Lectures 3 and 4: The Central Limit Theorem

Swapping, the Lindeberg condition, and Stein's method

OUTLINE maximum entropy $\longrightarrow$ why a Gaussian at all $\longrightarrow$ what a limit in distribution is $\longrightarrow$ swap one summand at a time $\longrightarrow$ the CLT, and Lindeberg's condition

$\mathbb{E}[Wf(W)] = \mathbb{E}[f'(W)]$ $\longrightarrow$ vector and matrix forms $\longrightarrow$ Stein's equation $\longrightarrow$ the CLT with a rate $\longrightarrow$ the same argument for Poisson

HOW TO READ Sections marked $*$ are not lectured. They are complete and they are examinable. Exercises are typed A, B and C; each week's quiz has one of each. What is collected is listed in a box at the end of each section; everything else is a pool — available and examinable, not collected.

1 Why Gaussians appear, and the CLT by swapping

Q. Why should the Gaussian, of all laws, be the one that appears?

1.1 Notation

Throughout, $X_1, X_2, \ldots$ are independent and identically distributed with mean 0 and variance 1, and

$$S_n \;:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

The normalization is chosen so that $\mathrm{Var}(S_n) = 1$ for every n. We write $g \sim \mathcal{N}(0,1)$, and γ for its law.

From §1.5* of Lectures 1–2. For a random variable X on $\mathbb{R}$ with density p, the *differential entropy* is $h(X) = -\int p \log p$, and the *relative entropy* of X from γ is $D(X\|\gamma) = \int p \log(p/\phi)$, where ϕ is the standard normal density. Both may be $+\infty$, and h is undefined when X has no density.

1.2 What "converging to a Gaussian" means

S_n does not converge to a random variable; what converges is its law, and we test that against functions.

Definition 1.1 (Convergence in distribution)**.** For random vectors in $\mathbb{R}^k$, $X_n \Rightarrow X$ means $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every bounded continuous $f : \mathbb{R}^k \to \mathbb{R}$.

Distribution functions need not converge at a jump: if $X_n \equiv 1/n$ and $X \equiv 0$ then $X_n \Rightarrow X$, yet $F_n(0) = 0$ for every n while $F(0) = 1$.

A smooth f can be Taylor expanded and $\mathbf{1}_{(-\infty,z]}$ cannot, so the proofs below test against smooth functions only. That is already enough.

Lemma 1.2. *Suppose $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every $f \in C^3(\mathbb{R})$ with f, f', f'', f''' bounded. Then $F_{X_n}(z) \to F_X(z)$ at every z where F_X is continuous.*

Proof. Fix such a z and $\delta > 0$. Let $\psi : \mathbb{R} \to [0,1]$ be smooth with $\psi = 1$ on $(-\infty, 0]$ and $\psi = 0$ on $[1,\infty)$, and put

$$u^+(x) = \psi\Big(\frac{x-z}{\delta}\Big), \qquad u^-(x) = \psi\Big(\frac{x-z+\delta}{\delta}\Big).$$

Both lie in the test class, and

$$\mathbf{1}_{(-\infty,z-\delta]} \ \le\ u^- \ \le\ \mathbf{1}_{(-\infty,z]} \ \le\ u^+ \ \le\ \mathbf{1}_{(-\infty,z+\delta]}.$$

Taking expectations at X_n and passing to the limit,

$$F_X(z-\delta) \ \le\ \mathbb{E}u^-(X) = \lim_n \mathbb{E}u^-(X_n) \ \le\ \liminf_n F_{X_n}(z),$$

$$\limsup_n F_{X_n}(z) \ \le\ \lim_n \mathbb{E}u^+(X_n) = \mathbb{E}u^+(X) \ \le\ F_X(z+\delta).$$

Let $\delta \downarrow 0$ and use continuity of F_X at z. ∎

A rate against smooth f is not a rate against probabilities; see §2.3.

Exercise 1.3 (TYPE C)**.** Let X_n, X be random vectors in $\mathbb{R}^k$ with $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every $f \in C^3(\mathbb{R}^k)$ whose derivatives up to order three are bounded. Show $X_n \Rightarrow X$. *Such an f is neither smooth nor compactly supported; repair each defect in turn, and note that the hypothesis already limits how much mass X_n can put far out.*

1.3 The moral answer

Section 1.5* of Lectures 1–2 supplies two facts. First, at fixed variance the Gaussian maximizes entropy: $h(X) \le h(g)$ whenever $\mathrm{Var}(X) = 1$. Second, for symmetric i.i.d. X, Y,

$$h\Big(\tfrac{X+Y}{\sqrt{2}}\Big) \ \ge\ h(X),$$

because $(x,y) \mapsto \big(\tfrac{x+y}{\sqrt{2}}, \tfrac{x-y}{\sqrt{2}}\big)$ is orthogonal, orthogonal maps preserve differential entropy, and entropy is subadditive.

Since $S_{2^k} = \frac{1}{\sqrt{2}}(S'_{2^{k-1}} + S''_{2^{k-1}})$ for two independent copies, $h(S_{2^k})$ increases in k and never exceeds $h(g)$. So it converges, and the maximizer is the only candidate for the limit.

1.4 The entropic CLT

Theorem 1.4 (Barron, 1986)**.** *If $D(S_m\|\gamma) < \infty$ for some m, then $D(S_n\|\gamma) \to 0$; equivalently $h(S_n) \to h(g)$.*

Theorem 1.5 (Artstein–Ball–Barthe–Naor, 2004)**.** *$h(S_n)$ is increasing in n, not merely along powers of two.*

We will not prove either.

Example 1.6 (The entropic CLT says nothing about a coin flip)**.** Let $X_i = \pm 1$ with probability $\frac{1}{2}$ each. Then S_n is supported on a finite set for every n, so it has no density, $h(S_n)$ is undefined, and $D(S_n\|\gamma) = \infty$ for every m and n. The hypothesis of Barron's theorem never holds.

The classical central limit theorem holds for ± 1 regardless, so we need a proof that covers Example 1.6.

1.5 Lindeberg's swapping argument

Theorem 1.7 (Central limit theorem)**.** *Suppose in addition* $\mathbb{E}|X_1|^3 < \infty$*. Then for every* $f \in C^3(\mathbb{R})$ *with bounded third derivative,*

$$\big|\mathbb{E}f(S_n) - \mathbb{E}f(g)\big| \;\le\; \frac{\|f'''\|_\infty}{6\sqrt{n}}\Big(\mathbb{E}|X_1|^3 + \mathbb{E}|g|^3\Big).$$

Proof. Let $g_1, \dots, g_n$ be i.i.d. standard Gaussians, independent of the X_i, and interpolate:

$$Z_k \;=\; \frac{1}{\sqrt{n}}\big(g_1 + \cdots + g_k + X_{k+1} + \cdots + X_n\big), \qquad k = 0, \dots, n,$$

so $Z_0 = S_n$ and $Z_n \sim \mathcal{N}(0,1)$. Telescoping,

$$\mathbb{E}f(S_n) - \mathbb{E}f(g) \;=\; \sum_{k=1}^{n}\Big(\mathbb{E}f(Z_{k-1}) - \mathbb{E}f(Z_k)\Big).$$

Fix k and write $W_k = \frac{1}{\sqrt{n}}\big(g_1 + \cdots + g_{k-1} + X_{k+1} + \cdots + X_n\big)$, which is independent of both X_k and g_k. Then $Z_{k-1} = W_k + X_k/\sqrt{n}$ and $Z_k = W_k + g_k/\sqrt{n}$, and Taylor's theorem gives

$$f(W_k + u) = f(W_k) + uf'(W_k) + \tfrac{u^2}{2}f''(W_k) + R(u), \qquad |R(u)| \le \tfrac{\|f'''\|_\infty}{6}|u|^3.$$

Take expectations of both expansions and subtract. The constant terms cancel. The linear terms cancel because W_k is independent of X_k and g_k and both have mean 0. The quadratic terms cancel because both have variance 1. Only the remainders survive:

$$\big|\mathbb{E}f(Z_{k-1}) - \mathbb{E}f(Z_k)\big| \;\le\; \frac{\|f'''\|_\infty}{6\,n^{3/2}}\Big(\mathbb{E}|X_k|^3 + \mathbb{E}|g_k|^3\Big).$$

Summing the n terms gives the claim. ■

The proof used only Taylor's theorem; in particular it covers $X_i = \pm 1$.

