A Short Introduction to Kernelized Stein Discrepancy

Qiang Liu (Dartmouth College)

Machine learning and statistics are essentially about understanding data using models (typically probabilistic models). Discrepancy measures that can tell the consistency between data and models are extremely useful, and provide foundations for algorithms for all kinds of tasks, including model evaluation (telling how well a model fits the data), frequentist parameter learning (finding the model that minimizes the discrepancy with data), as well as sampling for Bayesian inference (finding a set of points ("data") to approximate the posterior distribution). See Figure 1

For practical machine learning, we also need the discrepancy measure to be tractably computable, especially for the very complex models such as graphical models and deep learning models that are widely used in machine learning these days. The familiar KL divergence KL(q, p) = \int d\mathbf{x} \log \frac{q(\mathbf{x})}{p(\mathbf{x})}, for example, is not ideal for our purpose, because it needs to calculate the log-likelihood log p(x), which is often very difficult to do in practice. This is because many models we encounter are defined up to a normalization constant, e.g., p(x) = \exp(-E(x))/Z where Z is a normalization constant (known as the partition function), which is notoriously difficult to calculate. Another difficulty related to KL divergence is that it is not straightforward how to talk about the KL divergence KL(\{x_i\}, p) between a set of data \{x_i\} (e.g., drawn from an unknown q) and model p due to need for estimating the entropy term \int d\mathbf{x} \log q (since q is known only through data \{x_i\}). We can drop the entropy term and get the logarithmic scoring rule \int d\mathbf{x} \log p (or log-likelihood) when the goal is to learn or compare different models p for a given dataset. But for the purpose of goodness-of-fit evaluation or picking the "best data points" for Bayesian inference (Figure 1), we need to have the entropy term.

Kernelized Stein discrepancy (KSD) provides a convenient way to directly assess the compatibility of data-model pairs, even for models with intractable normalization constant. This allows us to derive a host of new (sometimes surprising) algorithms for various learning and inference tasks. The basic idea of KSD comes from Stein’s identity by finding the equation, whose solution is

\[ f(x) = \frac{1}{p(x)} \int \nabla \log p(x) \cdot \xi \, d\xi \]  

in one dimension.

For notation, we denote by

\[ A_p f(x) = f(x) \nabla \log p(x) + \nabla f(x), \]

Fig. 1: Computable data-model discrepancies, such as kernelized Stein discrepancy (KSD), provide foundations for developing efficient algorithms for model checking, Bayesian inference, and frequentist parameter learning.

Computable discrepancy between data and model:

\[ S(\{x_i\}_{i=1}^n, p(x)) \]

Model Checking:
- Given both \{x_i\} and p(x)

Bayesian Inference:
- Given p(x), optimize \{x_i\}

Frequentist Learning:
- Given \{x_i\}, optimize p(x)
where $\mathcal{A}_p$ is considered as a functional operator and is called Stein operator. Note that $\mathcal{A}_p$ is something that we can actually calculate in practice, even for models with intractable normalization constants. This is because $\mathcal{A}_p$ depend on $p$ only through $\nabla_x \log p(x)$ (called the score function) which equals $-\partial_x E(x)$ when $p(x) = \exp(-E(x))/Z$.

Another critical property is that $\mathcal{A}_p$ is a linear operator, that is,

$$\mathcal{A}_p(f + g) = \mathcal{A}_p f + \mathcal{A}_p g.$$  

This property makes it easy to search for the “most discriminant $f'$”, in the sense that we discuss soon.

**Stein Discrepancy** It turns out the "reverse" of Stein’s identity is also true. By this, I mean that if we take the expectation under a distribution $q$ different from $p$ in (1), say consider

$$\mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)]$$

where the expectation is now taken under $q$ while the Stein operator $\mathcal{A}_p$ is still related to $p$, then there must exist a function $f$ so that the above quantity does not equal zero. This can be easily seen, by noting that

$$\mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] = \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] - \mathbb{E}_{x \sim q}[\mathcal{A}_q f(x)] \quad \text{//assuming Stein’s identity holds for $q$.} \quad (2)$$

$$= \mathbb{E}_{x \sim q}[f(x)(\nabla_x \log p(x) - \nabla_x \log q(x))], \quad (3)$$

that is, the Stein operator is effectively the inner product with the difference of the score function $\nabla_x \log p(x) - \nabla_x \log q(x)$ between $q$ and $p$. Therefore, (3) can be made non-zero unless $\nabla_x \log p(x) = \nabla_x \log q(x)$, implying $q = p$. In this way, Stein’s identity provides a mechanism to compare two different distributions. When $q$ is observed through a set of data $\{x_i\}$, we can consider the empirical averaging $\sum_i \mathcal{A}_p f(x_i)/n$, which will concentrate on zero if $p = q$. This allows us to measure the compatibility between data $\{x_i\}$ and model $p$, without calculating the normalization constant of $p$.

Since the above argument depends on the function $f$, it is more convenient to consider the most discriminant $f$ that maximizes the violation of Stein’s identity. This gives leads to the notion of Stein discrepancy for measuring the difference between two distributions $p$ and $q$:

$$\sqrt{S(q, p)} = \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)], \quad (4)$$

where $\mathcal{F}$ is a proper set of functions that we optimize over. Obviously, the choice of function set $\mathcal{F}$ is critical. First, it needs to be broad enough to include those functions that can actually discriminant $p$ and $q$ (from (5), these functions should have positive correlation with $\nabla_x \log p - \nabla_x \log q$). On the other hand, $\mathcal{F}$ should be chosen so that the functional optimization in (4) can be easily solvable. Traditionally, in Stein’s method developed for theoretical purpose, $\mathcal{F}$ is often chosen to be sets of functions with some bounded Lipschitz norms, which allows Stein discrepancy strong enough to upper bound other probability metrics or divergences of interest (such as total variation metric) (since $\mathcal{F}$ is large), but such $\mathcal{F}$ are not practically computable unless further approximation is applied [10]. We need to search for better sets that are both broad and solvable.

