Lectures 3 and 4: The Central Limit Theorem

Swapping, the Lindeberg condition, and Stein's method

1 Why Gaussians appear, and the CLT by swapping

Lectures 1 and 2 said what a Gaussian is and how to condition one. Neither said why one keeps meeting them.

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

1.1 Notation

Throughout, $X_1, X_2, \dots$ 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 Homework 1.** 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 The moral answer

Homework 1 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 $(x,y) \mapsto \big(\frac{x+y}{\sqrt{2}}, \frac{x-y}{\sqrt{2}}\big)$ 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.

1.3 The entropic CLT

The heuristic is a theorem.

Theorem 1.1 (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 1.2 (Artstein–Ball–Barthe–Naor, 2004)**.** *$h(S_n)$ is increasing in n, not merely along powers of two.*

We will not prove either.

Example 1.3 (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. We need a proof that covers Example 1.3.

1.4 Lindeberg's swapping argument

Exchange the summands for Gaussians one at a time, and check that each exchange is cheap.

Theorem 1.4 (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 Taylor's theorem and nothing else. In particular it never objected to $X_i = \pm 1$.

Imported. Convergence of $\mathbb{E}f(S_n)$ for every $f \in C^3$ with bounded derivatives implies it for every bounded continuous f, by a smoothing argument. We define $X_n \Rightarrow X$ to mean $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous f.

The classical argument — expand $\varphi_{X_1}(t/\sqrt{n})$ and take nth powers — has been available since Lecture 1. Homework 3 carries it out and locates what it costs.

1.5 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 verbatim. What it did use is that no single summand carries much of the variance.

Definition 1.5 (Lindeberg's condition)**.** Let $\{X_{n,i}\}_{i\le k_n}$ be a triangular array, independent within rows, with $\mathbb{E}X_{n,i}=0$ and $s_n^2=\sum_i \mathrm{Var}(X_{n,i})$. The array satisfies *Lindeberg's condition* if for every $\varepsilon>0$

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

Under this condition $s_n^{-1}\sum_i X_{n,i} \Rightarrow \mathcal{N}(0,1)$, by the same interpolation with the third-moment bound replaced by a truncation.

This is not a technical hypothesis inserted to make a proof work. It separates two regimes: many small contributions give a Gaussian, and a few rare ones give a Poisson. Homework 3 computes both sides of the boundary; §2.4 proves the Poisson statement, with a rate.

Remark 1.6 (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.

1.6 A wide network at initialization

The multidimensional statement is the one that pays.

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 $\mathrm{Var}\,f_m(x)$ does not move as the width grows.

Proposition 1.7. *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*

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

Proof. Put $V_j=\big(a_j\sigma(\langle w_j,x_1\rangle),\dots,a_j\sigma(\langle w_j,x_k\rangle)\big)\in\mathbb{R}^k$. The V_j are i.i.d.; $\mathbb{E}V_j=0$ because a_j has mean 0 and is independent of w_j; and $\mathrm{Cov}(V_j)=K$ by the same independence. The vector in question is $m^{-1/2}\sum_j V_j$, so this is the multidimensional central limit theorem. Alternatively, apply Theorem 1.4 to $\langle v,V_j\rangle$ for each fixed $v\in\mathbb{R}^k$ and appeal to Corollary 1.5 of Lecture 1. ∎

The limit is Gaussian whatever σ and whatever the law of the weights: those enter only through K, and the rest of the architecture is erased. And K is computable — for $\sigma=\max(\cdot,0)$ and $w\sim\mathcal{N}(0,I_d)$,

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

Remark 1.8. Proposition 1.7 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 2.3 of Lecture 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 1.7, which Lecture 4 supplies.

2 Stein's method

Theorem 1.4 describes the Gaussian as a limit. There is another way to pin it down.

> **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 absolutely continuous f 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 (We have used this already). Take $f(x) = e^{itx}$. Stein's identity becomes $\mathbb{E}[We^{itW}] = it\,\mathbb{E}[e^{itW}]$, that is $\varphi'(t) = -t\varphi(t)$, which is the differential equation solved in Lemma 1.2. The first computation of the course was Stein's lemma with an exponential test function.

One can run a proof of the central limit theorem this way, splitting off one summand and Taylor-expanding to reach $\varphi'_{S_n} \approx -t\varphi_{S_n}$; that route is Homework 3. Exponential test functions give neither a rate in a probability metric nor any tolerance for dependence, which is what the rest of this lecture supplies.

2.2 Stein's equation

Fix a test function h and look for f solving

$$f'(w) - wf(w) \;=\; h(w) - \mathbb{E}h(g). \tag{1}$$

Lemma 2.3. *Equation* (1) *is solved by* $f_h(w) = e^{w^2/2}\int_{-\infty}^{w}\big[h(x) - \mathbb{E}h(g)\big]e^{-x^2/2}\,dx$.

