The contents comes from the paper
"A kernelized Stein Discrepancy for Goodness-of-fit Test"
Some notations are different (e.g. p and q are switched when they co-appear)

• Stein's class

1° let $\begin{cases} \mathcal{X} \subseteq \mathbb{R}^d \\ p(x) \end{cases}$ be a continuously differentiable distribution supported on $\mathcal{X}$

a function $f: \mathcal{X} \mapsto \mathbb{R}$ is in the Stein's class of p if

and $\begin{cases} f \text{ is smooth} \\ \int_{x\in\mathcal{X}} \nabla_x (p(x) f(x))\, dx = 0 \end{cases}$

2° $f: \mathcal{X} \mapsto \mathbb{R}^m$ is in the Stein's class of p if f_i is $\forall i = 1, \cdots, m$

3° f can be shown to be in Stein's class of p if

1. $\lim_{\|x\| \to \infty} f(x) p(x) = 0$ eg. when $p(x)$ is bounded and $\lim_{\|x\| \to \infty} f(x) = 0$

or

2. $\begin{cases} \mathcal{X} \text{ is compact with piecewise continuous boundary} \\ p(x) f(x) = 0 \quad \forall x \in \partial\mathcal{X} \end{cases}$

⋮

• Stein's Identity

1° let p be a smooth distribution on $\mathcal{X} \subseteq \mathbb{R}^d$, $f: \mathbb{R}^d \mapsto \mathbb{R}^m$

the Stein's operator of p is a linear operator

$$(\mathcal{A}_p f)(x) = \nabla \log p(x)\, f(x)^T + D f(x)^T \qquad D f(x)^T = \begin{bmatrix} \nabla f_1 & \cdots & \nabla f_m \end{bmatrix}$$

2° Assume p is a smooth density supported on $\mathcal{X}$, f is in Stein's class of p

then $\mathbb{E}_{x\sim p}[(\mathcal{A}_p f)(x)] = \mathbb{E}_{x\sim p}[\nabla \log p(x) f(x)^T + D f(x)^T] = 0$

$$\begin{aligned} & \int_{\mathcal{X}} p(x) \nabla \log p(x)\, f_i(x) + p(x) \nabla f_i(x)\, dx \\ = & \int_{\mathcal{X}} \nabla p(x) f_i(x) + p(x) \nabla f_i(x)\, dx \\ = & \int_{\mathcal{X}} \nabla (p(x) f_i(x))\, dx \\ = & \, 0 \end{aligned}$$

3° Suppose p and q are distributions on $\mathcal{X}$
f is in Stein's class of q

then $E_{x\sim q}[(A_p f)(x)] = E_{x\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x)) f(x)^T]$

$$E_{x\sim q}[(A_p f)(x)] = E_{x\sim q}[(A_p f)(x) - (A_q f)(x)]$$
$$= E_{x\sim q}[\nabla \log p(x) f(x)^T - \nabla f(x)^T - \nabla \log q(x) f(x)^T + \nabla f(x)^T]$$
$$= E_{x\sim q}[(\nabla \log p(x) - \nabla \log q(x)) f(x)^T]$$

Also $E_{x\sim q}[\mathrm{trace}((A_p f)(x))] = \underbrace{E_{x\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x))^T f(x)]}_{\text{Expected difference in score function, measured in } f \text{ direction}}$ (when $f: \mathbb{R}^d \mapsto \mathbb{R}^d$)

4° $p \neq q \Rightarrow \exists f$ s.t. $E_{x\sim q}[(A_p f)(x)] \neq 0$

· Stein Discrepency

1° define the Stein's Discrepency between p and q
$$\sqrt{S(q,p)} = \max_{f\in\mathcal{F}} E_{x\sim q}[\mathrm{trace}((A_p f)(x))]$$

e.g. $f_w(x) = \sum_i w_i f_i(x)$
f is a linear combination of simple f_i's

$$E_{x\sim q}[\mathrm{trace}((A_p f)(x))] = E_{x\sim q}[(\nabla \log p(x) - \nabla \log q(x))^T f(x)]$$
$$= \sum_i w_i \underbrace{E_{x\sim q}[(\nabla \log p(x) - \nabla \log q(x))^T f_i(x)]}_{\beta_i}$$

