A short survey of Stein's method

Sourav Chatterjee

Stanford University

Assumption

- Since this is a general audience talk, I will assume that the audience has a passing familiarity with the definitions of random variable, expected value, variance and central limit theorem, but not much more.

Common pursuits in probability theory

- Some of the most common things that probabilists do are: compute or estimate expected values and variances of random variables, and prove generalized central limit theorems, sometimes on function spaces.
- A large fraction of papers in probability theory may be classified into one of the above categories, although they may not say it explicitly.
- There are many important problems where we do not know how to compute or estimate expected values (e.g. almost any statistical mechanics model in three and higher dimensions, such as the Ising model); many important problems where we do not know how to compute the order of the variance (e.g. almost any random combinatorial optimization problem, such as traveling salesman); and many important problems where we do not know how to prove generalized central limit theorems (e.g. quantum field theories).

What is Stein's method?

- Stein's method is a sophisticated technique for proving generalized central limit theorems, pioneered in the 1970s by Charles Stein, one of the leading statisticians of the 20th century.
- Recall the ordinary central limit theorem: If $X_1, X_2, \ldots$ are independent and identically distributed random variables, then

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le x\right) = \int_{-\infty}^x \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du\,,$$

  where $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$.
- Usual method of proof: The probability on the left-hand side is computed using Fourier transforms. Independence of the summands implies that the Fourier transform decomposes as a product. The rest is analysis.
- Stein's motivation: What if the X_i's are not exactly independent? Fourier transforms are usually not very helpful.

Stein's idea

- Let W be any random variable and Z be a standard Gaussian random variable. That is,

$$\mathbb{P}(Z \le x) = \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du\,.$$

- Suppose that we wish to show that W is "approximately Gaussian", in the sense that $\mathbb{P}(W \le x) \approx \mathbb{P}(Z \le x)$ for all x.
- Or more generally, $\mathbb{E}h(W) \approx \mathbb{E}h(Z)$ for any well-behaved h. (To see that this is a generalization, recall that $\mathbb{P}(W \le x) = \mathbb{E}h(W)$, where $h(u) = 1$ if $u \le x$ and 0 otherwise.)
- To put this in context, imagine that

$$W = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}\,.$$

Stein's idea (contd.)

- We have a random variable W, a standard Gaussian random variable Z, and we wish to show that $\mathbb{E}h(W) \approx \mathbb{E}h(Z)$ for all well-behaved h.
- In the Fourier-theoretic approach, we write both expectations in terms of Fourier transforms and then use analysis to show that they are approximately equal.
- The problem with this approach is that it may be hard or useless to write $\mathbb{E}h(W)$ in terms of Fourier transforms if W is a relatively complicated random variable.
- For example: Toss a coin n times, and let W be the number of times where a certain pattern, say HTHH, appears.

Stein's idea (contd.)

- Stein's idea:
  - Given h, obtain a function f by solving the differential equation
$$f'(x) - xf(x) = h(x) - \mathbb{E}h(Z)\,.$$
  - Show that $\mathbb{E}(f'(W) - Wf(W)) \approx 0$ using the properties of W.
  - Since
$$\mathbb{E}(f'(W) - Wf(W)) = \mathbb{E}h(W) - \mathbb{E}h(Z)\,,$$
    conclude that $\mathbb{E}h(W) \approx \mathbb{E}h(Z)$.
- What's the gain? Stein's method is a local-to-global approach for proving central limit theorems; often, it is possible to prove $\mathbb{E}(f'(W) - Wf(W)) \approx 0$ using small local perturbations of W.
- The Fourier-theoretic method (and most other methods of proving central limit theorems) are non-perturbative.

Stein's idea (contd.)

- Indeed, Stein's method can be successfully implemented in various examples by showing that for some function g related to f,

$$\mathbb{E}(f'(W) - Wf(W)) \approx \alpha\, \mathbb{E}(g(W') - g(W))\,,$$

where α is a real number and W' is a small perturbation of W, that has the same probability distribution as W.

