Depth First Learning: SVGD

Week 3: A Kernelized Stein Discrepancy

May 29, 2019

The main theoretical meat comes from a single paper titled Kernelized Stein Discrepancy (KSD) [4]. KSD takes the powerful Stein's Identity, and uses RKHS theory to define a tractable discrepancy between a ground truth distribution and samples from an arbitrary one. Most importantly, KSD defines a discrepancy function that does not involve calculating the normalizing constant, allowing it to be much more widely applicable in practical tasks. We will discuss the difference between likelihood-free and likelihood-based methods in machine learning, how this normalization constant proves to be problematic in machine learning, and how KSD allows us to sidestep this issue with a new, tractable discrepancy. KSD will serve as the launch pad for the algorithm at the focus of this curriculum, Stein Variational Gradient Descent.

1 Table of Contents

Note: In the resources linked, The roles of p and q are switched in resources 1 and 3.

2 A Stein Discrepancy

We start by recapping last week, with some minor differences. The Stein Operator, is defined as:

$$\mathcal{A}_p\mathbf{f}(x) = s_p(x)\mathbf{f}(x)^T + \nabla_x\mathbf{f}(x) \tag{1}$$

where $s_p(x)$ is the (Stein) score function of p: $s_p(x) = \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$ and both $s_p(x)$ and $\mathcal{A}\mathbf{f}$ is now a vector-valued function, mapping from $\mathcal{X} \to \mathbf{R}^d$. We say f is in the *Stein class* of smooth density p if f is smooth and satisfies:

$$\int_{x\in\mathcal{X}} \nabla_x(f(x)p(x))dx = 0 \tag{2}$$

The vector-valued function $\mathbf{f}$ is in the Stein class of p if each individual dimensional component of the function is in the Stein class of p. The first equation now has a few extra subscripts (the $_p$

subscripts), which now denote which density the operator is defined on. We are again worrying about two densities: a ground truth one p, and then an empirical one, q.

3 Why? Goodness of Fit, Its Importance, and Its Difficulties

Before we start playing around with the Stein Operator again, its important to now understand the problem setup. From Week 1's final section, we're going to be working in the general framework of model checking, and more specifically, we're concerned about *goodness-of-fit*.

Goodness of fit is absolutely crucial in the field of statistics, and encompasses the tools and methods that answer the general question of "how well does my model fit the observed data"? Given some empirical data, defined as $\{x_i\} \sim q$, and our statistical model, we're looking to either accept or reject the following null hypothesis:

$$H_0 : p = q \tag{3}$$

meaning that our empirical samples *were* drawn from the same distribution as samples from our statistical model.

However, in practice, this can become quite hairy quite quickly. Our standard goodness-of-fit tests often only work with low-dimensional data, or require extra assumptions such as being able to bin our data, making these tests restrictive for most practical settings. Other times, sampling from the ground truth distribution can become impractical due to the dreaded normalizing constant, Z, which turns any probability function into a density. Calculating this can be intractable, and most methods - including the one discussed here - will look for ways around doing so.

With the operator equation, and the null hypothesis, our first step towards a Stein discrepancy is to switch the expectation:

$$\mathbf{E}_q[\mathcal{A}_p f(x)] = 0, \ \forall f \Leftrightarrow p = q \tag{4}$$

Intuititvely, this makes sense - if p and q are truly the same, then it shouldn't matter which density we calculate the expectation under. This implies that for some $p \neq q$, there exists some f such that $\mathbf{E}_q[\mathcal{A}_p f(x)] \neq 0$. From [3], we know that we can then rewrite the operator as function of the difference in score functions, given smooth densities p and q supported on $\mathcal{X}$.

$$\mathbf{E}_q[\mathcal{A}_p f(x)] = \mathbf{E}_q[\mathcal{A}_p f(x)] - \mathbf{E}_q[\mathcal{A}_q f(x)] = \mathbf{E}_q[(s_p(x) - s_q(x))^T f(x)] \tag{5}$$

This implies that *unless* $\nabla_x \log p(x)$ is exactly $\nabla_x \log q(x)$, we can always find some function f where the above equation is nonzero. This allows us to (finally) define the (squared) Stein Discrepancy, which we write as:

$$\sqrt{\mathbb{S}(q,p)} = \max_{f \in \mathcal{F}} [\text{trace}(\mathcal{A}_p f(x))] \tag{6}$$

Wondering where the trace comes from? You can find it right before Section 3 in the KSD paper, but give it a shot yourself!

This is usually as far as Stein's Method in statistics may have gone. While it gives a functional form for optimization, choices of family $\mathcal{F}$ were usually made with desirable properties (i.e bounded norms), rather than being practically optimizable. Essentially, it all boils down to the choice of functional family: rich enough to capture the intricacies we're looking for, especially in high dimensions, but easy enough to *actually* optimize. [2] provided the first attempt at making this optimization tractable, but here, we focus on the choice of $\mathcal{F}$ being a Reproducing Kernel Hilbert Space, as done in the Kernelized Stein Discrepancy paper.

