An Introduction to Kernel Stein Discrepancy

Lanya Yang & Rui-Yang Zhang

Contents

Preface

Notes based on the STOR-i Masterclass "Stein's Methods as a Computational Tool" in April 2024, delivered by FX Briol.

We cover a simple introduction to Stein's method and the kernel Stein discrepancy which is a commonly used tool in Computational Statistics and Machine Learning, with applications such as the Stein Variational Gradient Descent (SVGD) (Liu et al., 2016) and the Stein thinning for MCMC convergence diagnostics (Riabiz et al., 2022). This notes showcases the basics of Stein's method and the kernel Stein discrepancy.

The word 'Stein' has been used (merely!) **91** times in this notes.

Chapter 1

Introduction to Stein's Method

Measuring the distance between two probability distributions is one of the main tasks in computational statistics and machine learning. There are many types of distances/discrepancies being frequently used, such as the KL divergence, the total variation distance, the Wasserstein distance, etc. These notions are used both as a theoretical tool to study convergence rates of algorithms and as a conceptual guide to design algorithms. For example, the KL divergence is frequently used in variational inference (Blei et al., 2017) as a loss function to enable optimisations. At this same time, it is also used in the context of being the objective function in Wasserstein gradient flow to draw links with the Langevin diffusion (Jordan et al., 1998).

In this notes, we will focus on one class of discrepancy measures based on Stein's method. The field of Stein's method began as a new way to prove the central limit theorem due to Stein (1972), and it has stayed as a topic for Probabilists for a long time (Chen et al., 2010). Only recently has the rich class of tools of Stein's methods found their way into the computational statistics and machine learning literature, with early work such as Oates et al. (2017) and Liu et al. (2016).

There are some key properties of Stein's methods that make it accessible and desirable. First, (major classes of) discrepancy measures based on Stein's method require only the unnormalised version of probability density functions. Second, these discrepancy measures can be computed for a wide range of practical and complex probabilistic models within reasonable computational time. Third, there exists a lot of existing research on the theoretical aspects of such discrepancies from a mathematical point of view, which could be helpful.

1.1 Stein Characterisation

Consider a probability distribution P and we would like to find a **characterisation** for it. A characterisation of a distribution here means a description that is in one-to-one correspondence to the distribution. Examples of common characterisations include the probability density function, the cumulative density function, and the moment generating functions. In the case of Stein's method, we have the following characterisation.

Definition 1.1. *A **Stein characterisation** for a probability distribution P is a pair $(\mathcal{S}_P, \mathcal{G}_P)$*

such that for any probability distribution Q, we have

$$\begin{aligned} Q = P &\iff \mathbb{E}_{X\sim Q}\left[\mathcal{S}_P[g](X)\right] = 0 \ \forall g \in \mathcal{G}_P \\ &\iff \mathbb{E}_{X\sim Q}\left[h(X)\right] = 0 \ \forall h = \mathcal{S}_P[g] \text{ with } g \in \mathcal{G}_P. \end{aligned}$$

Here, $\mathcal{S}_P$ is called the ***Stein operator*** *of P, $\mathcal{G}_P$ is called a* ***Stein class*** *of P, and $\mathbb{E}_{X\sim Q}\left[\mathcal{S}_P[g](X)\right] = 0 \ \forall g \in \mathcal{G}_P$ is called the* ***Stein identity*** *of P.*

One should realise that the Stein operator and the Stein class for a distribution do not have to be unique. For example, a non-zero scalar multiple of a Stein operator would still be a Stein operator. Also, if some g satisfies the Stein identity, any non-zero multiple of it would also satisfy it. It, therefore, implies that the Stein class is not unique.

1.2 Langevin Stein Operator

One example of the Stein operator for a given distribution P with density p is the Langevin Stein operator.

Definition 1.2. *A* ***Langevin Stein operator*** *for a distribution P with density p is defined as*

$$\mathcal{J}_P[g](x) := \langle \nabla_x \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle$$

for sufficiently regular g. Here, the term $\nabla_x \log p(x)$ is often called the (Stein) ***score function****.*

Note that the operator uses the density only via $\nabla \log p$, which will give the same value if we multiply the density by a nonzero constant C when we can only access the unnormalised density.

