Stein's method for Machine Learning

Session 1 : An introduction

Amirabbas Asadi

amir.asadi78@sharif.edu

October 2026

Let's start with the central limit theorem

Let $X_1, X_2, \ldots$ be i.i.d., with

$$\mathbb{E}X_i = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2 \in (0, \infty).$$

$$W_n := \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

But how close are we?

The CLT tells us what happens as $n \to \infty$.

In an actual problem, n is finite.

How different are W_n and Z?

And what if the contributions are not independent?

What kind of bound are we looking for?

There are different ways to describe a random variable.

- **Tail bounds:** how far can it wander?
  Markov, Chebyshev, and concentration inequalities control probabilities such as $\mathbb{P}(|W_n| > t)$.

- **Distributional approximation:** how close is it to a Gaussian?
  We want to compare probabilities under W_n and Z.

A Berry–Esseen-type bound asks for

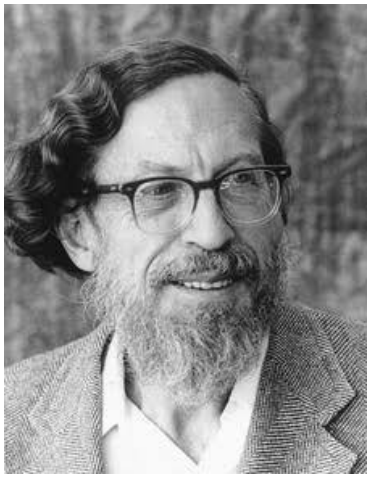
$$\sup_{a \in \mathbb{R}} |\mathbb{P}(W_n \leq a) - \Phi(a)| \leq \text{an explicit error depending on } n.$$

A different proof

While teaching probability, Charles Stein was looking for a different proof of the central limit theorem.

A proof that did not rely on Fourier transforms.

The starting point? A simple Gaussian identity.



Charles Stein

1920–2016

Stein's lemma

$$Z \sim \mathcal{N}(0,1)$$

$$\mathbb{E}\left[f'(Z) - Zf(Z)\right] = 0$$

for sufficiently regular test functions f.

First test: $f(z) = 1$

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

Then $f'(z) = 0$.

$$\mathbb{E}[0 - Z] = 0 \qquad \Longrightarrow \qquad \mathbb{E}Z = 0.$$

The identity knows the Gaussian's mean.

Not very surprising. Let's keep going.

Second test: $f(z) = z$

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

Now $f'(z) = 1$.

$$\mathbb{E}[1 - Z^2] = 0 \qquad \Longrightarrow \qquad \mathbb{E}Z^2 = 1.$$

So it also knows the variance.

But mean zero and variance one do not make a distribution Gaussian.

Third test: $f(z) = z^2$

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

$$\mathbb{E}[2Z - Z^3] = 0.$$

Since $\mathbb{E}Z = 0$,

$$\mathbb{E}Z^3 = 0.$$

The third moment vanishes.

That is consistent with symmetry, but does not prove symmetry.

Fourth test: $f(z) = z^3$

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

$$\mathbb{E}[3Z^2 - Z^4] = 0 \qquad \Longrightarrow \qquad \mathbb{E}Z^4 = 3.$$

Compare with a random sign:

$$\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}.$$

Both Y and Z have mean zero and variance one. But

$$\mathbb{E}Y^4 = 1 \neq 3.$$

This test can tell them apart.

All polynomial tests

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

Take $f(z) = z^k$, for $k \geq 1$.

$$\mathbb{E}Z^{k+1} = k\,\mathbb{E}Z^{k-1}.$$

One identity generates all Gaussian moments:

$$\mathbb{E}Z^{2m+1} = 0, \qquad \mathbb{E}Z^{2m} = (2m-1)!!.$$

But test functions need not be polynomials.

A more revealing test: $f(z) = e^{itz}$

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

Write $\psi(t) = \mathbb{E}e^{itZ}$. The identity gives

$$\mathbb{E}[Ze^{itZ}] = it\,\psi(t).$$

Therefore

$$\psi'(t) = i\mathbb{E}[Ze^{itZ}] = -t\psi(t).$$

With $\psi(0) = 1$,

$$\boxed{\psi(t) = e^{-t^2/2}.}$$

Equivalently, use the two real tests $\cos(tz)$ and $\sin(tz)$.