4 A Tractable Optimization

What choice of $\mathcal{F}$ is reasonable? Why not take an a kernel k and its corresponding (unit ball of) RKHS (See Week 1) to be $\mathcal{F}$? Of course, this is the choice that the authors made, and it has some nice properties that we'll take advantage of, but why this?:

1. The (kernelized) Stein Discrepancy is a valid discrepancy (proved in Definition 3.1 and Proposition 3.3 in [4]).

2. This choice removes the dependence of the normalization constant, and now only depends on p through its score function, which is easily calculatable.

3. It provides a practical and grounded *empirical* approximation to the density, through uses of a U-statistic.

While the paper is quite interesting on its own, some parts (connections to other discrepancies, empirical results on GoF) are less important to our roadmap, so we summarize the most important detail below.

With the RKHS as the family of functions, the solution to the optimization problem has a *closed form solution*, written as:

$$\mathbb{S}(q, p) = \mathbf{E}_{x,x' \sim q}[\kappa_p(x, x')] \tag{7}$$

where $\kappa_p = \text{trace}(\mathcal{A}_p^x \mathcal{A}_p^{x'} k(x, x'))$, with $\mathcal{A}_p^x$ being the Stein operator w.r.t x. This closed form solution allows for empirical estimation of the discrepancy, which allows for a natural goodness-of-fit test (based on whether the discrepancy is greater than some threshold or not).

However, this formulation has a much more relevant (to our discussion) interpretation by connecting the Stein Operator to the derivative of the KL-Divergence, leading to the centerpiece paper of our roadmap: Stein Variational Gradient Descent [5].

5 Intuition and Discussion

Below are two really interesting threads that occurred on Piazza after this week's discussion: one by Calvin Woo, and one by Sanyam Kapoor. However, before you start on the following, you may find it useful to read through the following items on:

- Integral Probability Metrics: [2, 6, 7]
- Adjoints: [1] (Chapter on Duals and Adjoints)

5.1 Sanyam's Intuition

Note: Sanyam also has an amazing blog post on Stein's Method in ML - check it out!
Going back to the integral probability metrics (IPMs) to define distance between an unknown target P and our model's estimate Q,

$$d(P, Q) = \sup_{h \in \mathcal{H}} ||\mathbf{E}_{X \sim Q}[h(X)] - \mathbf{E}_{X \sim P}[h(X)] \tag{8}$$

over an expressive enough family of test functions $\mathcal{H}$, this metric is pretty much impossible to compute because we don't know P in the first place. What Stein's method allows us to do is get rid of that second term P to get us one step closer to hopefully being clever about $\mathcal{H}$ and getting a tractable optimization problem (because Q is in our control). It gets rid of that second term containing the unknown distribution by choosing a family of functions in the Stein class of P (so that the term goes to zero).

5.2 Calvin's Alternate Analysis

Calvin Woo provided a nice post on our class Piazza, transcribed below.
To make things concrete, lets look at the case of Stein's identity for a Gaussian normal variable $X \sim \mathcal{N}(0, 1)$. Expanding out what's in the paper, we get that for any smooth function f of a suitable Hilbert space of functions, we have

$$\mathbf{E}_{x \sim \mathcal{N}(0,1)}[f'(x) - xf(x)] = 0 \tag{9}$$

Calvin's question to the class: How can we think about this? Can we prove this without resorting to a priori knowing the Stein operator?
Here were some hints:

1. The pdf of the Gaussian is:
$$p(x) = \frac{1}{\sqrt{2\pi}} \boldsymbol{e}^{\frac{-x^2}{2}} \tag{10}$$
so, $\mathcal{N}(0, 1)$ is a solution to the ordinary differential equation $p'(x) + xp(x) = 0$.

2. For L a linear operator on functions, we call L^* the adjoint (operator) equation if for any $f, g \in C_\infty(R)$ (smooth functions), we have $\langle Lf, g\rangle = \langle f, L^* g\rangle$, where the inner product is the standard function space inner product.

3. Use the two above ingredients to discover Stein's identity for the Gaussian. The first part is just a verification– now that we know it, we can write $L = \frac{d}{dx} + x$ and so the ODE becomes $Lp = 0$. Here L is thus a linear operator on functions, and so we can try to compute it's adjoint equation using the definition $\langle Lf, g\rangle = \langle f, L^* g\rangle$.

$$\begin{aligned} \langle Lf, g\rangle &= \int_{-\infty}^{\infty} [f'(x) + xf(x)]g(x)dx & (11) \\ &= \int_{-\infty}^{\infty} f(x)[g'(x) + xg(x)]dx & (12) \end{aligned}$$

where in expanding it, we used integration by parts and the boundary condition (decay of the function at infinity). Thus $L^* = \frac{d}{dx} + x$. Now Stein's identity follows for any such function g by noting that $L = 0$ implies $0 = \langle Lp, g\rangle = \langle p, L^* g\rangle$. Now just notice that $0 = \langle p, L^* g\rangle = 0$ is just the expectation value above.

References

[1] E. Chen. *The Napkin Project*. 2019.

[2] J. Gorham and L. Mackey. Measuring sample quality with stein's method, 2015.

[3] C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electron. Commun. Probab.*, 18:14 pp., 2013.

[4] Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation, 2016.

[5] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2016.

[6] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. A note on integral probability metrics and $\phi$-divergences. *CoRR*, abs/0901.2698, 2009.

[7] D. Sutherland. Two-sample tests, integral probability metrics, and gan objectives. University Lecture.