Before providing some intuitions on the construction of the Langevin Stein operator, we first try a toy example. Consider the case where the distribution P is $N(0, \sigma^2)$. Using the Langevin Stein operator, we have

$$\begin{aligned} \mathcal{J}_P[g](x) &= \langle \nabla_x \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle \\ &= \left(-\nabla_x \frac{x^2}{2\sigma^2}\right) g(x) + g'(x) \\ &= -xg(x)/\sigma^2 + g'(x). \end{aligned}$$

First, we will show that this is indeed a Stein operator.

Proposition 1.3. *The operator $\mathcal{J}_P$ defined by $\mathcal{J}_P[g](x) = -xg(x)/\sigma^2 + g'(x)$ is a Stein operator for $N(0, \sigma^2)$ where its Stein class is the set of all differentiable functions.*

Proof. First, we show that

$$\mathbb{E}_{X\sim N(0,\sigma^2)}[\mathcal{J}_P[g](X)] = 0.$$

Substituting the value of this operator and rearranging the terms, we have

$$\mathbb{E}_X[Xg(X)] = \sigma^2 \mathbb{E}_[g'(X)].$$

The right-hand side of the above equation, using integration by parts, gives us

$$\begin{aligned}
\sigma^2\mathbb{E}_[g'(X)] &= \frac{\sigma^2}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} g'(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
&= \frac{\sigma}{\sqrt{2\pi}}\left[g'(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)\right]_{-\infty}^{\infty} - \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)\left(-\frac{x}{\sigma^2}\right)dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} g(x)\exp\left(-\frac{x^2}{2\sigma^2}\right)xdx \\
&= \mathbb{E}_X[Xg(X)]
\end{aligned}$$

as desired.

Next, for the reverse direction, we first realise that if

$$xg(x) - \sigma^2 g'(x) = 1_{X\leq z}(x) - \sigma\Phi(z)$$

where Φ is the CDF of $N(0,1)$ and z is a fixed constant, then

$$0 = \mathbb{E}_{X\sim Q}[Xg(X) - \sigma^2 g'(X)] = \mathbb{E}_{X\sim Q}[1_{X\leq z}(x) - \sigma\Phi(z)] = \mathbb{P}(Q\leq z) - \sigma\Phi(z)$$

implies that Q is $N(0,\sigma^2)$, as required. It can be shown that the solution g to the equality

$$xg(x) - \sigma^2 g'(x) = 1_{X\leq z}(x) - \sigma\Phi(z)$$

has to be

$$g_z(w) = \begin{cases} \sqrt{2\pi}e^{w^2/2}\Phi(w)[1-\Phi(z)] & \text{if } w\leq z \\ \sqrt{2\pi}e^{w^2/2}\Phi(z)[1-\Phi(w)] & \text{if } w\geq z \end{cases}$$

which is certainly differentiable and will be in the Stein class. Therefore, since the reverse direction of the Stein identity holds for any g in Stein class, it will hold for g_z, which gives us the desired condition that Q is $N(0,\sigma^2)$. □

Regarding the intuition behind this operator, it is due to the infinitesimal generator of the Markov semigroup induced by the Langevin diffusion, which is also why this approach of constructing Stein operators is often called the **generator approach**.

Consider a time-homogeneous Ito diffusion process $\{X_t\}_t$ characterised by the following stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t$$

where μ is the drift term, σ is the volatility term, and $\{W_t\}_t$ is a Brownian motion. The Ito diffusion process $\{X_t\}_t$ is time-homogeneous as both μ and σ have no direct dependency on t. A nice property of this process is that it is a Markov process, and we can define a Markov transition operator P_t associated with $\{X_t\}$ by

$$P_t f(x) := \mathbb{E}[f(X_t)|X_0 = x]$$

for a sufficiently nice class of test functions f. Consequently, the set of operators $\{P_t\}_{t\geq 0}$ forms a **Markov semigroup**, which then allows one to define its **infinitesimal generator** $\mathcal{L}$ by

$$\mathcal{L}f := \lim_{t\to 0^+}\frac{P_t f - f}{t}$$

where f belongs to the same class of functions as the f before. More details on Markov processes, Markov operators, and Markov semigroup can be found in Bakry et al. (2014).