With Lemma 1.2, $\mathbb{P}(S_n \le z) \to \Phi(z)$ for every z, where Φ is the standard normal distribution function.

The classical argument — expand $\varphi_{X_1}(t/\sqrt{n})$ and take nth powers — is carried out in §1.9*.

1.6 What the proof actually needed

The argument never used that the X_i share a law, nor that they are one-dimensional: the same swap with multivariate Taylor gives the multidimensional central limit theorem. What it did use is that no single summand carries much of the variance.

Definition 1.8 (Lindeberg's condition)**.** Let $\{X_{n,i}\}_{i \le k_n}$ be a triangular array, independent within rows, with

$$\mathbb{E}X_{n,i} = 0, \qquad s_n^2 = \sum_{i \le k_n} \mathrm{Var}(X_{n,i}), \qquad k_n \to \infty, \qquad s_n \to \infty.$$

The array satisfies *Lindeberg's condition* if for every $\varepsilon > 0$

$$L_n(\varepsilon) \;:=\; \frac{1}{s_n^2}\sum_i \mathbb{E}\Big[X_{n,i}^2\,\mathbf{1}\{|X_{n,i}| > \varepsilon s_n\}\Big] \;\longrightarrow\; 0.$$

Normalize once and for all:

$$Y_{n,i} = \frac{X_{n,i}}{s_n}, \qquad \sigma_{n,i}^2 = \mathrm{Var}(Y_{n,i}), \qquad T_n = \sum_{i \le k_n} Y_{n,i},$$

so that $\sum_i \sigma_{n,i}^2 = 1$ and $L_n(\varepsilon) = \sum_i \mathbb{E}[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon\}]$.

Theorem 1.9 (Lindeberg)**.** *Under Lindeberg's condition, $T_n \Rightarrow \mathcal{N}(0,1)$.*

Proof. Let $f \in C^3(\mathbb{R})$ with f'' and f''' bounded. Taylor's theorem to second and to third order gives, uniformly in w,

$$f(w+u) = f(w) + uf'(w) + \tfrac{u^2}{2} f''(w) + R(u), \qquad |R(u)| \le \min\Big(\tfrac{\|f'''\|_\infty}{6}|u|^3,\ \|f''\|_\infty u^2\Big), \tag{1}$$

the second bound from $R(u) = \frac{u^2}{2}\big(f''(w+\theta u) - f''(w)\big)$ for some $\theta \in (0,1)$; it costs a power of u and does not involve f'''.

Splitting the second moment of $Y_{n,i}$ at ε gives, for every i and every $\varepsilon > 0$,

$$\sigma_{n,i}^2 = \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| \le \varepsilon\}\big] + \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon\}\big] \ \le\ \varepsilon^2 + L_n(\varepsilon), \tag{2}$$

so $\limsup_n \max_i \sigma_{n,i}^2 \le \varepsilon^2$ for every ε, that is $\max_i \sigma_{n,i} \to 0$.

Let $g_{n,i} \sim \mathcal{N}(0, \sigma_{n,i}^2)$ be independent of each other and of the array, so that $\sum_i g_{n,i} \sim \mathcal{N}(0,1)$, and swap one summand at a time as in Theorem 1.7: with

$$W_{n,i} = g_{n,1} + \cdots + g_{n,i-1} + Y_{n,i+1} + \cdots + Y_{n,k_n},$$

which is independent of $Y_{n,i}$ and of $g_{n,i}$, the telescoping sum gives

$$\big|\mathbb{E}f(T_n) - \mathbb{E}f(g)\big| \ \le\ \sum_i \Big|\mathbb{E}f\big(W_{n,i} + Y_{n,i}\big) - \mathbb{E}f\big(W_{n,i} + g_{n,i}\big)\Big|.$$

Expand both arguments by (1). The constant, linear and quadratic terms cancel term by term, because $Y_{n,i}$ and $g_{n,i}$ have the same first two moments and both are independent of $W_{n,i}$; only remainders survive. On the Y side use the cubic bound where $|Y_{n,i}| \le \varepsilon$ and the quadratic one where $|Y_{n,i}| > \varepsilon$:

$$\sum_i \mathbb{E}\big|R(Y_{n,i})\big| \le \frac{\|f'''\|_\infty}{6}\varepsilon \sum_i \mathbb{E}[Y_{n,i}^2] + \|f''\|_\infty \sum_i \mathbb{E}\big[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon\}\big] = \frac{\|f'''\|_\infty}{6}\varepsilon + \|f''\|_\infty L_n(\varepsilon).$$

On the Gaussian side there is nothing to truncate: $\mathbb{E}|g_{n,i}|^3 = 2\sqrt{2/\pi}\,\sigma_{n,i}^3$ and $\sum_i \sigma_{n,i}^3 \le \max_i \sigma_{n,i}$, so

$$\sum_i \mathbb{E}\big|R(g_{n,i})\big| \ \le\ \frac{\|f'''\|_\infty}{3}\sqrt{\tfrac{2}{\pi}}\, \max_i \sigma_{n,i}.$$

Adding the two and letting $n \to \infty$: $L_n(\varepsilon) \to 0$ by hypothesis and $\max_i \sigma_{n,i} \to 0$ by (2), so

$$\limsup_{n\to\infty} \big|\mathbb{E}f(T_n) - \mathbb{E}f(g)\big| \ \le\ \frac{\|f'''\|_\infty}{6}\,\varepsilon$$

for every $\varepsilon > 0$, and the left side is 0. Lemma 1.2 converts this into convergence in distribution. ∎

Remark 1.10 (Feller's condition). Say the array satisfies *Feller's condition* if

$$\max_{i \le k_n} \frac{\mathrm{Var}(X_{n,i})}{s_n^2} = \max_{i \le k_n} \sigma_{n,i}^2 \longrightarrow 0.$$

Three precise statements relate it to Definition 1.8.

(a) Lindeberg's condition implies Feller's. This is exactly (2), and it is the only place the proof above used anything beyond $L_n(\varepsilon) \to 0$.

(b) The implication is strict: Exercise 1.14 exhibits an array satisfying Feller's condition and not Lindeberg's.

(c) *Converse, stated without proof.* If Feller's condition holds and $T_n \Rightarrow \mathcal{N}(0,1)$, then Lindeberg's condition holds.

With Theorem 1.9, (c) says that under Feller's condition Lindeberg's condition is equivalent to the central limit theorem. The pair is the *Lindeberg–Feller theorem.*

Remark 1.11 (Lyapunov's condition). In practice one checks a cruder condition. If for some $\delta > 0$

$$\frac{1}{s_n^{2+\delta}} \sum_i \mathbb{E}|X_{n,i}|^{2+\delta} \longrightarrow 0,$$

then Lindeberg's condition holds: on $\{|X_{n,i}| > \varepsilon s_n\}$ one may write $X_{n,i}^2 \le |X_{n,i}|^{2+\delta}/(\varepsilon s_n)^\delta$. For a weighted sum $X_{n,i} = a_{n,i}\xi_i$ of i.i.d. noise with $\mathbb{E}|\xi_1|^3 < \infty$ this reads

$$L_n(\varepsilon) \le \frac{\mathbb{E}|\xi_1|^3}{\varepsilon} \left(\frac{\|a_n\|_3}{\|a_n\|_2} \right)^3,$$

so such a sum is asymptotically Gaussian whenever $\|a_n\|_3/\|a_n\|_2 \to 0$. This fails exactly when a bounded number of coordinates carries a fixed share of $\|a_n\|_2^2$.

Lindeberg's condition separates two regimes: many small contributions give a Gaussian, a few rare ones give a Poisson. Section 1.9* computes both sides of the boundary; §2.4 proves the Poisson statement, with a rate.

Remark 1.12 (The limit is universal, the rate is not). Kolmogorov–Smirnov distance from S_{30} to $\mathcal{N}(0,1)$, by simulation: uniform 0.0014, exponential 0.025, ± 1 0.073. The limit does not depend on the law of X_1; the speed does, and lattice summands are slowest.