$\mathcal{F} = \{\sum_i w_i f_i(x) : \|w\| \leq 1\}$

then $\sqrt{S(q,p)} = \max_w w^T \beta$ s.t. $\|w\| \leq 1$

· Kernelized Stein Discrepancy

1° A kernel $k(x,x')$ is integrally strictly positive definite if

$\forall\ g$ s.t. $0 < \|g\|_{L^2}^2 < \infty$

we have $\int_{\mathcal{X}} \int_{\mathcal{X}} g(x)\, k(x,x')\, g(x')\, dx\, dx' > 0$

think this as a continuous version of $g^T k g = \sum_i \sum_j g_i k_{ij} g_j > 0$

2° the kernelized Stein discrepancy is defined as

$$S_k(q,p) = E_{x,x'\sim q}\left[\left(\nabla \log p(x) - \nabla \log q(x)\right)^T k(x,x') \left(\nabla \log p(x') - \nabla \log q(x')\right) \right]$$

3° let $g_{q,p}(x) = q(x)\left(\nabla \log p(x) - \nabla \log q(x)\right)$

assume $\begin{cases} k(x,x') \text{ is integrally strictly positive definite} \\ p\ q \text{ are continuous density s.t. } \|g_{q,p}\|_{L^2}^2 < \infty \end{cases}$

then $\begin{cases} S_k(q,p) \geq 0 \\ S_k(q,p) = 0 \Leftrightarrow p = q \end{cases}$

eg. satisfied when tail of q decays exponentially

eg. not satisfied when q has a heavy tail

(when q is a Cauchy distribution, p is a Gaussian)

if $p = q$, then $\nabla \log p(x) - \nabla \log q(x) = 0$, $g_{q,p}(x) = 0$, then $S_k(q,p) = 0$ obviously

if $S_k(q,p) = 0$

$$\begin{aligned} S_k(q,p) &= E_{x,x'\sim q}\left[\left(\nabla \log p(x) - \nabla \log q(x)\right)^T k(x,x') \left(\nabla \log p(x') - \nabla \log q(x')\right) \right] \\ &= \iint q(x) \left(\nabla \log p(x) - \nabla \log q(x)\right)^T k(x,x') \left(\nabla \log p(x') - \nabla \log q(x')\right) q(x')\, dx\, dx' \\ &= \iint g_{q,p}(x)^T k(x,x')\, g_{q,p}(x')\, dx\, dx' \\ &= 0 \end{aligned}$$

then $\|g_{q,p}\|_{L^2}^2 = 0$, then $p = q$

4° A kernel $k(x,x')$ is in the Stein class of p if

and $\begin{cases} k \text{ has continuous 2nd partial derivatives} \\ \text{both } k(x,\cdot) \text{ and } k(\cdot,x) \text{ are in the Stein class of } p\ \forall x \end{cases}$

eg. RBF kernel is in the Stein class for smooth densities supported on $\mathbb{R}^n$

$\forall u$, $\lim_{\|x\| \to \infty} k(x,u) p(x) = 0$ since $\begin{cases} k(x,u) \text{ bounded } \forall u \\ \lim_{\|x\|\to\infty} p(x) = 0 \end{cases}$

5° if $k(\cdot,x'): \mathbb{R}^d \to \mathbb{R}$ is in the Stein class of p $\forall x'$

can show $\nabla_{x'} k(\cdot,x'): \mathbb{R}^d \to \mathbb{R}^d$ is also in the Stein class

$$\int_x \nabla_x \left(p(x) \nabla_{x'} k(x,x') \right) dx$$
$$= \int_x \nabla p(x) \nabla_{x'} k(x,x')^T + p(x) \nabla_{xx'} k(x,x') \, dx$$
$$= \int_x \nabla_{x'} \left[\nabla p(x) k(x,x') + p(x) \nabla_x k(x,x') \right] dx$$
$$= \nabla_{x'} \int_x \nabla_x \left[p(x) k_{x'}(x) \right] dx$$
$$= \nabla_{x'} 0 = 0$$

6° Assume $\begin{cases} p, q \text{ are smooth densities} \\ k(x,x') \text{ is in the Stein's class of } q \end{cases}$

then $S_k(q,p) = E_{x,x'\sim q}\left[\nabla \log p(x)^T k(x,x') \nabla \log p(x') + \nabla \log p(x)^T \nabla_{x'} k(x,x') + \nabla_x k(x,x')^T S_p(x') + \mathrm{trace}(\nabla_{xx'} k(x,x')) \right]$

evaluate $S_k(q,p)$ in a way st only $\nabla \log p(x)$ is need, no $\nabla \log q(x)$

$$S_k(q,p) = E_{x,x'\sim q}\left[(\nabla \log p(x) - \nabla \log q(x))^T k(x,x') (\nabla \log p(x') - \nabla \log q(x')) \right]$$
$$= E_{x,x'\sim q}\left[(\nabla \log p(x) - \nabla \log q(x))^T \left(\nabla \log p(x') k(x,x') + \nabla_{x'} k(x,x') - \nabla \log q(x') k(x,x') - \nabla_{x'} k(x,x') \right) \right]$$

use the notation $k(x,\cdot) = k_x(\cdot)$

$$= E_{x,x'\sim q}\left[(\nabla \log p(x) - \nabla \log q(x))^T \left((A_p k_x)(x') - (A_q k_x)(x') \right) \right]$$
$$= E_{x,x'\sim q}\left[(\nabla \log p(x) - \nabla \log q(x))^T (A_p k_x)(x') \right]$$

