Notes on Stein's Method (221201)

Jiachun Jin

School of Information Science and Technology

ShanghaiTech University

This note will contain some core concepts required to understand the Stein's method.

Contents

1 Kernelized Stein discrepancy

1.1 Background [Liu, 2016]

Given data: $\{\mathbf{x}_i\}_{i=1}^n$, and model: $p(\mathbf{x})$. We want some discrepancy measures that can tell the consistency between data and models. They have wide applications in:

- Model evalution: $\{\mathbf{x}_i\}_{i=1}^n$ and $p(\mathbf{x})$ are both given, (discrepancy measures tell us how well a model fits data).
- Frequentist parameter learning: $\{\mathbf{x}_i\}_{i=1}^n$ is given and we optimize $p(\mathbf{x})$, (find the model that minimizes the discrepancy with data).

- Sampling for Bayesian inference: $p(\mathbf{x})$ is given and we want to optimize $\{\mathbf{x}_i\}_{i=1}^n$, (find a set of points ("data") to approximate the posterior distribution).

The discrepancy measure should to be tractably computable, the famous KL divergence $D_{\mathrm{KL}}[p(\mathbf{x}) \parallel q(\mathbf{x})] = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right]$ is not ideal for this case because:

- $\log q(\mathbf{x})$ is required, however, a lot models are only known up to a normalization constant, e.g. energy based models (EBMs): $q(\mathbf{x}) = \exp(-E(\mathbf{x}))/Z$, where $Z = \int_{\mathcal{X}} \exp(-E(\mathbf{x}))\,\mathrm{d}\mathbf{x}$ is the normalization constant.
- It is not straightforward to talk about the KL divergence $D_{\mathrm{KL}}(\{\mathbf{x}_i\}_{i=1}^n \parallel p(\mathbf{x}))$ between a set of data points (drawn from a distribution q) and the model, since in this way we have to do density estimation (or entropy estimation) for $\{\mathbf{x}_i\}_{i=1}^n$.

Kernelized Stein discrepancy (KSD) [Liu et al., 2016] provides a convenient way to directly assess the compatibility of data-model pairs, even for models with intractable normalization constant.

For simplicity, in the following $f(\cdot)$ is always referred to a scalar-valued function, and the data points $\mathbf{x}$'s are also scalars.

1.2 Stein's identity

For distributions with smooth density $p(\mathbf{x})$ and function $f(\mathbf{x})$ (supported on $\mathbb{R}$) that satisfies $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})f(\mathbf{x}) = 0$, we have:

$$\mathbb{E}_{p(\mathbf{x})}[\nabla_{\mathbf{x}} \log p(\mathbf{x}) f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})] = 0, \quad \forall f. \tag{1}$$

Proof.

$$\begin{aligned}
\int p(\mathbf{x})[\nabla_{\mathbf{x}} \log p(\mathbf{x}) f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})] &= \int [\nabla_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}) + p(\mathbf{x}) \nabla_{\mathbf{x}} f(\mathbf{x})]\,\mathrm{d}\mathbf{x} \\
&= \int \nabla_{\mathbf{x}}[f(\mathbf{x})p(\mathbf{x})]\,\mathrm{d}\mathbf{x} \\
&= \lim_{\mathbf{x}\to\infty} p(\mathbf{x})f(\mathbf{x}) - \lim_{\mathbf{x}\to-\infty} p(\mathbf{x})f(\mathbf{x}) \\
&= 0.
\end{aligned} \tag{2}$$

□

Here we define $\mathcal{A}_p f(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})$, where $\mathcal{A}_p$ is called the *Stein operator*. And we say that a function $f : \mathcal{X} \to \mathbb{R}$ is in the *Stein class* of p if f is smooth and satisfies:

$$\int_{\mathbf{x}\in\mathcal{X}} \nabla_{\mathbf{x}}(f(\mathbf{x})p(\mathbf{x}))\,\mathrm{d}\mathbf{x} = 0. \tag{3}$$

1.3 (Kernelized) Stein discrepancy

Consider $\mathbb{E}_q[\mathcal{A}_p f(\mathbf{x})] = \mathbb{E}_q[\mathcal{A}_p f(\mathbf{x})] - \mathbb{E}_q[\mathcal{A}_q f(\mathbf{x})] = \mathbb{E}_{q(\mathbf{x})}[f(\mathbf{x})(\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x}))]$ (the equation holds because of Lemma 1). In this way, Stein's identity provides a mechanism to compare two different distributions. It is convenient to consider the most discriminant f that maximizes the violation of Stein's identity, this leads to the notion of Stein discrepancy for measuring the difference between two distributions p and q:

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

where $\mathcal{F}$ is a proper set of functions that we optimize over.