Exercise 1.13 (TYPE A)**.** Let $X_1, X_2, \ldots$ be i.i.d. with mean 0 and variance $\sigma^2 \in (0,\infty)$, and set $X_{n,i} = X_i$ for $i \le n$. Show $L_n(\varepsilon) = \sigma^{-2}\mathbb{E}[X_1^2 \mathbf{1}\{|X_1| > \varepsilon\sigma\sqrt{n}\}]$ and that it tends to 0. *Dominated convergence, with dominating variable X_1^2.* Conclude that Theorem 1.9 contains Theorem 1.7 without the third moment.

Exercise 1.14 (TYPE C)**.** Let $X_{n,i}$, $i \le n$, be independent, equal to ± 1 with probability $\frac{1}{2}(1 - \frac{1}{n})$ each and to $\pm\sqrt{n}$ with probability $\frac{1}{2n}$ each. Show $s_n^2 = 2n - 1$, that Feller's condition holds, and that $L_n(\varepsilon) \to \frac{1}{2}$ for every $\varepsilon \le 1/\sqrt{2}$. Then compute $\lim_n \varphi_{T_n}(t)$ and identify the limit law. *Remark 1.10(c) shows the limit cannot be $\mathcal{N}(0,1)$ before any computation. Half the variance sits in the ± 1's and half in the rare $\pm\sqrt{n}$'s.*

1.7 Records, and why random search plateaus

The hypothesis $s_n \to \infty$ of Definition 1.8 has the following consequence.

Lemma 1.15. *If the array of Definition 1.8 is uniformly bounded — $|X_{n,i}| \le C$ for all n and i — then Lindeberg's condition holds.*

Proof. Fix $\varepsilon > 0$. Since $s_n \to \infty$ we have $\varepsilon s_n > C$ for all large n, and then the event $\{|X_{n,i}| > \varepsilon s_n\}$ is empty for every i, so $L_n(\varepsilon) = 0$. ∎

The next array is not a rescaled sum of identically distributed summands. Let $X_1, X_2, \dots$ be i.i.d. with a continuous law, so that $\mathbb{P}(X_i = X_j) = 0$ for $i \ne j$, and let

$$B_i = \mathbf{1}\big\{X_i > \max(X_1, \dots, X_{i-1})\big\}, \qquad B_1 = 1,$$

so that $R_n = \sum_{i \le n} B_i$ counts the *records* in the first n observations: the times at which the running maximum improves.

Lemma 1.16 (Rényi)**.** *$B_1, \dots, B_n$ are independent, and $\mathbb{P}(B_i = 1) = 1/i$.*

Proof. Two rank vectors. Let

$$r_i = \#\{j \le n : X_j \le X_i\}, \qquad \rho_i = \#\{j \le i : X_j \le X_i\},$$

so $r \in S_n$ records the relative order of the whole sample, ρ_i is the rank of X_i among the first i observations only, and $B_i = \mathbf{1}\{\rho_i = i\}$. Note $\rho_i \in \{1, \dots, i\}$, so ρ takes values in

$$\mathcal{C} = \{1\} \times \{1, 2\} \times \cdots \times \{1, \dots, n\}.$$

(For $n = 3$ with $X_2 < X_3 < X_1$: $r = (3, 1, 2)$, $\rho = (1, 1, 2)$, and only X_1 is a record.)

Step 1: r is uniform on S_n. The X_i are i.i.d. and ties have probability 0, so permuting the sample does not change its law; hence $\mathbb{P}(r = \pi)$ is the same for all $\pi \in S_n$, and there are $n!$ of them.

Step 2: $r \mapsto \rho$ is a bijection $S_n \to \mathcal{C}$. Both sets have $n!$ elements, so it is enough that the map be injective, i.e. that r be recoverable from ρ. Induct on n. Since $\rho_n = r_n$, the position of the last observation in the order is known. Deleting it leaves $X_1, \dots, X_{n-1}$, whose own rank vector maps to $(\rho_1, \dots, \rho_{n-1})$ — each ρ_i with $i < n$ involves only the first i observations — and is therefore determined by induction. Reinserting X_n at rank ρ_n recovers r.

Step 3. A bijection carries the uniform law to the uniform law, so ρ is uniform on $\mathcal{C}$. Being uniform on a product set is exactly being independent with uniform coordinates: $\rho_1, \dots, \rho_n$ are independent and ρ_i is uniform on $\{1, \dots, i\}$. Hence the $B_i = \mathbf{1}\{\rho_i = i\}$ are independent with $\mathbb{P}(B_i = 1) = 1/i$. ∎

Theorem 1.17 (Records)**.** *Let $H_n = \sum_{i \le n} \frac{1}{i}$ and $s_n^2 = H_n - \sum_{i \le n} \frac{1}{i^2}$. Then*

$$\frac{R_n - H_n}{s_n} \implies \mathcal{N}(0, 1).$$

Proof. Apply Theorem 1.9 to the array $X_{n,i} = B_i - \frac{1}{i}$, $i \le n$, which is independent by Lemma 1.16 and centered, with $\mathrm{Var}(X_{n,i}) = \frac{1}{i}(1 - \frac{1}{i})$ and hence the stated s_n^2. Since $s_n^2 \ge H_n - \frac{\pi^2}{6} \to \infty$ and $|X_{n,i}| \le 1$, Lemma 1.15 supplies Lindeberg's condition. ∎

The variances $\frac{1}{i}(1-\frac{1}{i})$ are not comparable to one another and s_n grows like $\sqrt{\log n}$ rather than $\sqrt{n}$, so no rescaling turns R_n into S_m for any m.

Remark 1.18 (Bernoulli arrays on both sides). Both regimes of §1.6 are arrays of independent Bernoullis; only the variances differ. For $B_{n,i} \sim \text{Bernoulli}(\lambda/n)$, $i \le n$, the total variance $\lambda(1-\frac{\lambda}{n})$ stays bounded, Definition 1.8 does not apply at all, and the limit is Poisson(λ). For $B_i \sim \text{Bernoulli}(1/i)$ the total variance diverges, and the limit is Gaussian. The hypothesis $s_n \to \infty$ separates the two.

Remark 1.19 (Random search). Score n hyperparameter configurations drawn independently, and keep the best so far. The number of times the running best improves is precisely R_n, so it is H_n in expectation, with fluctuations of order $\sqrt{\log n}$: $\mathbb{E}R_{100} = 5.19$ and $\mathbb{E}R_{10^4} = 9.79$. A hundredfold increase in n adds $\log 100 = 4.6$ to the expected number of improvements.

Remark 1.20 (A slow theorem). The effective sample size is $s_n^2 \approx \log n$, not n, so the error should be of order $1/s_n$. Exact Kolmogorov–Smirnov distance from $(R_n - H_n)/s_n$ to $\mathcal{N}(0,1)$:

n	10^2	10^3	10^4	10^5
KS	0.131	0.106	0.089	0.080

against $1/s_n = 0.53,\ 0.41,\ 0.35,\ 0.31$. Halving the error requires raising n to the fourth power. As in Remark 1.12, R_n is also lattice-valued.

Exercise 1.21 (TYPE A)**.** Let $B_{n,i} \sim \text{Bernoulli}(p_{n,i})$, $i \le k_n$, be independent and put $\sigma_n^2 = \sum_i p_{n,i}(1-p_{n,i})$. Show that if $\sigma_n \to \infty$ then

$$\frac{1}{\sigma_n}\sum_i \big(B_{n,i} - p_{n,i}\big) \implies \mathcal{N}(0,1).$$

One line, given Lemma 1.15. Where does the array of Remark 1.18 with $p_{n,i} = \lambda/n$ fail the hypothesis, and what is the limit there instead?

