A PRIMER ON STEIN'S METHOD

MAURICE DIESENDRUCK

Stein's method is used to compare the unknown distribution of a random variable $Q \sim q(x)$ to the known distribution of a random variable $P \sim p(x)$, both with support $\mathcal{I}$. Beginning with the integration by parts identity over definite integrals, **Stein's method produces a bound on a functional of the random variable Q, which measures the discrepancy between q and p.**

Throughout this primer, [1, 2] are used as a references.

1. Integration by parts identity

Integration by parts identity.
$$\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b v(x)u'(x)dx \tag{1.1}$$

When mass is zero in the limits, $[u(x)v(x)]_a^b = 0$
$$\int_a^b u(x)v'(x)dx = -\int_a^b v(x)u'(x)dx \tag{1.2}$$

Let $u(x) = 1$.
$$\int_a^b 1 * v'(x)dx = -\int_a^b v(x)1'dx = 0 \tag{1.3}$$

Let $v(x) = f(x)p(x)$.
$$\int_a^b (f(x)p(x))'dx = 0 \tag{1.4}$$

Introduce extra $p(x)$.
$$\int_a^b \frac{(f(x)p(x))'}{p(x)}p(x)dx = 0 \tag{1.5}$$

Write as expectation wrt $p(x)$.
$$\mathbb{E}_{p(x)}\Big[\frac{(f(x)p(x))'}{p(x)}\Big] = 0 \tag{1.6}$$

Expand derivative of product.
$$\mathbb{E}_{p(x)}\Big[\frac{f(x)p'(x) + f'(x)p(x)}{p(x)}\Big] = 0 \tag{1.7}$$

Simplify fraction.
$$\mathbb{E}_{p(x)}\Big[f(x)\frac{p'(x)}{p(x)} + f'(x)\Big] = 0 \tag{1.8}$$

Let $\mathcal{A}_p$ represent the functional (1.9)

$f \mapsto f(x)\dfrac{p'(x)}{p(x)} + f'(x).$
$$\mathbb{E}_{p(x)}\big[\mathcal{A}_p f(x)\big] = 0. \tag{1.10}$$

Keep equations 1.4 and 1.10 in mind, as they will be used later in this primer. Equation 1.4 defines the *Stein class* of functions $\mathcal{F}$, and equation 1.10 is a critical piece of the *Stein operator.*

2. Stein operator

The operator $\mathcal{A}_p$ is called a *Stein operator* for an unknown distribution $q(x)$ with respect to $p(x)$, when the equivalence in distributions implies equation 1.10, and vice versa:
$$q \sim p \text{ if and only if } \mathbb{E}_{q(x)}\big[\mathcal{A}_p f(x)\big] = 0 \ \forall \ f \in \mathcal{F}. \tag{2.1}$$

The *Stein equation* states that for a class $\mathcal{H}$ of measure-determining functions (i.e. real-valued functions), for each $h \in \mathcal{H}$, there is a corresponding $f = f_h \in \mathcal{F}$ such that
$$h(x) - \mathbb{E}[h(P)] = \mathcal{A}_p f(x). \tag{2.2}$$

By taking the expectation of both sides, and substituting our variable of interest Q for x, we get

$$\mathbb{E}[h(Q)] - \mathbb{E}[h(P)] = \mathbb{E}[\mathcal{A}_p f(Q)] = \begin{cases} 0, & \text{if } q \sim p, \text{ by eq. 1.10} \\ \mathbb{R}_{\neq 0}, & \text{otherwise.} \end{cases} \tag{2.3}$$

Already, equation 2.3 states that when $q \sim p$, the expectation of the operator (over the variable of interest Q) equals zero. If the result is non-zero, Q and P are not equal in distribution, and a discrepancy measure is needed.

3. Stein discrepancy

The *Stein discrepancy* measures the difference between distributions q and p, and is derived from the definition of the *integral probability metric*, which takes the following form:

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}[h(Q)] - \mathbb{E}[h(P)]| \tag{3.1}$$

From 3.1, we get the Stein discrepancy

$$d_{\mathcal{H}}(Q, P) \leq \sup_{f \in \mathcal{F}(\mathcal{H})} |\mathbb{E}[\mathcal{A}_p f(Q)]|, \tag{3.2}$$

where $\mathcal{F}(\mathcal{H}) = \{f_h | h \in \mathcal{H}\}$, is the set of solutions to the Stein equation in 2.2.

3.1. **Distances as integral probability metrics.** The integral probability metric is interesting because it admits several popular distance metrics, based on the function class $\mathcal{H}$ [3, 4]. For example:

$$\text{Wasserstein} \qquad \mathcal{F} = \{f : \|f\|_L \leq 1\} \qquad \text{“1-Lipschitz”} \tag{3.3}$$

$$\text{Total variation} \qquad \mathcal{F} = \{f : \|f\|_\infty \leq 1\} \qquad \text{“1-Bounded”} \tag{3.4}$$

$$\text{Kolmogorov} \qquad \mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}^d\} \tag{3.5}$$

$$\text{Maximum mean discrepancy} \qquad \mathcal{F} = \{f \in \mathcal{H} : \|f\|_\infty \leq 1\} \qquad \text{“RKHS”} \tag{3.6}$$

3.2. **Utility of Stein's method.** When optimizing in Variational Inference or in Generative Adversarial learning, Kullback-Leibler (KL) divergence is often used to compare distributions. In spite of its popularity, it has been shown to underestimate the variance of the posterior, and produce unstable results when the supports of the distributions do not align [5]. The Stein discrepancy enables the use of a variety of divergence measures, which each produce a different distance topology.

References

[1] Christophe Ley, Gesine Reinert, Yvik Swan, et al. Steins method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

[2] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2370–2378, 2016.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. Non-parametric estimation of integral probability metrics. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1428–1432. IEEE, 2010.

[5] Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.