When f can be represented as a linear combination $f(\cdot) = \sum_i w_i f_i(\cdot)$ of a set of **known** basis functions $f_i(\cdot)$, with unknown coefficients w_i (give an example of Fourier series here). In this case we have:

$$\begin{aligned} \mathbb{E}_q\left[\mathcal{A}_p f\right] &= \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \sum_i w_i f_i(\mathbf{x})\right] \\ &= \sum_i w_i \beta_i, \end{aligned} \tag{5}$$

where $\beta_i = \mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_p f_i(\mathbf{x})\right]$, which is a fixed scalar when $\mathbf{x}$ is a scalar. Then the optimization problem delivered in equation 4 becomes to:

$$\max_{\mathbf{w}} \sum_i w_i \beta_i, \quad s.t. \quad \|\mathbf{w}\| \leq 1, \tag{6}$$

and the optimal solution with closed form can be easily got as $w_i^* = \beta_i / \|\beta_i\|$.

Kernelized Stein discrepancy (KSD) takes $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS) with kernel $k(\cdot,\cdot)$. (The RKHS $\mathcal{H}$ related to $k(\cdot,\cdot)$ contains functions of form $f(\cdot) = \sum_i w_i k(\mathbf{x}_i, \cdot)$. Q: what is $\mathbf{x}_i$? A: related to the reproducing property.) And KSD is defined as:

$$\sqrt{S(q,p)} = \max_{f\in\mathcal{H}} \mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_p f(\mathbf{x})\right], \quad s.t. \quad \|f\|_{\mathcal{H}} \leq 1. \tag{7}$$

To use a RKHS $\mathcal{H}$ as $\mathcal{F}$, we should make sure that $\forall f \in \mathcal{H}$ is in the *Stein class* of p, and this is carefully discussed in Section 3 of [Liu et al., 2016], in the following we simply assume $k(\mathbf{x},\cdot)$ and $k(\cdot,\mathbf{x})$ are in the *Stein class* of p for any fixed $\mathbf{x}$.

Our goal is to derive a computational tractable closed form solution to equation 7. First, by the reproducing property of RKHS [Sejdinovic and Gretton, 2012], we have:

$$f(\mathbf{x}) = \langle f(\cdot), k(\mathbf{x},\cdot)\rangle_{\mathcal{H}}, \tag{8}$$
$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \langle f(\cdot), \nabla_{\mathbf{x}} k(\mathbf{x},\cdot)\rangle_{\mathcal{H}}, \tag{9}$$

with the reproducing property and the definition of Stein's operator, we have:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_p f(\mathbf{x})\right] &= \mathbb{E}_{q(\mathbf{x})}\left[\nabla_{\mathbf{x}} \log p(\mathbf{x}) f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})\right] && (10) \\ &= \mathbb{E}_{q(\mathbf{x})}\left[\nabla_{\mathbf{x}} \log p(\mathbf{x}) \langle f(\cdot), k(\mathbf{x},\cdot)\rangle_{\mathcal{H}} + \langle f(\cdot), \nabla_{\mathbf{x}} k(\mathbf{x},\cdot)\rangle_{\mathcal{H}}\right] && (11) \\ &= \langle f(\cdot), \mathbb{E}_{q(\mathbf{x})}\left[k(\mathbf{x},\cdot)\nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} k(\mathbf{x},\cdot)\right]\rangle_{\mathcal{H}} && (12) \\ &= \langle f(\cdot), \mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_p k(\mathbf{x},\cdot)\right]\rangle_{\mathcal{H}} && (13) \\ &= \langle f(\cdot), \beta_{q,p}(\cdot)\rangle_{\mathcal{H}}, && (14) \end{aligned}$$