Exercise 1.22 (TYPE B)**.** Deduce from Theorem 1.17 that $(R_n - \log n)/\sqrt{\log n} \Rightarrow \mathcal{N}(0,1)$ as well. *First prove the tool: if $Z_n \Rightarrow Z$ with $\sup_n \mathbb{E}Z_n^2 < \infty$, and $a_n \to 1$, $b_n \to 0$ are deterministic, then $a_n Z_n + b_n \Rightarrow Z$. For bounded uniformly continuous f, split at $|Z_n| \le M$ and bound the far piece by $2\|f\|_\infty \mathbb{P}(|Z_n| > M) \le 2\|f\|_\infty M^{-2}\sup_n \mathbb{E}Z_n^2$.* Then use $H_n - \log n \to \gamma$ and $s_n^2/\log n \to 1$. Why is Lemma 1.2 not enough on its own here?

Exercise 1.23 (TYPE B)**.** Build a uniformly random permutation of $[n]$ in cycle form by inserting $1, 2, \ldots, n$ in turn: element i opens a new cycle with probability $1/i$, and otherwise is placed immediately after one of the $i-1$ elements already present, each with probability $1/i$. Show that every permutation arises from exactly one sequence of choices and that each sequence has probability $1/n!$, so the construction is uniform. Conclude that the number of cycles K_n has the law of $\sum_{i\le n}\text{Bernoulli}(1/i)$ with the summands independent, and hence that $(K_n - \log n)/\sqrt{\log n} \Rightarrow \mathcal{N}(0,1)$.

This is Gončarov's theorem, and the bijection is the same counting argument as Lemma 1.16.

1.8 ∗ A wide network at initialization

Not lectured. Fix a nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ and let

$$f_m(x) \;=\; \frac{1}{\sqrt{m}}\sum_{j=1}^m a_j\,\sigma\big(\langle w_j, x\rangle\big), \qquad x \in \mathbb{R}^d,$$

with $w_1, \dots, w_m$ i.i.d. in $\mathbb{R}^d$ and $a_1, \dots, a_m$ i.i.d. with mean 0 and variance σ_a^2, independent of the w_j. This is a one-hidden-layer network of width m at initialization, normalized so that $\operatorname{Var} f_m(x)$ does not move as the width grows.

Proposition 1.24. *Fix finitely many inputs $x_1, \dots, x_k$ and suppose $\mathbb{E}[\sigma(\langle w, x_i\rangle)^2] < \infty$ for each i. Then*

$$\big(f_m(x_1), \dots, f_m(x_k)\big) \implies \mathcal{N}(0, K), \qquad K(x, x') = \sigma_a^2\, \mathbb{E}\big[\sigma(\langle w, x\rangle)\, \sigma(\langle w, x'\rangle)\big].$$

Exercise 1.25 (TYPE B)**.** Prove it. *Put $V_j = \big(a_j\sigma(\langle w_j, x_1\rangle), \dots, a_j\sigma(\langle w_j, x_k\rangle)\big) \in \mathbb{R}^k$ and identify $\big(f_m(x_1), \dots, f_m(x_k)\big)$ as $m^{-1/2}\sum_j V_j$.* Check $\mathbb{E}V_j = 0$ and $\operatorname{Cov}(V_j) = K$, saying where the independence of a_j from w_j is used in each. Then either quote the multidimensional central limit theorem, or apply Theorem 1.7 to the scalars $\langle v, V_j\rangle$ for each fixed $v \in \mathbb{R}^k$, read off convergence of the transform on $\mathbb{R}^k$, and finish with Lemma 1.29. Say which of the two routes needs a third moment.

Exercise 1.26 (TYPE A)**.** Show that K is a valid covariance — that $\sum_{i,l} v_i v_l K(x_i, x_l) \ge 0$ for every $v \in \mathbb{R}^k$ — for any σ with $\mathbb{E}[\sigma(\langle w, x_i\rangle)^2] < \infty$. *One line.*

The limit is Gaussian whatever σ and whatever the law of the weights; those enter only through K.

Exercise 1.27 (TYPE C)**.** Let $\sigma = \max(\cdot, 0)$ and $w \sim \mathcal{N}(0, I_d)$. Show that for $x, x' \neq 0$ with angle $\theta = \angle(x, x')$,

$$\mathbb{E}\big[\sigma(\langle w, x\rangle)\, \sigma(\langle w, x'\rangle)\big] \;=\; \frac{\|x\|\,\|x'\|}{2\pi}\Big(\sin\theta + (\pi - \theta)\cos\theta\Big).$$

Check the answer at $\theta = 0$ against $\mathbb{E}[\max(0, g)^2]$, and at $\theta = \pi/2$ against independence. *This is the arccosine kernel.*

Remark 1.28. Proposition 1.24 concerns finitely many inputs. Holding for every finite subset of $\mathbb{R}^d$ at once is what it means to call the limit a *Gaussian process* with kernel K — Lecture 9. Conditioning it on data is Theorem 2.9 of Lectures 1–2 with Σ replaced by the kernel matrix, so an infinitely wide network fitted by exact Bayesian inference is kernel regression with kernel K. At finite width f_m is not Gaussian, and the size of the discrepancy is a question about the *rate* in Proposition 1.24, which Lecture 4 supplies.

1.9 ∗ The proof we did not give

Not lectured. Characteristic functions turn sums into products, and a product of n nearly-equal factors is an exponential. That is the older proof of Theorem 1.7.

Lemma 1.29 (Continuity theorem)**.** *Let Y_n and Y be random vectors in $\mathbb{R}^k$ with $\varphi_{Y_n}(t) \to \varphi_Y(t)$ for every $t \in \mathbb{R}^k$. Then $Y_n \Rightarrow Y$.*

Proof. Let f be continuous with f and $\widehat{f}$ both in $L^1(\mathbb{R}^k)$; then f is bounded, being the inverse transform of an integrable function. By Fourier inversion and Fubini, legitimate because $\widehat{f} \in L^1$,

$$\mathbb{E}f(Y_n) = \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} \widehat{f}(\xi)\, \mathbb{E}\big[e^{i\langle \xi, Y_n\rangle}\big]\, d\xi = \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} \widehat{f}(\xi)\, \varphi_{Y_n}(-\xi)\, d\xi.$$

Since $|\varphi_{Y_n}| \le 1$ and $\widehat{f} \in L^1$, dominated convergence gives $\mathbb{E}f(Y_n) \to \frac{1}{(2\pi)^k}\int \widehat{f}(\xi)\varphi_Y(-\xi)d\xi = \mathbb{E}f(Y)$.

The functions used in Exercise 1.3 all lie in this class: a smooth cutoff χ_M is compactly supported, so $\widehat{\chi_M}$ is Schwartz; and $g_\delta = g * \phi_\delta$ is integrable with $\widehat{g_\delta} = \widehat{g}\,\widehat{\phi_\delta}$ a bounded function times a Gaussian. That argument therefore applies verbatim and gives $Y_n \Rightarrow Y$. (The imported statements of Lectures 1–2 are all in $\mathbb{R}^k$.) ■

Exercise 1.30 (TYPE A)**.** For $\mathbb{E}|X|^k < \infty$, differentiate under the expectation k times — exhibit the dominating variable — to get $\varphi_X^{(k)}(0) = (-i)^k \mathbb{E}[X^k]$. Check it against $\varphi(t) = e^{-t^2/2}$ by computing $\varphi''(0)$ and $\varphi^{(4)}(0)$, recovering $\mathbb{E}g^2 = 1$ and $\mathbb{E}g^4 = 3$.

Lemma 1.31. *If $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$ then $\varphi_X(t) = 1 - \frac{t^2}{2} + r(t)$ with $r(t)/t^2 \to 0$.*

Proof. For real θ, $|e^{i\theta} - 1 - i\theta + \frac{\theta^2}{2}| \le \min\left(\frac{|\theta|^3}{6}, \theta^2\right)$: the first bound is the integral form of the Taylor remainder, the second follows from $|e^{i\theta} - 1 - i\theta| \le \frac{\theta^2}{2}$. Put $\theta = -tX$ and take expectations, using $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$:

$$|r(t)| \;\le\; \mathbb{E}\min\left(\tfrac{|t|^3|X|^3}{6},\; t^2X^2\right).$$

Divide by t^2. The integrand is bounded by X^2, which is integrable, and tends to 0 pointwise, so dominated convergence gives $r(t)/t^2 \to 0$. The second branch of the minimum is what supplies the dominating variable without a third moment. ■

Exercise 1.32 (TYPE B)**.** Show $(1+z_n/n)^n \to e^z$ whenever $z_n \to z$ in $\mathbb{C}$ *(for $|a|, |b| \le M$, $|a^n - b^n| \le nM^{n-1}|a-b|$; take $b = e^{z_n/n}$).* Combine with Lemma 1.31 to get $\varphi_{S_n}(t) = [\varphi_{X_1}(t/\sqrt{n})]^n \to e^{-t^2/2}$ and conclude $S_n \Rightarrow \mathcal{N}(0,1)$. Note where independence was used, and where identical distribution was used.