- Since $\mathbb{E}g(W') = \mathbb{E}g(W)$, we may conclude that $\mathbb{E}(f'(W) - Wf(W)) \approx 0$. This is known as the method of exchangeable pairs.

Why does the method work?

- If we replace $\mathbb{E}h(Z)$ by some other constant in the differential equation
$$f'(x) - xf(x) = h(x) - \mathbb{E}h(Z)\,,$$
the f that we get is not well-behaved: it will blow up badly at infinity.
- There is a relationship between this differential equation and the Gaussian distribution.
- In fact, a random variable X has the standard Gaussian distribution if and only if $\mathbb{E}(f'(X) - Xf(X)) = 0$ for all f.
- For this reason, the differential operator T, defined as $Tf(x) = f'(x) - xf(x)$, is called a characterizing operator for the standard Gaussian distribution.
- Stein's method can be (and has been) generalized to prove other probabilistic limit theorems, even on function spaces, by working with suitable characterizing operators and solving the related differential equations.

A simple example

- Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_i^2) = 1$.
- Let $S = n^{-1/2}(X_1 + \ldots + X_n)$.
- For each i let $S_i = S - n^{-1/2}X_i$. Then S_i and X_i are independent.
- Therefore, for any f, $\mathbb{E}(X_i f(S_i)) = \mathbb{E}(X_i)\mathbb{E}(f(S_i)) = 0$.
- By first order Taylor expansion, this gives

$$\begin{aligned}\mathbb{E}(X_i f(S)) &= \mathbb{E}(X_i(f(S) - f(S_i))) \approx \mathbb{E}(X_i(S - S_i)f'(S)) \\ &= n^{-1/2}\mathbb{E}(X_i^2 f'(S))\,.\end{aligned}$$

- Thus,

$$\begin{aligned}\mathbb{E}(Sf(S)) &= n^{-1/2}\sum_i \mathbb{E}(X_i f(S)) \approx n^{-1}\sum_i \mathbb{E}(X_i^2 f'(S)) \\ &= \mathbb{E}\big((n^{-1}\sum X_i^2) f'(S)\big)\,.\end{aligned}$$

- By the law of large numbers, $n^{-1}\sum X_i^2 \approx 1$. Therefore, $\mathbb{E}(Sf(S)) \approx \mathbb{E}(f'(S))$. By Stein's method, this shows that S is approximately Gaussian.

History of Stein's method

- Too much to summarize in this talk. See my ICM article for a recap of the main developments.
- Many variants: exchangeable pairs, size-biased couplings, zero-biased couplings, dependency graphs, multivariate and function space versions, Poisson approximation, etc.
- Key figures in the historical development: Charles Stein, Louis Chen, Andrew Barbour, Erwin Bolthausen, Larry Goldstein, Gesine Reinert, Yosef Rinott, Vladimir Rotar, ...
- Several young probabilists working on developing the theory Stein's method these days, including myself. Many more are using Stein's method in their research.
- Significant opportunities for new results and new applications.

In the remainder of this talk, I will talk about my interpretation of Stein's method.

What causes central limit behavior?

- Suppose that we are investigating whether a random variable W is "approximately Gaussian".
- Traditional wisdom: If W is built out of many small random influences which are approximately independent of each other, then one may expect W to exhibit Gaussian behavior.
- It is clear what this means if W is a sum of independent or approximately independent random variables, but not otherwise.
- Examples where central limit behavior is expected but the above heuristic does not make sense in any obvious way: Minimal spanning trees, traveling salesman problem, optimal matching,....

Minimal spanning tree

- Suppose that we have an undirected graph, e.g. a discrete grid.
- On each edge of the grid, there is a random weight. The weights are independent.
- A minimal spanning tree (MST) of this graph is a subtree (i.e. cycle free subgraph) that minimizes the sum of edge weights among all subtrees.
- It is expected that under fairly general conditions, the total edge weight of a minimal spanning tree should obey a central limit theorem.
- Proved on integer lattices by Alexander (1995, for 2D) and Kesten and Lee (1996, general dimensions).
- It is not clear how to weight of MST may be seen as a sum total of many small influences that are approximately independent.