$E_{x'\sim q}\left[(A_q k_x)(x') \right] = 0$

$$= E_{x,x'\sim q}\left[\nabla \log p(x)^T \left(\nabla \log p(x') k(x,x') + \nabla_{x'} k(x,x') \right) \right] - E_{x,x'\sim q}\left[\nabla \log q(x)^T \left(\nabla \log p(x') k(x,x') + \nabla_{x'} k(x,x') \right) \right]$$

$$= \underbrace{E_{x,x'\sim q}\left[\nabla \log p(x)^T k(x,x') \nabla \log p(x') \right] + E_{x,x'\sim q}\left[\nabla \log p(x)^T \nabla_{x'} k(x,x') \right]}_{①} - E_{x,x'\sim q}\left[\nabla \log q(x)^T k(x,x') \nabla \log p(x') \right] - E_{x,x'\sim q}\left[\nabla \log q(x)^T \nabla_{x'} k(x,x') \right]$$

$$= ① - E_{x,x'\sim q}\left[\left(\nabla \log q(x)^T k(x,x') + \nabla_x k(x,x')^T - \nabla_x k(x,x')^T \right) \nabla \log p(x') \right] + \underbrace{E_{x,x'\sim q}\left[\mathrm{trace}(\nabla_{xx'} k(x,x')) \right]}$$

$\nabla_{x'} k(\cdot,x')$ in Stein class of q

$$= ① + E_{x,x'\sim q}\left[\nabla_x k(x,x')^T \nabla \log p(x') \right] + E_{x,x'\sim q}\left[\mathrm{trace}(\nabla_{xx'} k(x,x')) \right]$$

$$= E_{x,x'\sim q}\left[\begin{array}{l} \nabla \log p(x)^T k(x,x') \nabla \log p(x') \\ + \nabla \log p(x)^T \nabla_{x'} k(x,x') + \nabla_x k(x,x')^T \nabla \log p(x') \\ + \mathrm{trace}(\nabla_{xx'} k(x,x')) \end{array} \right]$$

· RKHS Interpretation

1° if $k(x,x')$ is in the Stein class of p, $\mathcal{H}$ is the RKHS induced by k
then $\forall f \in \mathcal{H}$, f also satisfies $E_{x\sim p}[(A_p f)(x)] = 0$

if $f \in \mathcal{H}$, then $\nabla f(x) = \langle f, \nabla_x k(x,\cdot)\rangle_{\mathcal{H}}$

refer to theorem 1(b) in the paper

"Derive reproducing property for kernel method in machine learning"

$\forall f \in \mathcal{H}$

$$
\begin{aligned}
E_{x\sim p}[(A_p f)(x)] &= E_{x\sim p}[\nabla \log p(x) f(x) + \nabla f(x)] \\
&= E_{x\sim p}[\nabla \log p(x) \langle f, k(x,\cdot)\rangle_{\mathcal{H}} + \langle f, \nabla_x k(x,\cdot)\rangle_{\mathcal{H}}] \\
&= \langle f, E_{x\sim p}[\nabla \log p(x) k(x,\cdot) + \nabla_x k(x,\cdot)] \rangle_{\mathcal{H}} \\
&= \langle f, E_{x\sim p}[A_p k_x] \rangle_{\mathcal{H}} \\
&= 0
\end{aligned}
$$

A more clear notation, let $f = \sum_i \alpha_i k(x_i,\cdot)$

$$
\begin{aligned}
& E_{x\sim p}[\nabla \log p(x) f(x) + \nabla f(x)] \\
&= E_{x\sim p}[\nabla \log p(x) \textstyle\sum_i \alpha_i k(x,x_i) + \sum_i \alpha_i \nabla k(x,x_i)] \\
&= \textstyle\sum_i \alpha_i E_{x\sim p}[\nabla \log p(x) k_{x_i}(x) + \nabla k_{x_i}(x)] \\
&= 0
\end{aligned}
$$