equation 12 holds because of the linearity of expectation and inner product operation, in equation 14 we define $\beta_{q,p}(\cdot) = \mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_p k(\mathbf{x},\cdot)\right]$, and similar to equation 6, we have the optimal solution to equation 7:

$$f^*(\cdot) = \beta_{q,p}(\cdot) / \|\beta_{q,p}(\cdot)\|_{\mathcal{H}}, \tag{15}$$

and $\sqrt{S(q,p)} = \|\beta_{q,p}(\cdot)\|_{\mathcal{H}}$, $S(q,p) = \|\beta_{q,p}(\cdot)\|^2_{\mathcal{H}}$. Thus, we have:

$$\begin{aligned}
S(q,p) &= \langle \beta_{q,p}(\cdot), \beta_{q,p}(\cdot)\rangle_{\mathcal{H}} && (16)\\
&= \langle \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p k(\mathbf{x},\cdot)\right], \mathbb{E}_{\mathbf{x}'\sim q}\left[\mathcal{A}_p k(\mathbf{x}',\cdot)\right]\rangle_{\mathcal{H}} && (17)\\
&= \langle \mathbb{E}_{\mathbf{x}\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))k(\mathbf{x},\cdot)\right], \mathbb{E}_{\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}') - s_q(\mathbf{x}'))k(\mathbf{x}',\cdot)\right]\rangle_{\mathcal{H}} && (18)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top \underbrace{k(\mathbf{x},\mathbf{x}')(s_p(\mathbf{x}') - s_q(\mathbf{x}'))}_{①}\right], && (19)
\end{aligned}$$

we use $s_p(\mathbf{x})$ in equation 18 to denote $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, and the equality holds because of Lemma 1. The form in equation 19 still contains the intractable $s_q(\cdot)$, we will further make it computationally tractable.

First, note that we can apply Lemma 1 to ① in equation 19 by keeping $\mathbf{x}$ fixed (denote $k(\mathbf{x},\mathbf{x}') = k_{\mathbf{x}}(\mathbf{x}')$ in this case), then we have:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top k_{\mathbf{x}}(\mathbf{x}')(s_p(\mathbf{x}') - s_q(\mathbf{x}'))\right] && (20)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top \mathcal{A}_p k_{\mathbf{x}}(\mathbf{x}')\right] && (21)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top \left(k_{\mathbf{x}}(\mathbf{x}')\nabla_{\mathbf{x}'} \log p(\mathbf{x}') + \nabla_{\mathbf{x}'} k_{\mathbf{x}}(\mathbf{x}')\right)\right] && (22)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top v(\mathbf{x},\mathbf{x}')\right], && (23)
\end{aligned}$$

where we denote $v(\mathbf{x},\mathbf{x}') = \mathcal{A}_p^{\mathbf{x}'} k_{\mathbf{x}}(\mathbf{x}') = k_{\mathbf{x}}(\mathbf{x}')\nabla_{\mathbf{x}'} \log p(\mathbf{x}') + \nabla_{\mathbf{x}'} k_{\mathbf{x}}(\mathbf{x}') \in \mathbb{R}^d$, and $v_{\mathbf{x}'}(\mathbf{x})$ is also in the Stein class, thus Lemma 2 is applicable to equation 23, and we can have:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top v_{\mathbf{x}'}(\mathbf{x})\right] && (24)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathrm{trace}\left(\mathcal{A}_p^{\mathbf{x}} v_{\mathbf{x}'}(\mathbf{x})\right)\right] && (25)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathrm{trace}\left(\mathcal{A}_p^{\mathbf{x}}\mathcal{A}_p^{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\right)\right] && (26)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathrm{trace}\left(\nabla_{\mathbf{x}} \log p(\mathbf{x}) v_{\mathbf{x}'}(\mathbf{x})^\top + \nabla_{\mathbf{x}} v_{\mathbf{x}'}(\mathbf{x})\right)\right] && (27)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathrm{trace}\left(\nabla_{\mathbf{x}} \log p(\mathbf{x})^\top v_{\mathbf{x}'}(\mathbf{x})\right) + \mathrm{trace}\left(\nabla_{\mathbf{x}} v_{\mathbf{x}'}(\mathbf{x})\right)\right], && (28)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[s_p(\mathbf{x})^\top k(\mathbf{x},\mathbf{x}') s_p(\mathbf{x}') + s_p(\mathbf{x})^\top \nabla_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}') + \mathrm{trace}\left(\nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{x}') s_p(\mathbf{x}')^\top\right) + \mathrm{trace}\left(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\right)\right] && (29)\\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[s_p(\mathbf{x})^\top k(\mathbf{x},\mathbf{x}') s_p(\mathbf{x}') + s_p(\mathbf{x})^\top \nabla_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}') + s_p(\mathbf{x}')^\top \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{x}') + \mathrm{trace}\left(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\right)\right], && (30)
\end{aligned}$$

