scientific, 2008 — Ch. 4 (transforms, covariance), Ch. 5 (Markov, Chebyshev, the weak law).
- G. Grimmett, D. Stirzaker, *Probability and Random Processes*, 4th ed., Oxford University Press, 2020 — Ch. 5, for the generating-function material we skipped.
- A. N. Shiryaev, *Probability-1*, 3rd ed., Graduate Texts in Mathematics 95, Springer, 2016 — the rigorous classic.

## Concentration inequalities — Chernoff, Hoeffding and beyond

- L. Wasserman, *All of Statistics*, Springer, 2004 — Ch. 4, a compact tour of the inequalities used in ML.
- S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013 — Ch. 2, the Chernoff method done properly.
- R. Vershynin, *High-Dimensional Probability*, Cambridge University Press, 2018 — Ch. 2 (sub-gaussian variables, Hoeffding); free draft on the author's page.