Proof. Differentiate: $f_h'(w) = wf_h(w) + \big[h(w) - \mathbb{E}h(g)\big]$. ■

Stated without proof. If h is bounded then f_h is bounded and absolutely continuous with $\|f_h\|_\infty \le \sqrt{2\pi}\,\|h - \mathbb{E}h(g)\|_\infty$ and $\|f_h'\|_\infty \le 4\|h - \mathbb{E}h(g)\|_\infty$; if h is moreover Lipschitz then $\|f_h''\|_\infty \le 2\|h'\|_\infty$. These bounds are the technical core of the method.

Evaluating (1) at W and taking expectations turns a distance between two laws into a single expectation:

$$\mathbb{E}h(W) - \mathbb{E}h(g) \;=\; \mathbb{E}\big[f_h'(W) - W f_h(W)\big]. \tag{2}$$

This also settles the converse in Theorem 2.1. If W satisfies the identity, then for every bounded Lipschitz h the right side of (2) vanishes, since f_h is an admissible test function; so $\mathbb{E}h(W) = \mathbb{E}h(g)$ for all such h, and hence $W \overset{d}{=} g$.

2.3 The central limit theorem, with a rate

Theorem 2.4. *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 bounded Lipschitz h,*

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

for an absolute constant C.

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

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

and expanding $f(W)$ about $W^{(i)}$ gives, for each i,

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

with the remainder bounded by $\|f''\|_\infty \mathbb{E}|X_i|^3/n$. Summing over i turns the middle terms into $\mathbb{E}f'(W)$ up to an error of the same order, so

$$\big|\mathbb{E}\big[f'(W) - Wf(W)\big]\big| \;\le\; \frac{C'\|f''\|_\infty\,\mathbb{E}|X_1|^3}{\sqrt{n}},$$

and (2) with the bound $\|f_h''\|_\infty \le 2\|h'\|_\infty$ finishes the proof. ■

Remark 2.5. Theorem 1.4 produced the rate $n^{-1/2}$ as a by-product of its error term; here the rate is what the method computes. Taking h to approximate $\mathbf{1}_{(-\infty,z]}$, and controlling the approximation, upgrades Theorem 2.4 to the Berry–Esseen theorem: $\sup_z |\mathbb{P}(S_n \le z) - \mathbb{P}(g \le z)| \le C\,\mathbb{E}|X_1|^3/\sqrt{n}$. The extra work is in the indicator, not in the method.

2.4 The same argument for Poisson

Nothing above was about Gaussians in particular. It was about having a characterizing identity.

Theorem 2.6 (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.

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

Running §2.2 and §2.3 with this identity in place of Stein's, and total variation in place of a smooth metric, gives:

Theorem 2.7 (Le Cam)**.** *Let $B_1, \dots, B_n$ be independent with $B_i \sim \mathrm{Bernoulli}(p_i)$, and let $\lambda = \sum_i p_i$. Then*

$$d_{\mathrm{TV}}\Big(\sum_i B_i,\ \mathrm{Poisson}(\lambda)\Big) \ \leq\ \sum_i p_i^2.$$

Corollary 2.8 (Poisson limit theorem)**.** *Let $B_{n,i} \sim \mathrm{Bernoulli}(p_{n,i})$, $i \leq k_n$, be independent, with $\max_i p_{n,i} \to 0$ and $\sum_i p_{n,i} \to \lambda$. Then $\sum_i B_{n,i} \Rightarrow \mathrm{Poisson}(\lambda)$.*

Proof. $\sum_i p_{n,i}^2 \leq (\max_i p_{n,i}) \sum_i p_{n,i} \to 0$, and $\mathrm{Poisson}(\lambda_n) \Rightarrow \mathrm{Poisson}(\lambda)$ when $\lambda_n \to \lambda$. ∎

This is the Poisson side of the Lindeberg boundary. At $p_i = \lambda/n$ the rate is λ^2/n; with $\lambda = 3$ and $n = 100$ the bound is 0.09, against an actual discrepancy of about 0.002 in $\mathbb{P}(\sum_i B_i = 0)$.

Lecture 3 met the Gaussian/Poisson boundary as a phenomenon. Here a single technique sits on both sides of it, and only the characterizing identity changes.

2.5 Dependence

The proof of Theorem 2.4 used independence only to evaluate $\mathbb{E}[X_i f(W^{(i)})]$. Any other route to that expectation will do.

Definition 2.9 (Exchangeable pair)**.** A pair (W, W') is exchangeable if $(W, W') \stackrel{d}{=} (W', W)$. It is a *Stein pair* if in addition $\mathbb{E}[W' - W \mid W] = -aW$ for some $a > 0$.

For a Stein pair one bounds $\mathbb{E}[f'(W) - Wf(W)]$ in terms of the conditional variance of $W' - W$, with no independence assumed. This, and the dependency-graph variants, is why Stein's method rather than swapping is the tool when the summands interact.

Homework 3 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; Homework 4 develops Stein's equation and the exchangeable-pair bound.