Exercise 1.33 (TYPE A)**.** Let $B \sim \mathrm{Bernoulli}(p)$; show $\varphi_B(t) = 1 + p(e^{-it} - 1)$. For independent $B_{n,i} \sim \mathrm{Bernoulli}(\lambda/n)$ with $N_n = \sum_i B_{n,i}$, show $\varphi_{N_n}(t) \to \exp(\lambda(e^{-it} - 1))$ and identify the limit as $\mathrm{Poisson}(\lambda)$ by summing the series. *The same three lines as Exercise 1.32; only the normalization differs.*

Exercise 1.34 (TYPE B)**.** For the centered array $X_{n,i} = B_{n,i} - \lambda/n$, show $s_n^2 = \lambda(1 - \lambda/n)$ and that $L_n(\varepsilon) \to 1$ for every $\varepsilon < 1/\sqrt{\lambda}$, so Definition 1.8 fails. Compute $L_n(\frac{1}{2})$ exactly at $\lambda = 3$ for $n = 10, 100, 10^4$; you should get 0, 0.970, 1.000, and the first should be exactly 0 — say why.

Exercise 1.35 (TYPE C)**.** Take $X_{n,i} = \pm 1$ for $i \le n$ and one extra summand $X_{n,n+1} = \pm\sqrt{n}$, all independent and symmetric. Show $s_n^2 = 2n$, that $L_n(\varepsilon) \ge \frac{1}{2}$ for $\varepsilon < 1/\sqrt{2}$, and that the limiting characteristic function of $s_n^{-1}\sum_i X_{n,i}$ is $e^{-t^2/4}\cos(t/\sqrt{2})$. Exhibit a t at which it vanishes, and conclude that $s_n^{-1}\sum_i X_{n,i}$ converges in distribution to no Gaussian at all. *Feller's condition fails here as well, so Remark 1.10(c) does not apply and the transform must be computed. Why does one zero settle it, without knowing in advance that a limit exists?*

Exercise 1.36 (TYPE C)**.** Show that for $X_i = \pm 1$ no bound $\sup_z |\mathbb{P}(S_n \le z) - \Phi(z)| \le Cn^{-\alpha}$ can hold with $\alpha > \frac{1}{2}$. *For n even, $\mathbb{P}(S_n = 0) = \binom{n}{n/2}2^{-n}$, which Stirling puts at $\sim \sqrt{2/\pi n}$; now compare $\mathbb{P}(S_n \le z)$ with $\Phi(z)$ near $z = 0$.* Explain in one sentence why lattice summands were the slowest row of the table in §1.6.

Exercise 1.37 (TYPE C)**.** Show that for every real random variable Y and every $u > 0$,

$$\mathbb{P}\big(|Y| \ge 2/u\big) \;\le\; \frac{1}{u}\int_{-u}^{u} \big(1 - \varphi_Y(t)\big)\,dt.$$

Deduce that if $\varphi_{Y_n} \to \psi$ pointwise with ψ continuous at 0, then $\sup_n \mathbb{P}(|Y_n| > K) \to 0$ as $K \to \infty$. *This is the half of the continuity theorem that does not need the limit law in advance. The other half — that ψ is then the transform of some random variable — is a compactness argument we do not need, since Lemma 1.29 is always applied with the limit already in hand.*

ASSIGNED **Lecture 3**, due at the start of Lecture 4: Exercises 1.3, 1.13, 1.14, 1.21, 1.22, 1.25, 1.32, 1.35 — two A, three B, three C. Every other exercise in §1 is a pool problem: available and examinable, not collected. *If you are short of time, do the* C *problems.*

2 Stein's method

Theorem 1.7 describes the Gaussian as a limit. It can also be characterized by an identity.

> **Q.** Is there an identity satisfied by the standard Gaussian and by no other law?
>
> If so, then showing that a law is approximately Gaussian becomes showing that it approximately satisfies that identity.

2.1 The characterization

Theorem 2.1 (Stein's lemma). *A random variable W is standard Gaussian if and only if*

$$\mathbb{E}\big[Wf(W)\big] = \mathbb{E}\big[f'(W)\big]$$

for every bounded $f \in C^1(\mathbb{R})$ with bounded derivative.

Proof of the forward direction. Let $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$, so that $\phi'(x) = -x\phi(x)$. Then

$$\mathbb{E}[f'(g)] = \int f'\phi = -\int f\phi' = \int xf(x)\phi(x)\,dx = \mathbb{E}[gf(g)].$$

∎

The converse is proved in §2.2.

Remark 2.2 (The exponential test function). With $f(x) = e^{-itx}$, Stein's identity becomes

$$\mathbb{E}\big[We^{-itW}\big] = -it\,\mathbb{E}\big[e^{-itW}\big],$$

that is $\varphi'(t) = -t\varphi(t)$, the differential equation of Lemma 1.10 of Lectures 1–2.

Splitting off one summand and Taylor-expanding to reach $\varphi'_{S_n} \approx -t\varphi_{S_n}$ proves the central limit theorem; that route is §1.9*.

Exercise 2.3 (TYPE B). *Vector form.* Let $X \sim \mathcal{N}(0,\Sigma)$ on $\mathbb{R}^d$, with $\Sigma \succeq 0$ possibly singular, and let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $\hat{f}$ and $|\xi|\,\hat{f}(\xi)$ are integrable, with $\hat{f}$ and the inversion formula as in Proposition 1.1 of Lectures 1–2.

(a) Show

$$\mathbb{E}\big[X_j f(X)\big] = \sum_k \Sigma_{jk}\,\mathbb{E}\big[\partial_k f(X)\big], \qquad \text{that is} \qquad \mathbb{E}\big[Xf(X)\big] = \Sigma\,\mathbb{E}\big[\nabla f(X)\big].$$

Two facts about the transform. Multiplication by X_j acts on $\xi \mapsto \mathbb{E}[e^{i\langle X,\xi\rangle}]$ as $-i\partial_{\xi_j}$, and by Definition 1.3 of Lectures 1–2 that function is $e^{-\frac{1}{2}\xi^\top \Sigma\xi}$, so

$$\mathbb{E}\big[X_j e^{i\langle X,\xi\rangle}\big] = i\,(\Sigma\xi)_j\, e^{-\frac{1}{2}\xi^\top \Sigma\xi};$$

and $\widehat{\partial_k f}(\xi) = i\xi_k \hat{f}(\xi)$. Write both sides of the claim by Fourier inversion and compare the integrands. Fubini is justified by $\hat{f} \in L^1(\mathbb{R}^d)$ and $\mathbb{E}|X_j| < \infty$, and $\xi_k \hat{f} \in L^1$ is what lets $\partial_k f$ be recovered by inversion. No nondegeneracy of Σ is used anywhere.

(b) Show conversely that if $\mathbb{E}[\langle W, F(W)\rangle] = \mathbb{E}[\mathrm{tr}(\Sigma\, DF(W))]$ for every bounded $F \in C^1(\mathbb{R}^d;\mathbb{R}^d)$ with bounded derivative, then $W \sim \mathcal{N}(0,\Sigma)$. *Take $F(x) = v\, f(\langle v, x\rangle)$, reduce to Theorem 2.1 for $\langle v, W\rangle$, and finish with Corollary 1.7 of Lectures 1–2. Treat $v^\top \Sigma v = 0$ separately.*

(c) *Matrix form.* Let M be a random matrix whose entries are jointly Gaussian with mean 0, and let F satisfy the hypothesis above on $\mathbb{R}^{n\times m}$. Show

$$\mathbb{E}\big[M_{ij}F(M)\big] = \sum_{k,l} \mathbb{E}\big[M_{ij}M_{kl}\big]\,\mathbb{E}\Big[\frac{\partial F}{\partial M_{kl}}(M)\Big].$$

One line from (a): a matrix is a vector whose index is a pair, with covariance $\Sigma_{(ij),(kl)} = \mathbb{E}[M_{ij}M_{kl}]$, and nothing in (a) used the shape of the index set.