**Solving the Optimization** A key observation is that the Stein operator $\mathcal{A}_p$ is linear, and hence the objective in (4) is a linear function whenever $f$ can be represented as a linear combination $f(x) = \sum_i w_i f_i(x)$ of a set of known basis functions $f_i(x)$ with unknown coefficient $w_i$. In this case, we have

$$\mathbb{E}_q[\mathcal{A}_p f] = \mathbb{E}_q[\mathcal{A}_p \sum_i w_i f_i(x)] = \sum_i w_i \beta_i,$$

where

$$\beta_i = \mathbb{E}_{x \sim q}[\mathcal{A}_p f_i(x)].$$

Therefore, our optimization objective is in fact a linear objective on $w_i$; the optimal coefficient $w_i$ can be founded easily by:

$$\max_w \sum_i w_i \beta_i, \quad \text{s.t. } \|w\| \leq 1,$$
where $\| \cdot \|$ is certain norm (such as L2) of $w$ to prevent infinite solution. This allows us to solve $w_i$ easily, even in closed form! For example, taking $\|w\|$ to be $L_2$ norm, we have $w_i = \beta_i/\|\beta_i\|$. This is great. But to make $\mathcal{F}$ to be broad enough, we may need to use an infinite number of basis functions, and this leads us to kernelized Stein discrepancy (KSD) which takes $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS). Briefly speaking, let $k(x, x')$ be a positive definite kernel, the RKHS $\mathcal{H}$ related to $k(x, x')$ includes functions of form $f(x) = \sum_i w_i k(x, x_i)$, equipped with inner product $(f, g)_{\mathcal{H}} = \sum_{ij} w_i v_j k(x_i, x_j)$ for $g = \sum_j v_j k(x, x_j)$ and RKHS norm $\|f\|_{\mathcal{H}}^2 = \sum_{ij} w_i w_j k(x_i, x_j)$. Kernelized Stein discrepancy is defined as

$$
\sqrt{\mathbb{E}_{x \sim q}[A_p f(x)]} \leq ||f||_{\mathcal{H}} \leq 1.
$$

A critical property of such RKHS space is the reproducing property: $f(x) = (f(\cdot), k(x, \cdot))_{\mathcal{H}}$ (this is analogous to the linear combination form $f = \sum_i w_i f_i(x)$ earlier). In addition, we also have $\nabla_x f(x) = (f(\cdot), \nabla_x k(x, \cdot))_{\mathcal{H}}$ where the derivative operator is “shifted” to the kernel. Therefore, we have

$$
\mathbb{E}_{x \sim q}[A_p f(x)] = (f(\cdot), \mathbb{E}_{x \sim q}[A_p k(\cdot, x)])_{\mathcal{H}},
$$

where we shift both the expectation and Stein operator to the kernel function. Define

$$
\beta_{q,p}(\cdot) = \mathbb{E}_{x \sim q} A_p k(\cdot, x).
$$

(5)

The optimization of KSD is then framed into

$$
\max_f \langle f, \beta_{q,p} \rangle_{\mathcal{H}}, \quad \text{s.t.} \quad ||f||_{\mathcal{H}} \leq 1.
$$

It is then easy to see that the optimal $f$ should be a normalized version of $\beta_{q,p}$, that is, $f = \beta_{q,p} / ||\beta_{q,p}||_{\mathcal{H}}$, and $S(q, p) = ||\beta_{q,p}||_{\mathcal{H}}^2$. With some easy work, we can further get

$$
S(q, p) = \mathbb{E}_{x, x' \sim q} |\kappa_p(x, x')|, \quad \text{where} \quad \kappa_p(x, x') = A^x_p A^x_p k(x, x').
$$

where $A^x_p$ and $A^x_p$ represent the Stein operator w.r.t. variable $x$ and $x'$, respectively, and $\kappa_p(x, x')$ is a new “Steinialized” kernel obtained by applying Stein operator on $k(x, x')$ twice, and can be easily calculated given $\nabla_x \log p(x)$. In practice, $q$ is observed through data $\{x_i\}$ and we can replace the expectation with empirical averaging This allows us to tell if $\{x_i\}$ is drawn from $p (q = p)$ by accessing if $\mathbb{E}(\{x_i\}, p)$ is significantly different from zero using hypothesis test [1, 3].

KSD is also closely connected with several other important discrepancy measures, including maximum mean discrepancy, KL divergence, and Fisher divergence (See Figure 2 and [1]). You can find more information about Stein discrepancy and its applications in the following papers:

[1]: both works (developed independently and simultaneously by two groups of people!) introduced KSD and applied it for goodness-of-fit test.

[1]: first developed the combination of Stein’s identity with RKHS, and used it as a control variate for variance reduction.

[10]: Another form of computable Stein discrepancy that does not use kernels.

[2]: It turns out the most discriminative function $\beta^*_{q,p}$ in (5) is also the steepest descent direction of KL divergence in a special sense, which allows us to derive an surprising variational inference algorithm.
References