Intuitively, the generator $\mathcal{L}$ can be viewed as the gradient of the Markov operator at time 0, which serves as a fundamental tool in other equations involving the Markov process such as the Fokker-Planck equation. The Fokker-Planck equation associated with the SDE characterisation of the Ito diffusion $\{X_t\}$ is given by

$$\frac{\partial}{\partial t} p_t = \mathcal{L}^* p_t$$

where p_t is the density of the Markov operator P_t (and it is also the transition density of the Markov process), and $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$ defined by

$$\int \mathcal{L} f(x) g(x) dx = \int f(x) \mathcal{L}^* g(x) dx$$

for any sufficiently nice f, g. While the SDE characterisation represents the motions of the positions and states of the Markov process, the Fokker-Planck characterisation represents the motions of the density of the states of the Markov process. It can be shown that these two characterisations are equivalent, see standard references such as Oksendal (2013).

In the special case of the Ito diffusion being the Langevin diffusion, characterised by the SDE

$$dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dW_t \tag{1}$$

where $\{W_t\}$ is a Brownian motion and π is a probability distribution. A key property of the Langevin diffusion is that the equilibrium distribution of X_t is, in fact, the distribution π, due to the steady state derived from the Fokker-Planck equation. For the Langevin diffusion, its generator is provided by

$$\mathcal{L} f = \frac{1}{2} \nabla \log \pi(x) \frac{\partial f}{\partial x} + \frac{1}{2} \frac{\partial^2 f}{\partial x \partial x} = \frac{1}{2} \langle \nabla \log \pi, \nabla f \rangle + \frac{1}{2} \langle \nabla, \nabla f \rangle$$

for some sufficiently nice f. At this stage, it is obvious that if we replace $\nabla f / 2$ by g, we would recover the Langevin Stein operator of Definition 1.2. This is why there is 'Langevin' in the name, and this approach of generating Stein operators is called the 'generator approach'.

The fact that this generator approach is a sensible one is established in Barbour (1988). Consider the equation

$$\mathbb{E}_{X \sim Q}[\mathcal{L} g(X)]$$

where g is some function of the Stein class and $\mathcal{L}$ is the generator of a Markov chain with equilibrium P. When Q is the same as P, we have, where $X \sim Q = P$,

$$\mathcal{L} g(X) = \mathbb{E}[\lim_{t \to 0^+} [P_t g(X) - g(X)]/t] = \mathbb{E}[\lim_{t \to 0^+} [g(X) - g(X)]/t] = 0$$

as X follows the equilibrium distribution of the Markov chain with transition kernel P_t. The reverse of the above statement, i.e. the expectation is zero implies $Q = P$, is slightly harder to show, but it should make intuitive sense. Additionally, the generator approach allows us to even construct Stein operators in cases where we are working on a manifold instead of the Euclidean space (which is the default here).

1.3 Stein Discrepancy

Using the Stein characterisation of a distribution P stated in Definition 1.1, we can construct a way to measure the distance between two distributions P and Q. The term 'distance' here does not mean 'metric' in the full mathematical sense as not all conditions of a metric are satisfied.

A general discrepancy for distributions is the **integral probability metric** (IPM), which is defined by

$$d_H(P,Q) := \sup_{h \in H} |\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)]| \tag{2}$$

where P, Q are probability measures with the same support $\mathcal{X}$ and H is a class of measurable test functions. A lot of commonly used discrepancies can be viewed as special cases of IPMs. For example, if we choose

$$H = \left\{ h : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \sup_{x \neq y \in \mathbb{R}^d} \frac{|h(x) - h(y)|}{\|x - y\|_2} \leq 1 \right\}$$

to be the set of functions with Lipschitz constant at most 1, then we would recover the L^1-Wasserstein distance. If we choose

$$H = \{f : \mathcal{X} \to [0,1]\},$$

then we would get the total variation distance. One should note that the class of discrepancies that can be phrased as an IPM does not include the KL divergence.