Search for a better heuristic

- Often, the random variable W of interest may be written as a function $f(X_1, \ldots, X_n)$ of a collection of independent random variables $X_1, \ldots, X_n$.
- For example, the weight of the MST is a function of the edge weights, which are independent.
- Let $\partial_i W$ be the change in W if X_i is perturbed (typically, replaced by an independent copy).
- The typical size of $\partial_i W$ is often called the "influence" of X_i on W. Widespread use in theoretical CS, machine learning and other areas.
- It is true that if the influences are small, W exhibits central limit behavior?

Search for a better heuristic (contd.)

- Not quite.
- For example, let $X_1, \ldots, X_n$ be independent random variables, with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$.
- Let
$$W = \frac{1}{n} \sum_{1 \le i < j \le n} X_i X_j \, .$$
- W has a limiting distribution that is not Gaussian.
- However,
$$\partial_i W = \frac{1}{n} \sum_{j \ne i} X_j \, ,$$
which is typically quite small.

The perturbative heuristic for central limit behavior

- It turns out that one can still formulate a heuristic for central limit behavior using influences.
- In a 2008 paper I showed — roughly speaking — that if the influences $\partial_i W$ are not only small, but also approximately independent, then W is approximately Gaussian.
- The proof uses Stein's method. Will give precise statement and proof shortly.

Coming back to MST

- Does this heuristic work for the MST? Yes.
- If the weight of an edge e is perturbed, then the change $\partial_e W$ in the total weight W may be shown to be approximately equal (with high probability) to a function of the weights of the edges close to e.
- Consequently, if e and e' are two edges that are far apart, then $\partial_e W$ and $\partial_{e'} W$ are approximately independent.
- Since most pairs of edges are far apart from each other, the perturbative heuristic implies that W is approximately Gaussian.
- This argument has been recently used to obtain rates of convergence in central limit theorems for minimal spanning trees in C. & Sen (2013).

Open problem: Traveling salesman

- n points are uniformly distributed in the unit square.
- L_n is the length of the minimum length path that touches every point.
- It is widely believed that L_n should satisfy a central limit theorem.
- According to the perturbative heuristic, one should be able to solve this open question if one shows that the change in L_n that occurs upon perturbing a single point depends (approximately) only on a small number of points in its neighborhood.
- There are many other similar questions about central limit behavior in combinatorial optimization that modern probability theory is unable to tackle. The perturbative heuristic gives a possible unified approach to attack such problems, if someone can figure out how to verify the required condition (approximate independence of influences).

Formal statement of the perturbative heuristic

- Let $X = (X_1, \ldots, X_n)$ be a vector of independent random variables.
- Let $W = f(X)$ for some function f.
- Let $X' = (X_1', \ldots, X_n')$ be an independent copy of X.
- Let $[n] = \{1, \ldots, n\}$, and for each $A \subseteq [n]$, define the random vector X^A as

$$X_i^A = \begin{cases} X_i' & \text{if } i \in A, \\ X_i & \text{if } i \notin A. \end{cases}$$

- Let

$$\partial_i f := f(X) - f(X^{\{i\}}),$$

and for each $A \subseteq [n]$ and $i \notin A$, let

$$\partial_i f^A := f(X^A) - f(X^{A \cup \{i\}}).$$

Formal statement of the perturbative heuristic (contd.)

- For each proper subset A of $[n]$ define

$$\nu(A) := \frac{1}{n\binom{n-1}{|A|}}.$$

- Let T_i be the weighted average

$$\frac{1}{2}\sum_{A\subseteq[n]\setminus\{i\}} \nu(A)\partial_i f \partial_i f^A.$$

- This T_i will be our measure of influence of X_i on W.
- The heuristic is that if the T_i's are approximately uncorrelated, then W is approximately Gaussian. To be made more precise in the next slide.

Formal statement of the perturbative heuristic (contd.)

- Let $T := \sum_{i=1}^n T_i$.
- The order of fluctuations of T is a measure of the uncorrelatedness of the T_i's. If the T_i's have small correlation, then T has a small variance.