2.2 Stein's equation

Theorem 2.1 says that $W \sim \mathcal{N}(0,1)$ exactly when

$$\mathbb{E}\big[f'(W) - Wf(W)\big] = 0 \qquad \text{for every admissible } f.$$

What is wanted is a quantitative form of this: a bound on $\mathbb{E}h(W) - \mathbb{E}h(g)$ for a given test function h.

Suppose that for this h we can produce a function f with

$$f'(w) - wf(w) = h(w) - \mathbb{E}h(g) \qquad \text{for every } w \in \mathbb{R}.$$

Both sides are functions of a real variable, and g enters only through the constant $\mathbb{E}h(g)$; so the identity may be evaluated at $w = W$ and averaged, giving

$$\mathbb{E}h(W) - \mathbb{E}h(g) = \mathbb{E}\big[f'(W) - Wf(W)\big].$$

Producing f from h is therefore the whole problem. Fix h and look for f solving

$$f'(w) - wf(w) \;=\; h(w) - \mathbb{E}h(g). \tag{3}$$

Throughout §2.2 and §2.3, $h \in C^3(\mathbb{R})$ with h', h'' and h''' bounded; this is the class of Lemma 1.2. For $s \geq 0$ set

$$(T_s h)(w) = \mathbb{E}\Big[h\big(e^{-s}w + \sqrt{1-e^{-2s}}\, g\big)\Big].$$

Independent Gaussians add variances, and $e^{-2t}(1-e^{-2s}) + 1 - e^{-2t} = 1 - e^{-2(s+t)}$, so $T_sT_t = T_{s+t}$: this is the *Ornstein–Uhlenbeck semigroup*. It interpolates between $T_0h = h$ and $T_sh \to \mathbb{E}h(g)$, and its generator is

$$Lu = \lim_{s\downarrow 0} \frac{T_s u - u}{s} = u'' - wu',$$

computed in the proof below, and $\partial_s T_s = LT_s$.

Two observations now produce f_h. Write $\tilde{h} = h - \mathbb{E}h(g)$, so that $T_s\tilde{h} = T_s h - \mathbb{E}h(g)$ and $T_s\tilde{h} \to 0$. First, telescoping in s,

$$\tilde{h} \;=\; T_0\tilde{h} - \lim_{s\to\infty} T_s\tilde{h} \;=\; -\int_0^\infty \partial_s T_s\tilde{h}\,ds \;=\; -L\int_0^\infty T_s\tilde{h}\,ds.$$

Second, Stein's operator $\mathcal{A}f = f' - wf$ is L with one derivative factored out:

$$\mathcal{A}(u') = (u')' - wu' = Lu.$$

So $\mathcal{A}f = \tilde{h}$ is solved by $f = u'$ with $u = -\int_0^\infty T_s\tilde{h}\,ds$, that is by

$$f_h = -\int_0^\infty \partial_w T_s h\,ds,$$

the constant $\mathbb{E}h(g)$ disappearing under ∂_w. Each w-derivative of T_s produces a factor e^{-s}, which is what makes this integral converge and what gives the bounds below.

Lemma 2.4 (The solution, and its derivatives)**.** *Equation* (3) *has exactly one bounded solution, namely*

$$f_h \;=\; -\int_0^\infty \partial_w T_s h\,ds, \qquad \textit{that is} \qquad f_h(w) = -\int_0^\infty e^{-s}\,\mathbb{E}\Big[h'\big(e^{-s}w + \sqrt{1-e^{-2s}}\,g\big)\Big]\,ds,$$

and it satisfies

$$\|f_h\|_\infty \le \|h'\|_\infty, \qquad \|f_h'\|_\infty \le \tfrac{1}{2}\|h''\|_\infty, \qquad \|f_h''\|_\infty \le \tfrac{1}{3}\|h'''\|_\infty.$$

In particular $f_h \in C^2(\mathbb{R})$.

Proof. Put $b_s = \sqrt{1-e^{-2s}}$ and $u(s,w) = (T_s h)(w) = \mathbb{E}[h(e^{-s}w + b_s g)]$, so that $u(0,w) = h(w)$ and $u(s,w) \to \mathbb{E}h(g)$ as $s \to \infty$. Differentiating in s, and using $\dot{b}_s = e^{-2s}/b_s$,

$$\partial_s u = \mathbb{E}\Big[h'(e^{-s}w + b_s g)\Big(-e^{-s}w + \frac{e^{-2s}}{b_s}\,g\Big)\Big] = -e^{-s}w\,\mathbb{E}\big[h'\big] + \frac{e^{-2s}}{b_s}\,\mathbb{E}\big[g\,h'\big].$$

Theorem 2.1, applied to $g \mapsto h'(a + b_s g)$, gives $\mathbb{E}[g\,h'(a+b_s g)] = b_s\,\mathbb{E}[h''(a+b_s g)]$, which also removes the apparent singularity at $s = 0$. Since $\partial_w u = e^{-s}\mathbb{E}[h']$ and $\partial_w^2 u = e^{-2s}\mathbb{E}[h'']$, this says exactly

$$\partial_s u \;=\; \partial_w^2 u - w\,\partial_w u.$$

Now $f_h(w) = -\int_0^\infty \partial_w u(s,w)\,ds$ is the displayed formula, and

$$f_h'(w) - wf_h(w) = -\int_0^\infty \big(\partial_w^2 u - w\,\partial_w u\big)\,ds = -\int_0^\infty \partial_s u\,ds = u(0,w) - \lim_{s\to\infty} u(s,w) = h(w) - \mathbb{E}h(g),$$

so f_h solves (3). Two solutions differ by a solution of $f' = wf$, that is by $ce^{w^2/2}$, which is bounded only for $c = 0$; and f_h is bounded, by $\|h'\|_\infty \int_0^\infty e^{-s}ds$. Differentiating under the integral — legitimate since the kth integrand is dominated by $e^{-ks}\|h^{(k)}\|_\infty$ — gives

$$f_h'(w) = -\int_0^\infty e^{-2s}\,\mathbb{E}\big[h''(e^{-s}w + b_s g)\big]ds, \qquad f_h''(w) = -\int_0^\infty e^{-3s}\,\mathbb{E}\big[h'''(e^{-s}w + b_s g)\big]ds,$$

and $\int_0^\infty e^{-ks}ds = 1/k$. ∎

Remark 2.5. Solving (3) as a linear first-order equation instead gives the classical closed form $f_h(w) = e^{w^2/2} \int_{-\infty}^{w} \big[h(x) - \mathbb{E}h(g)\big] e^{-x^2/2} dx$. It is the same function, by uniqueness.

With $f = f_h$, the identity above reads

$$\mathbb{E}h(W) - \mathbb{E}h(g) \;=\; \mathbb{E}\big[f_h'(W) - W f_h(W)\big]. \tag{4}$$

This settles the converse in Theorem 2.1. Restricting the test functions there to C^1 costs nothing: any solution of (3) satisfies $f' = wf + h - \mathbb{E}h(g)$, whose right side is continuous, so every solution is automatically C^1. By Lemma 2.4, f_h is such a function, bounded and with bounded derivative, hence admissible. So if W satisfies the identity the right side of (4) vanishes; thus $\mathbb{E}h(W) = \mathbb{E}h(g)$ for every h in our class, and $W \overset{d}{=} g$ by Lemma 1.2 and Exercise 1.3.