Using this framework, a choice of the class of functions would be

$$H = \{h = \mathcal{S}_P[g] \text{ with } g \in \mathcal{G}_P\}$$

for Stein characterisation $(\mathcal{S}_P, \mathcal{G}_P)$ of distribution P. We would then obtain the **Stein discrepancy** $\mathcal{S}$ as

$$\mathcal{S}(P,Q) := \sup_{g \in \mathcal{G}_P} |\mathbb{E}_P[\mathcal{S}_P[g](X)] - \mathbb{E}_Q[\mathcal{S}_P[g](X)]| = \sup_{g \in \mathcal{G}_P} |\mathbb{E}_Q[\mathcal{S}_P[g](X)]| \,. \tag{3}$$

A justification of why this is a reasonable measurement is that, due to the Stein identity, the above quantity $\mathcal{S}(P,Q)$ would be small if Q is sufficiently close to P, and large if otherwise. This discrepancy, however, does not offer an intuitive interpretation like the total variation distance or the Wasserstein distance.

After identifying this discrepancy, a reasonable question to ask is, how should we compute this in practice? The quantity involves a supremum, which is taken over an infinite set, so a direct enumeration approach of computation is doomed to fail. In the next chapter, we will consider a computationally feasible and equivalent formulation to the Stein discrepancy which is the key to making it implementable.

Chapter 2

Kernel Stein Discrepancy

In the previous chapter, we have focused mostly on Stein operators, while putting little emphasis on the Stein class. If we wish to compute the Stein discrepancy in practice, we have to consider constructing a nice class of functions for the Stein class that is sufficiently large to be a Stein class, yet not too large to make computation infeasible. This relies on the concept of a reproducing kernel Hilbert space.

2.1 Reproducing Kernel Hilbert Space

2.1.1 Positive Definite Kernel

The building block of a reproducing kernel Hilbert space, the central topic of this section, is a (positive definite) kernel.

Definition 2.1. *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **positive definite kernel** if it satisfies:*

1. *(symmetric) $k(x, y) = k(y, x)$ $\forall x, y \in \mathcal{X}$.*
2. *(positive semi-definite) For all $n \in \mathbb{N}$, $c_1, c_2, \ldots, c_n \in \mathbb{R}$, and $x_1, x_2, \ldots, x_n \in \mathcal{X}$, we have*

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

Following the definition, we can represent k in a matrix form like

$$k = \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix}$$

which needs to be symmetric (i.e. $k = k^T$) and positive semi-definite.

Some of the direct properties of a kernel is that for two positive definite kernels k_1, k_2 defined on the same space $\mathcal{X} \times \mathcal{X}$, we can show that $k_1 + k_2$ and $k_1 \cdot k_2$ (where the addition and multiplication are defined pointwise) are positive definite kernels too. Also, a scaling of a positive definite kernel k using a function $f : \mathcal{X} \to \mathbb{R}$ defined by

$$k_f(x, y) := f(x)k(x, y)f(y)$$

for any $x, y \in \mathcal{X}$ will still give a positive definite kernel.

Examples of such kernels include the **squared exponential** kernel

$$k(x, y) = \lambda \exp\left(-\frac{\|x - y\|_2^2}{l}\right)$$

where $l > 0$ is the length scale and $\lambda > 0$ is the amplitude of the kernel, and the **polynomial** kernel

$$k(x, y) = \lambda(x^T y + c)^p$$

for amplitude $\lambda > 0$, degree $p > 0$, and intercept $c \geq 0$. These kernels are very closely linked to the covariance functions in Gaussian processes, and this indicates an intimate connection between kernels and Gaussian processes. This connection will not be explored here, and one could refer to Kanagawa et al. (2018) for a survey on it.