now the intractable $s_q(\mathbf{x})$ terms are removed from the formulation of KSD.

2 Stein Variational Gradient Descent

2.1 Multi-dimensional KSD

In the following, we will consider data points take values in $\mathcal{X} \subset \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$. We can apply the Stein identity in equation 1 again by taking $\boldsymbol{\phi}(\mathbf{x})$ as the $f(\mathbf{x})$, a tiny difference is now $\mathbf{x} \in \mathbb{R}^d$ and $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \cdots, \phi_d(\mathbf{x})]^\top$ are both d-dimensional vectors, and $\mathcal{A}_p\boldsymbol{\phi}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})\nabla_{\mathbf{x}} \log p(\mathbf{x})^\top + \nabla_{\mathbf{x}}\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{d\times d}$. We will also use $\mathcal{H}^d$ to denote the space of vector functions $\boldsymbol{f} = [f_1, \cdots, f_d]$ with $f_d \in \mathcal{H}$, whose inner product is given by $\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle f_i, g_i\rangle_{\mathcal{H}}$. And the Stein discrepancy which searches the $\boldsymbol{\phi}$ in the RKHS $\mathcal{H}^d$ is given by:

$$\sqrt{S(q,p)} = \max_{\boldsymbol{\phi}\in\mathcal{H}^d} \{\mathbb{E}_{\mathbf{x}\sim q}\left[\mathrm{trace}\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{x})\right)\right] \qquad s.t. \qquad \|\boldsymbol{\phi}\|_{\mathcal{H}^d} \le 1\}, \tag{31}$$

and the objective of equation 31 can be further written as:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{x})}\left[\operatorname{trace}\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{x})\right)\right] && (32)\\
=&\ \mathbb{E}_{q(\mathbf{x})}\left[\operatorname{trace}\left(\boldsymbol{\phi}(\mathbf{x})\nabla_{\mathbf{x}}\log p(\mathbf{x})^\top\right)+\operatorname{trace}\left(\nabla_{\mathbf{x}}\boldsymbol{\phi}(\mathbf{x})\right)\right] && (33)\\
=&\ \mathbb{E}_{q(\mathbf{x})}\left[\sum_{i=1}^{d}\left(\frac{\partial}{\partial\mathbf{x}_i}\phi_i(\mathbf{x})+\frac{\partial}{\partial\mathbf{x}_i}\log p(\mathbf{x})\phi_i(\mathbf{x})\right)\right], && (34)
\end{aligned}$$

and since every $\phi_i(\cdot)$ comes from the RKHS with reproducing kernel $k(\cdot,\cdot)$, by the reproducing property we can have:

$$\phi_i(\mathbf{x})=\langle\phi_i(\cdot),k(\mathbf{x},\cdot)\rangle_{\mathcal{H}}, \tag{35}$$

$$\frac{\partial}{\partial\mathbf{x}_i}\phi_i(\mathbf{x})=\langle\phi_i(\cdot),\frac{\partial}{\partial\mathbf{x}_i}k(\mathbf{x},\cdot)\rangle_{\mathcal{H}}, \tag{36}$$

thus equation 34 can be further derived as:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{x})}\left[\sum_{i=1}^{d}\left(\frac{\partial}{\partial\mathbf{x}_i}\phi_i(\mathbf{x})+\frac{\partial}{\partial\mathbf{x}_i}\log p(\mathbf{x})\phi_i(\mathbf{x})\right)\right] && (37)\\
=&\sum_{i=1}^{d}\langle\phi_i(\cdot),\mathbb{E}_{q(\mathbf{x})}\left[\frac{\partial}{\partial\mathbf{x}_i}\log p(\mathbf{x})k(\mathbf{x},\cdot)+\frac{\partial}{\partial\mathbf{x}_i}k(\mathbf{x},\cdot)\right]\rangle_{\mathcal{H}}, && (38)
\end{aligned}$$