2.3 The central limit theorem, with a rate

Theorem 2.6. *Let $X_1, \dots, X_n$ be i.i.d. with mean 0, variance 1 and $\mathbb{E}|X_1|^3 < \infty$. Then for every $h \in C^3(\mathbb{R})$ with bounded derivatives,*

$$\big|\mathbb{E}h(S_n) - \mathbb{E}h(g)\big| \;\le\; \frac{\|h'''\|_\infty\, \mathbb{E}|X_1|^3}{2\sqrt{n}}.$$

Proof. Write $W = S_n$, let $W^{(i)} = W - X_i/\sqrt{n}$, which is independent of X_i, and abbreviate $f = f_h$. Then

$$\mathbb{E}\big[W f(W)\big] = \frac{1}{\sqrt{n}} \sum_i \mathbb{E}\big[X_i f(W)\big],$$

and Taylor expanding $f(W) = f(W^{(i)}) + \frac{X_i}{\sqrt{n}} f'(W^{(i)}) + R_i$ with $|R_i| \le \frac{\|f''\|_\infty}{2} \frac{X_i^2}{n}$ gives, for each i,

$$\mathbb{E}\big[X_i f(W)\big] = \underbrace{\mathbb{E}[X_i]\, \mathbb{E}\big[f(W^{(i)})\big]}_{=0} + \frac{1}{\sqrt{n}} \underbrace{\mathbb{E}[X_i^2]}_{=1}\, \mathbb{E}\big[f'(W^{(i)})\big] + \mathbb{E}\big[X_i R_i\big],$$

where $\big|\mathbb{E}[X_i R_i]\big| \le \|f''\|_\infty \mathbb{E}|X_1|^3/2n$ and the first two expectations factor because X_i is independent of $W^{(i)}$. Summing,

$$\Big|\mathbb{E}\big[W f(W)\big] - \frac{1}{n} \sum_i \mathbb{E}\big[f'(W^{(i)})\big]\Big| \;\le\; \frac{\|f''\|_\infty \mathbb{E}|X_1|^3}{2\sqrt{n}}.$$

Each $\mathbb{E}[f'(W^{(i)})]$ differs from $\mathbb{E}[f'(W)]$ by at most $\|f''\|_\infty \mathbb{E}|X_i|/\sqrt{n}$, so

$$\big|\mathbb{E}\big[f'(W) - W f(W)\big]\big| \;\le\; \frac{\|f''\|_\infty}{\sqrt{n}} \Big(\mathbb{E}|X_1| + \tfrac{1}{2}\mathbb{E}|X_1|^3\Big) \;\le\; \frac{3\|f''\|_\infty\, \mathbb{E}|X_1|^3}{2\sqrt{n}},$$

since $\mathbb{E}|X_1| \le (\mathbb{E}X_1^2)^{1/2} = 1 \le (\mathbb{E}X_1^2)^{3/2} \le \mathbb{E}|X_1|^3$. Now insert $\|f''\|_\infty \le \frac{1}{3}\|h'''\|_\infty$ from Lemma 2.4 into (4). ∎

Remark 2.7. On the same class Theorem 1.7 gives $\frac{\|h'''\|_\infty}{6\sqrt{n}} (\mathbb{E}|X_1|^3 + 2\sqrt{2/\pi})$, so the two bounds agree up to a constant.

The same argument with a different characterizing identity gives a Poisson limit theorem with a rate in total variation (§2.4), and it does not require independence (§2.5*).

2.4 The same argument for Poisson

Nothing above was specific to the Gaussian; only the characterizing identity was used. Write $\pi_k = e^{-\lambda}\lambda^k/k!$ for the Poisson(λ) weights and $\pi(A) = \sum_{k\in A}\pi_k$, and recall that for laws on $\{0,1,2,\dots\}$

$$d_{\mathrm{TV}}(X,Y) = \sup_{A\subseteq\{0,1,2,\dots\}} \big|\mathbb{P}(X\in A) - \mathbb{P}(Y\in A)\big|.$$

Theorem 2.8 (Chen's lemma)**.** *A random variable N taking values in $\{0,1,2,\dots\}$ is* Poisson(λ) *if and only if*

$$\mathbb{E}\big[\lambda f(N+1) - Nf(N)\big] = 0$$

for every bounded $f : \{0,1,2,\dots\} \to \mathbb{R}$.

Proof of the forward direction. $\mathbb{E}[Nf(N)] = \sum_{k\geq 1} k\, e^{-\lambda}\dfrac{\lambda^k}{k!} f(k) = \lambda \sum_{j\geq 0} e^{-\lambda}\dfrac{\lambda^j}{j!} f(j+1) = \lambda\,\mathbb{E}[f(N+1)]$. ■

The converse, and the quantitative statement, come from the analogue of (3). Fix A and look for f solving

$$\lambda f(k+1) - kf(k) \;=\; \mathbf{1}_A(k) - \pi(A), \qquad k\geq 0. \tag{5}$$

Here the difference operator replaces the derivative and there is no semigroup to integrate: (5) is a first-order recursion, and it can be solved outright.

Lemma 2.9. *Set $f_A(0) = 0$. Then* (5) *has the unique solution*

$$f_A(k) = \frac{1}{k\pi_k}\sum_{j<k}\pi_j\big(\mathbf{1}_A(j) - \pi(A)\big) = -\frac{1}{k\pi_k}\sum_{j\geq k}\pi_j\big(\mathbf{1}_A(j) - \pi(A)\big), \qquad k\geq 1,$$

it is bounded, and

$$\big|(k-\lambda)f_A(k)\big| \leq 1 \;\; \textit{for every } k, \qquad \|\Delta f_A\|_\infty := \sup_k \big|f_A(k+1) - f_A(k)\big| \;\leq\; \frac{2}{\lambda}.$$

Proof. Write $g = \mathbf{1}_A - \pi(A)$, so $\sum_j \pi_j g(j) = 0$. Multiplying (5) by π_k and using $\lambda\pi_k = (k+1)\pi_{k+1}$ turns it into

$$(k+1)\pi_{k+1}f(k+1) - k\pi_k f(k) = \pi_k g(k),$$

a telescoping recursion for $k\mapsto k\pi_k f(k)$, which starts at 0 when $k=0$. Summing gives the first formula, and the second follows from $\sum_j \pi_j g(j) = 0$. Since $|g|\leq 1$,

$$|f_A(k)| \;\leq\; \frac{1}{k\pi_k}\min\Big(\mathbb{P}(N\leq k-1),\ \mathbb{P}(N\geq k)\Big), \qquad N\sim \text{Poisson}(\lambda). \tag{6}$$

For $k>\lambda$ the ratios $\pi_{k+m}/\pi_k = \prod_{r=1}^m \lambda/(k+r)$ are at most $(\lambda/k)^m$, so $\mathbb{P}(N\geq k)\leq \pi_k\, k/(k-\lambda)$ and (6) gives $|f_A(k)|\leq 1/(k-\lambda)$. For $1\leq k\leq\lambda$ the ratios π_{k-1-m}/π_{k-1} are at most $((k-1)/\lambda)^m$, so $\mathbb{P}(N\leq k-1)\leq \pi_{k-1}\lambda/(\lambda-k+1)$; since $\pi_{k-1} = k\pi_k/\lambda$, (6) gives $|f_A(k)|\leq 1/(\lambda-k+1)$. In both cases $|(k-\lambda)f_A(k)|\leq 1$, and the case $k=0$ is trivial. The first bound also shows $f_A(k)\to 0$, so f_A is bounded.

Finally, subtracting $\lambda f(k)$ from both sides of (5) gives

$$\Delta f_A(k) = \frac{(k-\lambda)f_A(k) + g(k)}{\lambda},$$

and both terms in the numerator are at most 1 in absolute value. ■

Evaluating (5) at $k = W$ and taking expectations is the analogue of (4):

$$\mathbb{P}(W \in A) - \pi(A) = \mathbb{E}\big[\lambda f_A(W+1) - W f_A(W)\big]. \tag{7}$$

This settles the converse in Theorem 2.8: if W satisfies the identity then the right side vanishes for every A, since each f_A is bounded, so $W \sim \text{Poisson}(\lambda)$.