In general, we have the general form of **radial kernel** which is given by

$$k(x, y) = \lambda\phi(-\|x - y\|_2^2/l^2)$$

for a bounded function $\phi : \mathcal{X} \to \mathbb{R}$, hyperparameters amplitude $\lambda > 0$ and length scale $l > 0$. This type of formulation highlights the translational invariance of the kernel.

2.1.2 Reproducing Kernel Hilbert Space

One could build a Hilbert space using the kernel such that the space can be equipped with an inner product possessing a nice (reproducing) property.

We first state some basic functional analysis definitions for completeness.

Definition 2.2. *An **inner product space** is a vector space X equipped with a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ called an **inner product** such that*

1. $\langle x, x \rangle \geq 0$ *for all* $x \in X$.
2. $\langle x, x \rangle = 0 \iff x = 0$.
3. $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ *for all* $\lambda, \mu \in \mathbb{R}$ *and* $x, y, z \in X$.
4. $\langle x, y \rangle = \langle y, x \rangle$ *for all* $x, y \in X$.

A consequence of having an inner product space structure is that we have the **Cauchy-Schwartz inequality**, i.e.

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\| \tag{4}$$

for any $x, y \in X$ and $\|x\| := \sqrt{\langle x, x \rangle}$ is the **norm**.

Definition 2.3. *An inner product space $(X, \langle \cdot, \cdot \rangle)$ is a **Hilbert space** if it is complete in the induced norm $\| \cdot \|$, i.e. every Cauchy sequence in the space converges.*

Next, we wish to define a specific type of Hilbert space that is the central concept of this section.

Definition 2.4. *A **reproducing kernel Hilbert space** (RKHS) H_k with positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on the set X needs to satisfies the following conditions*

1. $k(x, \cdot) \in H_k$ *for all* $x \in \mathcal{X}$.
2. *(reproducing)* $f(x) = \langle f, k(x, \cdot) \rangle_{H_k}$ *for all* $x \in \mathcal{X}$ *and* $f \in H_k$.

Here, $\langle \cdot, \cdot \rangle_{H_k}$ is the inner product equipped by the RKHS H_k.

The above definition of an RKHS is slightly mysterious and abstract. We will, in a relatively hand-wavy manner, construct an RKHS using a kernel k below to illustrate the point.

Building an RKHS

Fix a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and define the base space H_k^0 as

$$H_k^0 := \text{span}\{k(x, \cdot) \mid x \in \mathcal{X}\} = \left\{ f(\cdot) = \sum_{i=1}^{n} c_i k(x_i, \cdot) \mid n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}, x_1, \ldots, x_n \in \mathcal{X} \right\}.$$

It should be immediate that H_k^0 is a vector space over $\mathbb{R}$ by looking at its explicit constructions of each element f.

Next, we wish to equip this space with an inner product. We will force this inner product to have the reproducing property, in the sense that

$$f(x) = \langle f, k(x, \cdot) \rangle_{H_k} \ \forall f \in H_k^0, x \in \mathcal{X}.$$

This inner product with reproducing property is the defining feature of an RKHS. Consider some $f, g \in H_k^0$ that can be written in the following form

$$f = \sum_{i=1}^{n} a_i k(x_i, \cdot), \qquad g = \sum_{j=1}^{m} b_j k(y_j, \cdot).$$

We can calculate their inner product

$$\langle f, g \rangle_{H_k} = \sum_i \sum_j a_i b_j \langle k(x_i, \cdot), k(y_j, \cdot) \rangle_{H_k} = \sum_i \sum_j a_i b_j k(x_i, y_j)$$

which can be directly used as a definition of the inner product $\langle \cdot, \cdot \rangle_{H_k}$.

One could also use the above formulation to define the corresponding norm of the RKHS, which is

$$\|f\|_{H_k}^2 = \langle f, f \rangle_{H_k} = \sum_i \sum_j a_i a_j k(x_i, x_j).$$

Naturally, one would want to check if the inner product we just defined is indeed an inner product on H_k^0. Also, if the way of writing $f \in H_k^0$ in terms of sums is unique. It turns out that it is indeed an inner product, and any equivalent linear formulation of the function f would yield the same value under the inner product. The proofs are omitted here and can be found in Chapter 1 of Berlinet and Thomas-Agnan (2011).