Theorem (C., 2008)

Suppose that $\mathbb{E}(W) = 0$ *and* $\mathrm{Var}(W) = 1$. *Then*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(W \le x) - \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \right| \le 2\left(\sqrt{\mathrm{Var}(T)} + \frac{1}{4} \sum_{i=1}^{n} \mathbb{E}|\partial_i f|^3 \right)^{1/2}.$$

The first term measures the uncorrelatedness of the influences and the second term measures the smallness of the influences.

Proof sketch

- First, recall notation:
  - $X = (X_1, \ldots, X_n)$ is a vector of independent random variables.
  - $W = f(X)$ is some function of X. Assume that $\mathbb{E}(W) = 0$ and $\mathrm{Var}(W) = 1$.
  - $X' = (X'_1, \ldots, X'_n)$ is an independent copy of X.
  - For $A \subseteq [n]$, the random vector X^A is defined as

$$X_i^A = \begin{cases} X'_i & \text{if } i \in A, \\ X_i & \text{if } i \notin A. \end{cases}$$

  - Discrete derivatives are defined as

$$\partial_i f := f(X) - f(X^{\{i\}}), \quad \partial_i f^A := f(X^A) - f(X^{A \cup \{i\}}).$$

  - Let

$$T := \frac{1}{2} \sum_{i=1}^{n} \sum_{A \subseteq [n] \setminus \{i\}} \nu(A) \partial_i f \partial_i f^A$$

    where

$$\nu(A) := \frac{1}{n\binom{n-1}{|A|}}.$$

Proof sketch (contd.)

- For any g, f, A, and $i \notin A$,

$$\begin{aligned}
&\mathbb{E}(\partial_i g \partial_i f^A) \\
&= \mathbb{E}(g(X)\partial_i f^A) - \mathbb{E}(g(X^{\{i\}})\partial_i f^A) \\
&= 2\mathbb{E}(g(X)\partial_i f^A) \text{ by exchangeability of } (X_i, X_i').
\end{aligned}$$

- With $g = \varphi \circ f$, we have $\partial_i g \approx \varphi'(W)\partial_i f$, and hence

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}(\varphi'(W)\partial_i f \partial_i f^A) \\
&\approx \frac{1}{2}\mathbb{E}(\partial_i g \partial_i f^A) = \mathbb{E}(\varphi(W)\partial_i f^A).
\end{aligned}$$

- Thus,

$$\mathbb{E}(\varphi'(W)T) \approx \mathbb{E}\left(\varphi(W)\sum_{i=1}^{n}\sum_{A\subseteq[n]\setminus\{i\}}\nu(A)\partial_i f^A\right).$$

Proof sketch (contd.)

- Now note that

$$\sum_{i=1}^{n} \sum_{A \subseteq [n] \setminus \{i\}} \nu(A) \partial_i f^A = f(X) - f(X'),$$

  which is just an algebraic identity.
- Thus,

$$\begin{aligned} \mathbb{E}(\varphi'(W)T) &\approx \mathbb{E}[\varphi(W)(f(X) - f(X'))] \\ &= \mathbb{E}(\varphi(W)W). \end{aligned}$$

- Exact equality holds for $\varphi(u) = u$, which gives

$$\mathbb{E}(T) = \mathrm{Var}(W) = 1.$$

- Thus, if $\mathrm{Var}(T)$ is tiny, then

$$\mathbb{E}(\varphi(W)W) \approx \mathbb{E}(\varphi'(W)).$$

- One can now apply Stein's method to conclude that W is approximately Gaussian.

Summary

- Stein's method is a powerful tool for proving central limit theorems in the presence of dependence. Many variants.
- The perturbative heuristic is a version of Stein's method that intends to give a unified approach to understanding central limit behavior for complicated random variables.
- The heuristic says, roughly, that if a random variable W is a function of many independent random variables $X_1, \ldots, X_n$, and (a) the influence of each X_i on W, measured in a certain way, is small; and (b) the influences are approximately independent, then W may be expected to exhibit central limit behavior.
- The heuristic has been used to prove central limit theorems in statistics, random matrix theory, number theory, and other areas.
- Lots of open questions that may be approachable by this heuristic.