Depth First Learning: SVGD

Week 2: Stein's Method

May 22, 2019

Most of the theory we will see in this curriculum builds off the general theoretical framework of Stein's Method, a tool to obtain bounds on distances between distributions. In Machine Learning (as we shall later see), distances between distributions can be used to quantify how well (or poorly) a model is at approximating a certain distribution of interest. We shall start from Stein's Identity and Operator, while explaining their theoretical significance and working through some proofs to get an understanding of some terms (Steins Method, Stein's Discrepancy) we'll see in the coming weeks. Lastly, we will discuss why Stein's Method has historically been a theoretical tool, and hint at how ideas from Week 1 (particularly RKHS) can be used in combination with Stein's Method to build the tractable discrepancy measure at the center of Week 3's discussion.

1 Table of Contents

2 What is Stein's Method?

Stein's Method, while gaining popularity in the machine learning world, was originally developed by Charles Stein [2], and has been used countless amounts of times to bound distances between distributions, prove generalized central limit theorems, and play important roles in other parts of theoretical statistics. Generally, Stein's Method can be seen as a way to **bound the distance between probability distributions**.
Usually, we want to show that some arbitrary random variable W is *approximately* normal, meaning that, given Z - a standard Gaussian random variable:

$$\mathbf{P}(W \leq x) \approx \mathbf{P}(Z \leq x) \tag{1}$$

Generally, we're looking to show that $\mathbf{E}h(W) = \mathbf{E}h(Z)\ \forall h$ such that $\mathbf{E}(h'(Z)) < \infty$, where $\mathbf{E}[\cdot]$ is the usual expectation operator. In the following, we're going to be working probability measures P and Q on some measurable space $\mathcal{X}$, with random variables W and Y (which have distributions

P and Q respectively). In our setting, P will be a complicated distribution (in Week 3, we will define concretely on what is "complicated" about it), and Q is some simpler distribution - which we take as the standard normal - that we hope to use to approximate P.

3 The Stein Operator

Currently, our goal is to just decide if P is approximately equal to Q - in later weeks, when we discuss Stein's Method *in* a machine learning context, we will discuss on *how* to transform P into Q. We define an operator, called the *Stein Operator* $\mathcal{A} : \mathcal{F} \to \mathbf{R}$, that requires, for all functions $f \in \mathcal{F}$, the following proposition to hold:

$$\mathbf{E}[\mathcal{A}f(Y)] = 0 \quad \forall f \in \mathcal{F} \Leftrightarrow Y \sim Q \tag{2}$$

Basically, if the expectation of this operator acting on $f(Y)$ is zero for all functions in our family, we can say that the random variable Y has distribution Q. Stein's Lemma [1] for normally distributed random variables X with expectation μ and variance σ^2 tells us that, for some proper choice of g (where the expectation of g is finite), we have:

$$\mathbf{E}[g(X)(X - \mu)] = \sigma^2 \mathbf{E}[g'(X)] \tag{3}$$

Since we've chosen Q to be the standard normal, we can rewrite our original expectation as:

$$\mathbf{E}[f'(Y) - Yf(Y)] = 0 \quad \forall f \in C^1 \Leftrightarrow Y \text{ is normally distributed} \tag{4}$$

This gives us an equivalent form of our operator, allowing us to write:

$$\mathcal{A}f(x) = f'(x) - xf(x) \tag{5}$$

4 Stein's Identity

In our case, why do we care about this? The last form of the operator allows us to rewrite the operator as:

$$\mathcal{A}f(x) = f'(x) + f(x)s_p(x) \tag{6}$$

where $s_p(x)$ is the (Stein) score function of p: $s_p(x) = \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$. It's this form that will drive next week's analysis of the Kernelized Stein Discrepancy.

5 Some More Intuition

Calvin Woo provided our class with some really nice intuition regarding functions and operators in mathematics, transcribed here.

References

[1] J. E. Ingersoll. Theory of financial decision making. 1987.

[2] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602, Berkeley, Calif., 1972. University of California Press.