Finally, the current space H_k^0 is not complete, and we will make it complete by simply taking its closure with respect to the induced norm $\| \cdot \|_{H_k}$, which is formally stated as

$$H_k = \overline{H_k^0}.$$

This gives us a complete inner product space, which is simply called a Hilbert space, H_k with equipped inner product $\langle \cdot, \cdot \rangle_{H_k}$ satisfying the reproducing property - and we have finished our construction of an RKHS using a positive definite kernel k.

2.2 Kernel Stein Discrepancy

We will derive the kernel Stein discrepancy in this section.

2.2.1 Kernel Mean Embedding and Maximum Mean Discrepancy

Consider a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and its RKHS H_k. A direct application of an RKHS is we can represent probability measures supported on $\mathcal{X}$ as points in the RKHS.

Consider a probability measure P on $\mathcal{X}$ that admits density p. We define

$$\mu_P := \int k(x, \cdot) dP(x) = \int k(x, \cdot) p(x) dx$$

which is called the **kernel mean embedding**. Using this, for any $f \in H_k$, we have

$$\mathbb{E}_P[f(X)] = \int f(x) dP(x) = \int \langle f, k(x, \cdot) \rangle dP(x) = \left\langle f, \int k(x, \cdot) dP(x) \right\rangle = \langle f, \mu_P \rangle$$

which should illustrate the purpose of the above definition. The name of the kernel mean embedding is because we can consider μ_P as a function that takes a probability measure P to an element in the RKHS H_k using the mapping defined above, by focusing only on the mean of the distribution under the kernel. A direct consequence of this definition is that we can compare distributions P, Q supported on $\mathcal{X}$ by comparing their embeddings μ_P, μ_Q in H_k using the RKHS norm. This gives us the **maximum mean discrepancy** (MMD)

$$\text{MMD}(P \| Q) := \| \mu_P - \mu_Q \|_{H_k}.$$

Before drawing the connection between MMD and IPM, we first need to make some remarks on checking if μ_P, μ_Q are indeed elements of H_k.

Proposition 2.5. *If $\mathbb{E}_P[\sqrt{k(X, X)}] < \infty$, then $\mu_P \in H_k$ and $\int f dP = \langle f, \mu_P \rangle_{H_k}$ for any $f \in H_k$.*

Proof. Define an operator L as $Lf = \int f dP$. We have

$$\begin{aligned}
|Lf| &= \left| \int f(x) dP(x) \right| \\
&\leq \int |f(x)| dP(x) \qquad \textit{Jensen inequality} \\
&= \int |\langle f, k(x, \cdot) \rangle_{H_k}| \, dP(x) \qquad \textit{reproducing property} \\
&\leq \int \|f\|_{H_k} \|k(x, \cdot)\|_{H_k} dP(x) \qquad \textit{Cauchy Schwartz inequality} \\
&= \|f\|_{H_k} \int \sqrt{k(x, x)} dP(x) \qquad \textit{definition of norm} \\
&< \|f\|_{H_k} \cdot M
\end{aligned}$$

for some constant $M > \mathbb{E}_P[\sqrt{k(X, X)}]$. This indicates that L is a bounded (linear) operator from H_k to $\mathbb{R}$.

Using the Riesz representation theorem, since L is a bounded functional, there exists a unique element $g \in H_k$ such that

$$Lf = \langle f, g \rangle_{H_k}$$

for all $f \in H_k$. Additionally, if we pick $f(\cdot) = k(y, \cdot)$ for some $y \in \mathcal{X}$, we have

$$h(y) = \langle k(y, \cdot), h \rangle_{H_k} = \langle f, g \rangle_{H_k} = Lf = \int k(y, x) dP(x)$$

which means, by the symmetry of k,

$$H_k \ni h = \int k(x, \cdot) dP(x) = \mu_P$$

as desired. □

Next, we will show the connection between MMD and IPM, as promised. In fact, MMD is just IPM with the considered set of test function H being $H = \{h \in H_k, \|h\|_{H_k} \leq 1\}$.