One more test: toward Poincaré

$$Z \sim \mathcal{N}(0,1) \qquad \mathbb{E}[f'(Z) - Zf(Z)] = 0$$

Take a smooth, centered g and choose $f(z) = g(z)g'(z)$. Stein's identity gives the energy identity

$$\mathbb{E}[(g'(Z))^2] = -\mathbb{E}[g(Z)(g''(Z) - Zg'(Z))].$$

With a short Gaussian semigroup argument, this leads to

$$\boxed{\mathrm{Var}(g(Z)) \leq \mathbb{E}[(g'(Z))^2].}$$

Small derivatives mean small fluctuations.

The Gaussian Poincaré inequality. The single substitution is a starting point, not the whole proof.

Why is the lemma true?

The Gaussian density satisfies

$$\varphi'(z) = -z\varphi(z).$$

Integration by parts:

$$\begin{aligned}
\mathbb{E}f'(Z) &= \int_{\mathbb{R}} f'(z)\varphi(z)\,dz \\
&= [f(z)\varphi(z)]_{-\infty}^{\infty} - \int_{\mathbb{R}} f(z)\varphi'(z)\,dz \\
&= \int_{\mathbb{R}} zf(z)\varphi(z)\,dz \\
&= \mathbb{E}[Zf(Z)].
\end{aligned}$$

Assume f is continuously differentiable, the expectations are absolutely integrable, and $f(z)\varphi(z) \to 0$ at both ends.

Now replace the Gaussian

For another random variable W, consider

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

For a Gaussian, this is zero.

If it is small for enough test functions, could W be close to Gaussian?

This is where the identity becomes a method.

What do we mean by "close"?

Compare expectations:

$$\mathbb{E}h(W) - \mathbb{E}h(Z).$$

Different tests ask different questions.

- $h(x) = \mathbf{1}_{\{x \le a\}}$: compare CDFs.
- h is 1-Lipschitz: compare Lipschitz observables.

For the second choice,

$$d_{\mathrm{W}}(W, Z) := \sup_{\mathrm{Lip}(h) \le 1} |\mathbb{E}h(W) - \mathbb{E}h(Z)|.$$

Wasserstein–1 distance; assume finite first moments.

The bridge: Stein's equation

Given h, find f_h satisfying

$$f_h'(x) - x f_h(x) = h(x) - \mathbb{E}h(Z).$$

Then, exactly,

$$\mathbb{E}h(W) - \mathbb{E}h(Z) = \mathbb{E}[f_h'(W) - W f_h(W)].$$

A distributional comparison becomes an expectation we can try to calculate.

Does this equation have a solution?

Yes. One useful solution is

$$f_h(x) = e^{x^2/2} \int_{-\infty}^{x} (h(t) - \mathbb{E}h(Z)) e^{-t^2/2}\, dt.$$

The important part is not the formula itself.

It is control of the solution.

For a smooth 1-Lipschitz test h,

$$\boxed{\|f_h''\|_\infty \leq 2.}$$

We will use this standard bound without proving it. General Lipschitz tests follow by approximation.

Back to a sum

Let

$$W = \sum_{i=1}^{n} \xi_i, \qquad \mathbb{E}\xi_i = 0, \qquad \sum_{i=1}^{n} \mathbb{E}\xi_i^2 = 1,$$

where the ξ_i are independent and have finite third moments.

Remove one contribution:

$$W_i = W - \xi_i.$$

Now W_i and ξ_i are independent.

That small observation does most of the work.

The main cancellation

Because $\mathbb{E}[\xi_i f(W_i)] = 0$,

$$\mathbb{E}[Wf(W)] = \sum_i \mathbb{E}\left[\xi_i (f(W) - f(W_i))\right].$$

Taylor-expand:

$$f(W) - f(W_i) = \xi_i f'(W_i) + R_i,$$

where

$$|R_i| \leq \frac{1}{2}\|f''\|_\infty \, \xi_i^2.$$

The error: a short sketch

Write $v_i = \mathbb{E}\xi_i^2$. Independence gives

$$\mathbb{E}[Wf(W)] = \sum_i v_i \, \mathbb{E}f'(W_i) + r,$$

with

$$|r| \leq \frac{1}{2}\|f''\|_\infty \sum_i \mathbb{E}|\xi_i|^3.$$

Since $\sum_i v_i = 1$,

$$\left|\mathbb{E}f'(W) - \sum_i v_i \mathbb{E}f'(W_i)\right| \leq \|f''\|_\infty \sum_i v_i \mathbb{E}|\xi_i|.$$

Finally,

$$v_i \mathbb{E}|\xi_i| \leq \mathbb{E}|\xi_i|^3.$$

The remainder is controlled by third absolute moments.