the optimal unnormalized $\tilde{\boldsymbol{\phi}}(\cdot)$ is given by simply setting its i-th entry to $\mathbb{E}_{q(\mathbf{x})}\left[\frac{\partial}{\partial\mathbf{x}_i}\log p(\mathbf{x})k(\mathbf{x},\cdot)+\frac{\partial}{\partial\mathbf{x}_i}k(\mathbf{x},\cdot)\right]$, which means $\tilde{\boldsymbol{\phi}}^*(\cdot)=\mathbb{E}_{q(\mathbf{x})}\left[\mathcal{A}_pk(\mathbf{x},\cdot)\right]$ (note that $\mathcal{A}_pk(\mathbf{x},\cdot)\in\mathbb{R}^d$) and $\boldsymbol{\phi}^*(\mathbf{x})=\tilde{\boldsymbol{\phi}}^*(\mathbf{x})/\|\tilde{\boldsymbol{\phi}}^*(\cdot)\|_{\mathcal{H}^d}$.

2.2 Variational inference with smooth transforms

The general idea of Stein Variational Gradient Descent (SVGD) [Liu and Wang, 2016] is incrementally transforming a set of data points $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i\in\mathbb{R}^d$ sampled from a known initial distribution $q(\mathbf{x})$ to approximate a target distribution $p(\mathbf{x})=\tilde{p}(\mathbf{x})/Z$ which may be unnormalized. The transformation is in the form of: $\boldsymbol{T}(\mathbf{x})=\mathbf{x}+\epsilon\boldsymbol{\phi}(\mathbf{x})$, where $\boldsymbol{\phi}(\mathbf{x})\in\mathbb{R}^d$ is a smooth function that characterizes the direction and the scalar ϵ represents the magnitude.

Denote $q_{[\boldsymbol{T}]}$ as the density of the transformed points, when $|\epsilon|$ is sufficiently small, $\boldsymbol{T}$ is guranteed to be invertible, and denote $\mathbf{z}=\boldsymbol{T}(\mathbf{x})$, we have:

$$q_{[\boldsymbol{T}]}(\mathbf{z})=q(\boldsymbol{T}^{-1}(\mathbf{z}))\left|\det\left(J_{\boldsymbol{T}}^{-1}(\mathbf{z})\right)\right|. \tag{39}$$

SVGD proposes to use $q_{[\boldsymbol{T}]}(\mathbf{z})$ to do variational inference by updating the particles to get close to $p(\mathbf{x})$ in terms of KL divergence. And there is a surprising connection between *Stein operator* and the derivative of KL divergence w.r.t. the perturbation magnitude ϵ:

$$\begin{aligned}
&\nabla_\epsilon D_{\mathrm{KL}}\left(q_{[\boldsymbol{T}]}\parallel p\right)\big|_{\epsilon=0} && (40)\\
=&\ \nabla_\epsilon D_{\mathrm{KL}}\left(q\parallel p_{[\boldsymbol{T}^{-1}]}\right)\big|_{\epsilon=0} && (41)\\
=&\ \mathbb{E}_{\mathbf{x}\sim q}\left[-\nabla_\epsilon\log p_{[\boldsymbol{T}^{-1}]}(\mathbf{x})\right]\big|_{\epsilon=0} && (42)\\
=&\ \mathbb{E}_{\mathbf{x}\sim q}\left[-\nabla_\epsilon\left(\log p\left(\boldsymbol{T}_\epsilon(\mathbf{x})\right)+\log|\det J_{\boldsymbol{T}}(\mathbf{x})|\right)\right]\big|_{\epsilon=0} && (43)\\
=&\ -\mathbb{E}_{\mathbf{x}\sim q}\left[s_p(\boldsymbol{T}_\epsilon(\mathbf{x}))^\top\nabla_\epsilon\boldsymbol{T}_\epsilon(\mathbf{x})+\operatorname{trace}\left(J_{\boldsymbol{T}}(\mathbf{x})^{-1}\nabla_\epsilon J_{\boldsymbol{T}}(\mathbf{x})\right)\right]\big|_{\epsilon=0} && (44)\\
=&\ -\mathbb{E}_{\mathbf{x}\sim q}\left[s_p(\mathbf{x})^\top\boldsymbol{\phi}(\mathbf{x})+\operatorname{trace}\left(\boldsymbol{I}\nabla_{\mathbf{x}}\boldsymbol{\phi}(\mathbf{x})\right)\right] && (45)\\
=&\ -\mathbb{E}_{\mathbf{x}\sim q}\left[\operatorname{trace}\left(\mathcal{A}_p\boldsymbol{\phi}(\mathbf{x})\right)\right]. && (46)
\end{aligned}$$