2° let k be a kernel in Stein's class of q, $\mathcal{H}$ be the RKHS induced by k

let $\beta(x') = E_{x\sim q}[(A_p k_{x'})(x)]$, then $S_k(q,p) = \|\beta\|^2_{\mathcal{H}^d}$

$$
\begin{cases}
\beta(x') = E_{x\sim q}[\nabla \log p(x) k(x,x') + \nabla_x k(x,x')] \\
\beta_i(x') = E_{x\sim q}[\frac{\partial}{\partial x_i} \log p(x) k(x,x') + \frac{\partial}{\partial x_i} k(x,x')]
\end{cases}
$$

$$\beta(x') = E_{x\sim q}[(\nabla \log p(x) - \nabla \log q(x)) k(x,x')]$$

$$
\begin{aligned}
S_k(q,p) &= E_{x,x'\sim q}[(\nabla \log p(x) - \nabla \log q(x))^T k(x,x') (\nabla \log p(x') - \nabla \log q(x'))] \\
&= E_{x,x'\sim q}[(\nabla \log p(x) - \nabla \log q(x))^T \langle k(x,\cdot), k(\cdot,x')\rangle_{\mathcal{H}} (\nabla \log p(x') - \nabla \log q(x'))] \\
&= \sum_{i=1}^d E_{x,x'\sim q}[(\tfrac{\partial}{\partial x_i} \log p(x) - \tfrac{\partial}{\partial x_i} \log q(x)) \langle k(x,\cdot), k(\cdot,x')\rangle_{\mathcal{H}} (\tfrac{\partial}{\partial x'_i} \log p(x') - \tfrac{\partial}{\partial x'_i} \log q(x'))] \\
&= \sum_{i=1}^d \langle E_{x,x'\sim q}[(\tfrac{\partial}{\partial x_i} \log p(x) - \tfrac{\partial}{\partial x_i} \log q(x)) k(x,\cdot)], E_{x,x'\sim q}[(\tfrac{\partial}{\partial x'_i} \log p(x') - \tfrac{\partial}{\partial x'_i} \log q(x')) k(\cdot,x')] \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^d \langle \beta_i, \beta_i \rangle_{\mathcal{H}} \\
&= \|\beta\|^2_{\mathcal{H}^d}
\end{aligned}
$$

3° $\langle f, \beta \rangle_{\mathcal{H}^d} = E_{x\sim q}\left[\operatorname{trace}((\mathcal{A}_p f)(x))\right] \quad \forall f \in \mathcal{H}^d$

$$
\begin{aligned}
E_{x\sim q}\left[\operatorname{trace}((\mathcal{A}_p f)(x))\right] &= E_{x\sim q}\left[\operatorname{trace}\left(\nabla \log p(x) f(x)^T + \nabla f(x)^T\right)\right] \\
&= \sum_{i=1}^{d} E_{x\sim q}\left[\frac{\partial}{\partial x_i}\log p(x)\, f_i(x) + \frac{\partial f_i(x)}{\partial x_i}\right] \\
&= \sum_{i=1}^{d} E_{x\sim q}\left[\frac{\partial}{\partial x_i}\log p(x) \langle f_i, k(x,\cdot)\rangle_{\mathcal{H}} + \langle f_i, \frac{\partial}{\partial x_i} k(x,\cdot)\rangle_{\mathcal{H}}\right] \\
&= \sum_{i=1}^{d} \langle f_i, E_{x\sim q}\left[\frac{\partial}{\partial x_i}\log p(x)\, k(x,\cdot) + \frac{\partial}{\partial x_i} k(x,\cdot)\right]\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{d} \langle f_i, \beta_i \rangle_{\mathcal{H}} \\
&= \langle f, \beta \rangle_{\mathcal{H}^d}
\end{aligned}
$$

4° $\beta(x') = E_{x\sim q}\left[(\mathcal{A}_p k_{x'})(x)\right]$

$S_k(q,p) = \|\beta\|_{\mathcal{H}^d}^2$

$E_{x\sim q}\left[(\mathcal{A}_p f)(x)\right] = \langle f, \beta \rangle_{\mathcal{H}^d} \quad \forall f \in \mathcal{H}^d$

then $\sqrt{S_k(q,p)} = \|\beta\|_{\mathcal{H}^d} = \max\left\{\langle f, \beta\rangle_{\mathcal{H}^d} : \|f\|_{\mathcal{H}^d} \le 1\right\}$

$$= \max\left\{E_{x\sim q}\left[(\mathcal{A}_p f)(x)\right] : \|f\|_{\mathcal{H}^d} \le 1\right\}$$