A quantitative CLT bound

Combining the estimates,

$$|\mathbb{E}[f'(W) - Wf(W)]| \leq \frac{3}{2}\|f''\|_\infty \sum_i \mathbb{E}|\xi_i|^3.$$

Use $f = f_h$ and $\|f_h''\|_\infty \leq 2$:

$$d_{\mathrm{W}}(W, Z) \leq 3 \sum_{i=1}^{n} \mathbb{E}|\xi_i|^3.$$

A standard Stein-method bound with a convenient, non-sharp constant; not a verbatim statement of Stein's original 1972 theorem.

For the i.i.d. CLT

Suppose also that

$$\rho := \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3} < \infty.$$

Taking

$$\xi_i = \frac{X_i - \mu}{\sigma\sqrt{n}}$$

gives

$$d_{\mathrm{W}}(W_n, Z) \leq \frac{3\rho}{\sqrt{n}}.$$

And the usual CDF error?

The Wasserstein bound is not itself a CDF bound.

Using indicator tests in Stein's equation requires additional estimates.

With that extra work, Stein's method also gives Berry–Esseen-type bounds:

$$\sup_{a\in\mathbb{R}} |\mathbb{P}(W_n \le a) - \Phi(a)| \le C\,\frac{\rho}{\sqrt{n}},$$

for a universal constant C.

Same basic strategy. More delicate test functions.

Another tool: exchangeable pairs

Build a small random perturbation W' of W so that

$$(W, W') \stackrel{d}{=} (W', W).$$

Swapping the pair changes nothing. In particular,

$$\mathbb{E}\big[(W' - W)\{f(W') + f(W)\}\big] = 0.$$

This symmetry gives another route to Stein's cancellations, including in dependent problems.

We did not need it here: independence made the leave-one-out trick enough.

Stein's method: the recipe

1. **Find a Stein operator.** For the target Z, arrange that $\mathbb{E}[\mathcal{A}f(Z)] = 0$.

2. **Solve a suitable Stein equation.**

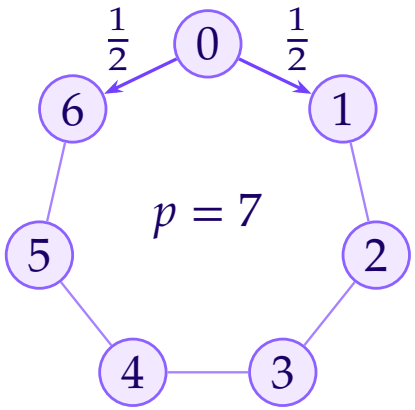
$$\mathcal{A}f_h = h - \mathbb{E}h(Z).$$

   Get useful bounds on f_h (or its derivatives).

3. **Do the dirty algebra.** Bound $\mathbb{E}[\mathcal{A}f_h(W)]$ using the structure of W.

Your turn: a walk around an odd cycle

A simple random walk on $\mathbb{Z}_p$, with $p \geq 3$ odd, converges to the uniform distribution.



Left or right, each with probability $\frac{1}{2}$.

Two Stein operators for the same target!

$$\mathcal{A}f(j) = f(j) - f(j-2),$$

$$\mathcal{L}f(j) = \frac{f(j+1) + f(j-1)}{2} - f(j).$$

The second gives a Stein equation:

$$\mathcal{L}f_h(j) = h(j) - \frac{1}{p}\sum_{k=0}^{p-1} h(k).$$

All indices are taken modulo p.

1. Can we use these to get a rate of convergence?
2. Does the second operator look familiar? is it an accident!?
3. How about its Stein equation?

The example is stolen from Persi Diaconis, *Stein's Method: Expository Lectures and Applications*.

One last question

For the standard Gaussian,

$$\mathcal{A}f(x) = f'(x) - xf(x), \qquad \mathbb{E}[\mathcal{A}f(Z)] = 0.$$

Can we construct Stein operators
for other random objects?

Chen-Stein

Target: $\mathrm{Poisson}(\lambda)$

$$\boxed{\mathcal{A}_\lambda f(k) = \lambda f(k+1) - kf(k)}$$

$k = 0, 1, 2, \dots, \qquad \lambda > 0.$

Langevin-Stein

Target: a smooth positive density p on $\mathbb{R}^d$

$$\boxed{\mathcal{T}_p f(x) = \nabla \cdot f(x) + \nabla \log p(x)^\top f(x)}$$

$f : \mathbb{R}^d \to \mathbb{R}^d$, with suitable regularity and boundary conditions.

Next session