Proposition 2.6. *For probability distributions P, Q, we have*

$$MMD(P\|Q) = \sup_{h \in H} |\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)]| \tag{5}$$

where $H = \{h \in H_k, \|h\|_{H_k} \leq 1\}$.

Proof. First, for any $h \in H_k$, we have

$$\begin{aligned} |\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)]| &= \left| \int h(x) dP(x) - \int h(y) dQ(y) \right| \\ &= |\langle h, \mu_P \rangle_{H_k} - \langle h, \mu_Q \rangle_{H_k}| \\ &= |\langle h, \mu_P - \mu_Q \rangle_{H_k}| \end{aligned}$$

using the definition of the kernel mean embedding. If we take the supremum of the above quantity over H, we can use Cauchy Schwartz inequality to show that the optimal choice of h needs to be linearly dependent of $\mu_P - \mu_Q$. Since we further have the condition that $\|h\|_{H_k} \leq 1$ due to our choice of H, we have the optimiser

$$h^* = (\mu_P - \mu_Q)/\|\mu_P - \mu_Q\|_{H_k},$$

which gives us

$$\sup_{h \in H} |\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)]| = |\langle h^*, \mu_P - \mu_Q \rangle_{H_k}| = \|\mu_P - \mu_Q\|_{H_k} = \text{MMD}(P\|Q)$$

as desired. □

2.2.2 Kernel Stein Discrepancy

We are one step away from defining the kernel Stein discrepancy. Notice the Stein discrepancy of Equation (3) is very similar to the maximum mean discrepancy of Equation (5) using RKHS. As long as we can show that the set $H = \{h \in H_k, \|h\|_{H_k} \leq 1\}$ can be the Stein class of some Stein operator for some choice of kernel, then the connection can be fully drawn.

We will use the Langevin Stein operator as a starting point. For distribution P with density p, its Langevin Stein operator $\mathcal{J}_P$, provided by Definition 1.2, is

$$\mathcal{J}_P f(x) = \langle \nabla_x \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle.$$

We will also use the following derivative kernel properties

$$\nabla f(x) = \langle f, \nabla k(x, \cdot) \rangle_{H_k}, \qquad \nabla k(x, x') = \langle k(x, \cdot), \nabla k(x', \cdot) \rangle_{H_k}$$

for differential kernel k and $f \in H_k$. These results are proved in Steinwart and Christmann (2008).

Next, we try to rewrite $\mathcal{J}_P$ using a differential kernel k and inner product $\langle \cdot, \cdot \rangle_{H_k}$ in its RKHS. We have

$$\begin{aligned}
\mathcal{J}_P f(x) &= \langle \nabla_x \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle \\
&= \langle f, \nabla_x \log p(x) k(x, \cdot) \rangle_{H_k} + \langle f, \nabla k(x, \cdot) \rangle_{H_k} \\
&= \langle f, [\nabla_x \log p(x) k(x, \cdot) + \nabla k(x, \cdot)] \rangle_{H_k} \\
&=: \langle f, \xi_k(x) \rangle_{H_k}.
\end{aligned}$$

Then, the Stein discrepancy using the Langevin Stein operator will be maximising

$$|\mathbb{E}_Q[\mathcal{J}_P f(X)]| = |\langle f, \mathbb{E}_Q[\xi_k(X)] \rangle_{H_k}|$$