Theorem 2.10 (Le Cam)**.** *Let $B_1, \dots, B_n$ be independent with $B_i \sim \text{Bernoulli}(p_i)$, and let $\lambda = \sum_i p_i$. Then*

$$d_{\text{TV}}\Big(\sum_i B_i,\ \text{Poisson}(\lambda)\Big) \ \le\ \frac{2}{\lambda}\sum_i p_i^2 \ \le\ 2\max_i p_i.$$

Proof. Put $W = \sum_i B_i$ and $W_i = W - B_i$, which is independent of B_i. Fix A and write $f = f_A$. Since $B_i f(W) = B_i f(W_i + 1)$,

$$\mathbb{E}\big[W f(W)\big] = \sum_i \mathbb{E}\big[B_i f(W_i+1)\big] = \sum_i p_i\,\mathbb{E}\big[f(W_i+1)\big],$$

while $\lambda = \sum_i p_i$ gives $\mathbb{E}[\lambda f(W+1)] = \sum_i p_i\,\mathbb{E}[f(W_i + B_i + 1)]$. Subtracting,

$$\mathbb{E}\big[\lambda f(W+1) - W f(W)\big] = \sum_i p_i\,\mathbb{E}\big[f(W_i+B_i+1) - f(W_i+1)\big].$$

The bracket vanishes when $B_i = 0$ and equals $\Delta f(W_i+1)$ when $B_i = 1$; as B_i is independent of W_i, the ith term is $p_i^2\,\mathbb{E}[\Delta f(W_i+1)]$. Hence by (7) and Lemma 2.9,

$$\big|\mathbb{P}(W \in A) - \pi(A)\big| \ \le\ \|\Delta f_A\|_\infty \sum_i p_i^2 \ \le\ \frac{2}{\lambda}\sum_i p_i^2,$$

uniformly in A. The second inequality of the theorem is $\sum_i p_i^2 \le (\max_i p_i)\sum_i p_i$. ■

Corollary 2.11 (Poisson limit theorem)**.** *Let $B_{n,i} \sim \text{Bernoulli}(p_{n,i})$, $i \le k_n$, be independent, with $\max_i p_{n,i} \to 0$ and $\sum_i p_{n,i} \to \lambda > 0$. Then $\sum_i B_{n,i} \Rightarrow \text{Poisson}(\lambda)$.*

Proof. The bound $2\max_i p_{n,i}$ tends to 0, and $\text{Poisson}(\lambda_n) \Rightarrow \text{Poisson}(\lambda)$ when $\lambda_n \to \lambda$. ■

This is the Poisson side of the boundary of §1.6, and unlike Theorem 2.6 it is a bound in total variation. At $p_i = \lambda/n$ the bound is $2\lambda/n$; with $\lambda = 3$ and $n = 100$ it reads 0.06, against a true total variation distance of 0.0076. The constant 2 in Lemma 2.9 can be sharpened to $1 - e^{-\lambda}$, which is attained.

The same technique covers both sides of the boundary of §1.6; only the characterizing identity changes.

2.5 $*$ Where independence was spent

Not lectured. The proof of Theorem 2.6 used independence in exactly one place, to evaluate $\mathbb{E}[X_i f(W^{(i)})] = \mathbb{E}[X_i]\mathbb{E}[f(W^{(i)})]$. Any other route to that expectation will do, and the cheapest is a coupling.

Definition 2.12 (Exchangeable pair)**.** A pair (W, W') is exchangeable if $(W, W') \overset{d}{=} (W', W)$. It is a *Stein pair* if in addition $\mathbb{E}[W' - W \mid W] = -aW$ for some $a > 0$.

Exercise 2.13 (TYPE A)**.** Let (W, W') be exchangeable and F antisymmetric, $F(x,y) = -F(y,x)$, with $\mathbb{E}|F(W,W')| < \infty$. Show $\mathbb{E}F(W,W') = 0$. *One line; everything else is the choice of F.*

Exercise 2.14 (TYPE B)**.** Apply that to $F(x,y) = (y-x)(f(x)+f(y))$, expand $f(W')$ about W to first order, and use Definition 2.12 to obtain

$$\mathbb{E}\big[f'(W) - Wf(W)\big] = \mathbb{E}\Big[f'(W)\Big(1 - \tfrac{1}{2a}\mathbb{E}\big[(W'-W)^2 \mid W\big]\Big)\Big] + \mathrm{err}, \quad |\mathrm{err}| \le \frac{\|f''\|_\infty}{4a}\mathbb{E}|W'-W|^3.$$

A Gaussian approximation now asks two things: that $\frac{1}{2a}\mathbb{E}[(W'-W)^2 \mid W]$ be close to 1, and that $W'-W$ be small.

Exercise 2.15 (TYPE B)**.** *The canonical pair.* With $W = S_n$, let I be uniform on $\{1,\dots,n\}$ and X_I' an independent copy of X_I; put $W' = W - (X_I - X_I')/\sqrt{n}$. Show (W,W') is exchangeable with $\mathbb{E}[W'-W \mid X_1,\dots,X_n] = -W/n$, so $a = 1/n$; then that $\mathbb{E}[(W'-W)^2 \mid X_1,\dots,X_n] = n^{-2}\sum_i(X_i^2+1)$, so that $\frac{1}{2a}\mathbb{E}[(W'-W)^2 \mid X_1,\dots,X_n] = \frac{1}{2}\big(1+\frac{1}{n}\sum_i X_i^2\big)$, which sits at L^2 distance $\frac{1}{2}\sqrt{\mathrm{Var}(X_1^2)/n}$ from 1. Assuming also $\mathbb{E}X_1^4 < \infty$, conclude a bound of the order of Theorem 2.6. *Independence enters twice — through the resampled copy, and in the variance of $\frac{1}{n}\sum_i X_i^2$ — and neither use is the product formula of §2.3.*

Now a sum to which no version of independence applies. Let π be a uniformly random permutation of $\{1,\dots,n\}$, let $I_i = \mathbf{1}\{\pi(i) = i\}$, and let $W = \sum_i I_i$ count fixed points.

Exercise 2.16 (TYPE A)**.** Show $\mathbb{E}W = \mathrm{Var}\, W = 1$ for $n \ge 2$, and that $\mathbb{P}(I_1 = I_2 = 1) \ne \mathbb{P}(I_1 = 1)\mathbb{P}(I_2 = 1)$. A student applies Theorem 2.10 with $p_i = 1/n$ and $\lambda = 1$ to conclude $d_{\mathrm{TV}}(W, \mathrm{Poisson}(1)) \le 2/n$. Name the hypothesis that fails, and say why the conclusion happening to be true does not rescue the argument.

> **Imported: Chen–Stein, coupling form.** Let $W = \sum_{i\le n} I_i$ with $I_i \sim \mathrm{Bernoulli}(p_i)$, *not necessarily independent*, and $\lambda = \sum_i p_i$. If for each i one can build, on the same space as W, a variable V_i distributed as $W-1$ conditioned on $I_i = 1$, then $d_{\mathrm{TV}}(W, \mathrm{Poisson}(\lambda)) \le \min(1, \lambda^{-1})\sum_i p_i\mathbb{E}|W - V_i|$.

Exercise 2.17 (TYPE C)**.** Build the coupling. Fix i; if $\pi(i) = i$ put $\sigma = \pi$, otherwise $\sigma = \pi \circ (i\ \ \pi^{-1}(i))$. Show $\sigma(i) = i$ always, that σ is uniform among permutations fixing i, and that $W(\sigma) - W(\pi) \in \{1,2\}$ when $\pi(i) \ne i$ — the value 2 exactly when π has a 2-cycle at i. Deduce $\mathbb{E}|W - V_i| \le 2/n$, hence $d_{\mathrm{TV}}(W, \mathrm{Poisson}(1)) \le 2/n$.

Inclusion–exclusion gives $\mathbb{P}(W = 0) = \sum_{k\le n}(-1)^k/k!$, whose distance from e^{-1} is 7.1×10^{-3}, 1.8×10^{-4} and 2.3×10^{-8} at $n = 4, 6, 10$. *By how many orders of magnitude is the method wrong at $n = 10$, and what does it have that inclusion–exclusion does not?*

---

§1.8* carries out the wide-network limit and computes the arccosine kernel; §1.9* gives the characteristic-function proof of the central limit theorem, the same computation for Poisson, and Lindeberg's condition on both sides of the boundary; §2.5* develops the exchangeable-pair bound and a case with no independence at all.

ASSIGNED **Lecture 4**, due at the start of Lecture 5: Exercises 2.13–2.17 — two A, two B, one C. Exercise 2.3 is a pool problem. *If you are short of time, do the C problem.*