We can see it is equivalent to the objective in equation 31, and when we consider $\phi(\cdot)$ in the unit ball of $\mathcal{H}^d$, the optimal direction that gives **the steepest descent on the KL divergence** has a closed form solution as $\phi^*_{q,p}(\cdot) = \beta_{q,p}(\cdot) = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p k(\mathbf{x}, \cdot)\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[\nabla_{\mathbf{x}} \log p(\mathbf{x}) k(\mathbf{x}, \cdot) + \nabla_{\mathbf{x}} k(\mathbf{x}, \cdot)\right]$, this is computationally tractable.

3 Amortizd SVGD

"SVGD and other particle based methods become ineficient when we need to apply them repeatedly on a large number of different, but similar target distributions for multiple tasks, because they can not leverage the similarity between the different distributions and may require a large memory to restore a large number of particles."

References

Q. Liu. A short introduction to kernelized stein discrepancy, 2016.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

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

D. Sejdinovic and A. Gretton. What is an rkhs? *Lecture Notes*, 2012.

A The reproducing property

Refer to [Sejdinovic and Gretton, 2012].

B Lemmas

Lemma 1 (First half of Lemma 2.3 of [Liu et al., 2016])**.** *Assume $p(\mathbf{x})$ and $q(\mathbf{x})$ are smooth densities supported on $\mathcal{X}$ and* **scalar-valued** *function $f(\mathbf{x})$ is in the Stein class of q, we have:*

$$\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p f(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))f(\mathbf{x})\right].$$

Lemma 2 (Second half of Lemma 2.3 of [Liu et al., 2016])**.** *Assume $p(\mathbf{x})$ and $q(\mathbf{x})$ are smooth densities supported on $\mathcal{X}$ and when $\boldsymbol{f}(\mathbf{x})$ is a $d \times 1$* **vector-valued** *function in the Stein class of q, we have:*

$$\mathbb{E}_{\mathbf{x}\sim q}\left[(s_p(\mathbf{x}) - s_q(\mathbf{x}))^\top \boldsymbol{f}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathrm{trace}\left(\mathcal{A}_p \boldsymbol{f}(\mathbf{x})\right)\right].$$

C Introduction to measure theory

- Limit of a sequence: a sequence $x_1, x_2, \cdots, x_n$ is said to converge to x or have limit if ...
- Cauchy sequence

- Algebraic structure
- measure space: $(\mathcal{X}, \mathcal{A}, \mu)$, where $\mathcal{X}$ is a set, $\mathcal{A}$ is a class of subsets of $\mathcal{X}$, and μ is a function that attach a nonnegative number to every set in $\mathcal{A}$.
- σ-algebra: $\mathcal{A}$ is call a σ-field of $\mathcal{X}$ if:
  - both $\emptyset$ and $\mathcal{X}$ in $\mathcal{A}$
  - if A in $\mathcal{A}$, then A^c in $\mathcal{A}$
  - if $A_1, \cdots, A_n$ is a countable collection of sets in $\mathcal{A}$, then both $\cup_i A_i$ and $\cap_i A_i$ in $\mathcal{A}$
- measure: a function μ defined on $\mathcal{A}$ is called a (countably additive, nonnegative) measure if: (1) (2) (3)
- $(\Omega, \mathcal{F}, \mathbb{P})$ used to denote a probability space
- countable additive
- metric space, complete metric space, normed space
- inner product on a vector space
- Hilbert space: a vector space where inner product is defined, and contains all the limits of Cauchy sequences of functions
- kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if exists a $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \forall x, x' \in \mathcal{X}$