If we maximise it over $H = \{h \in H_k, \|h\|_{H_k} \leq 1\}$, we then have

$$\sup_{f \in H} |\mathbb{E}_Q[\mathcal{J}_P f(X)]| = \sup_{f \in H} |\langle f, \mathbb{E}_Q[\xi_k(X)] \rangle_{H_k}| = \|\mathbb{E}_Q[\xi_k(X)]\|_{H_k}$$

which is

$$\begin{aligned}
\|\mathbb{E}_Q[\xi_k(X)]\|^2_{H_k} &= \|\mathbb{E}_Q\left[\nabla_X \log p(X) k(X, \cdot) + \nabla k(X, \cdot)\right]\|^2_{H_k} \\
&= \mathbb{E}_{X,Y \sim Q}[\nabla_X \log p(X) \nabla_Y \log p(Y) k(X, Y) + \nabla_X \log p(X) \nabla_Y k(X, Y) \\
&\quad + \nabla_Y \log p(Y) \nabla_X k(X, Y) + tr(\nabla_{XY} k(X, Y))].
\end{aligned}$$

This gives us the desired **kernel Stein discrepancy** (KSD) between probability measures P, Q.

Definition 2.7. *For two probability measures P, Q with support on $\mathcal{X}$ that admit densities p, q respectively, consider a differentiable (and sufficiently nice) kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and its corresponding RKHS H_k, the **kernel Stein discrepancy** with respect to P $KSD(P\|Q)$ is defined by*

$$KSD(P\|Q)^2 = \sup_{f \in \left\{ \substack{h \in H_k \\ \|h\|_{H_k} \leq 1} \right\}} |\mathbb{E}_Q[\mathcal{S}_P f(X)]|^2 = \mathbb{E}_{X,Y \sim Q}[k_P(X, Y)] \approx \frac{1}{n} \sum_{i,j=1}^{n} k_P(x_i, x_j) \tag{6}$$

where $H = \{h \in H_k, \|h\|_{H_k} \leq 1\}$ and

$$k_P(X, Y) = s_P(X) s_P(Y) k(X, Y) + s_P(X) \nabla_Y k(X, Y) + s_P(Y) \nabla_X k(X, Y) + tr(\nabla_{XY} k(X, Y)) \tag{7}$$

where $s_P(X) = \nabla_X \log p(X)$ and $s_P(Y) = \nabla_Y \log p(Y)$.

This has turned our Stein discrepancy from a supremum over an infinite set to an expectation - which is suddenly magnitudes better computationally. In the case of only having an empirical distribution Q_n of Q made up of samples $x_1, x_2, \ldots, x_n$, we have

$$\text{KSD}(P\|Q_n)^2 = \frac{1}{n} \sum_{i,j=1}^{n} k_P(x_i, x_j).$$

Note that we have ignored some technical details on the existence and correctness of certain things. These can be found in references of Anastasiou et al. (2023). Also, we limit our discussion to continuous distributions. The kernel Stein discrepancy can be extended to discrete distributions too, which we omit for now.

Bibliography

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q. et al. (2023). Stein's method meets computational statistics: A review of some recent developments, *Statistical Science* **38**(1): 120–139.

Bakry, D., Gentil, I., Ledoux, M. et al. (2014). *Analysis and Geometry of Markov Diffusion Operators*, Vol. 103, Springer.

Barbour, A. D. (1988). Stein's method and Poisson process convergence, *Journal of Applied Probability* **25**(A): 175–184.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert spaces in Probability and Statistics*, Springer Science & Business Media.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians, *Journal of the American statistical Association* **112**(518): 859–877.

Chen, L. H., Goldstein, L. and Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*, Springer Science & Business Media.

Jordan, R., Kinderlehrer, D. and Otto, F. (1998). The variational formulation of the Fokker–Planck equation, *SIAM journal on mathematical analysis* **29**(1): 1–17.

Kanagawa, M., Hennig, P., Sejdinovic, D. and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences, *arXiv preprint arXiv:1807.02582* .

Liu, Q., Lee, J. and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests, *International conference on machine learning*, PMLR, pp. 276–284.

Oates, C. J., Girolami, M. and Chopin, N. (2017). Control functionals for Monte Carlo integration, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(3): 695–718.

Oksendal, B. (2013). *Stochastic Differential Equations*, Springer Science & Business Media.

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L. and Oates, C. J. (2022). Optimal thinning of MCMC output, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**(4): 1059–1081.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 6, University of California Press, pp. 583–603.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*, Springer Science & Business Media.