Kernelized Stein Discrepancy.

• Stein Variational Gradient Descent.

(Oct 27th) For $E(\rho) = \int \rho \ln\left(\frac{\rho}{\sigma}\right) dx$.

$$R(v) = \frac{1}{2}\|v\|_H^2 + \int v \cdot (\nabla g + g \nabla U)\, dx \qquad (\sigma = e^{-U})$$

minimize v $\longrightarrow$ $\partial_t \rho = \mathrm{div}(\rho K * (\nabla \rho + \rho \nabla U))$

$(-V)$

For such E,

$$\frac{dE}{dt}(\rho) = \int v \cdot (\nabla \rho + \rho \nabla U)\, dx$$

$$= -\|v\|_H^2$$

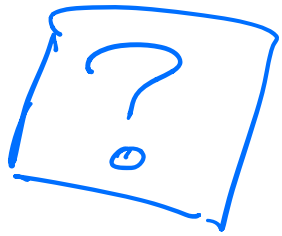
$$\overset{(*)}{=} -D(\rho \| \sigma)^2$$

Here $V(x) = \int \nabla \ln \sigma(y)\, k(x,y) + \nabla_y k(x,y)\, d\rho(y)$

$$D(\rho\|\sigma) = \max_{\phi \in \mathcal{H},\, \|\phi\|_{\mathcal{H}} \le 1} \{ E_\rho(A_\sigma(\phi)) \}$$

$$A_g(\phi) := \nabla \ln \sigma \cdot \phi + \nabla \cdot \phi$$

"K" Stein Discrepancy

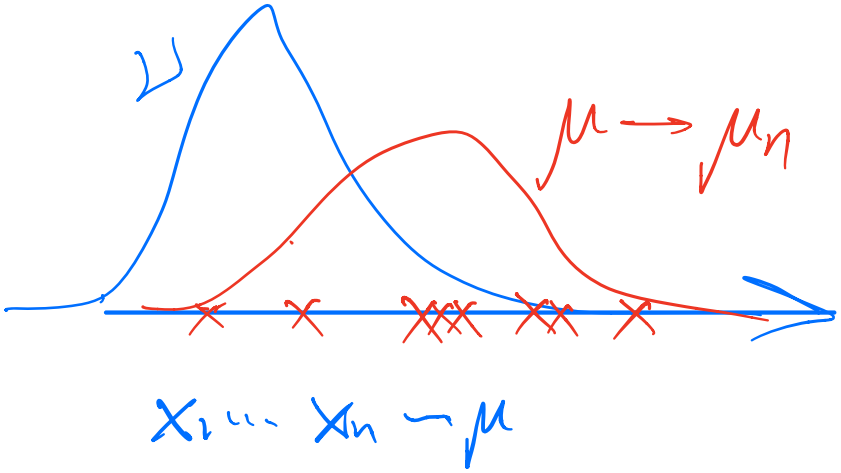


[A KSD for Goodness-of-fit Tests]

(Liu, Lee, Jordan, ICML16')

- "Goodness-of-fit Test"

Measuring how well do the observed data correspond to the fitted model.



$H_0: \mu = \nu$

$H_1: \mu \neq \nu$

ν known.

("Two-sample test":

$\begin{cases} x_1 \cdots x_n \sim \mu \\ y_1 \cdots y_n \sim \nu \end{cases} \Rightarrow \begin{cases} H_0 : \mu = \nu \\ H_1 : \mu \neq \nu \end{cases}$.)

— likelihood-based approaches

(Need to compute likelihoods & CDF)

hidden variable models & generative models ✗

L MCMC; variational methods

error large & hard to estimate

(Curse of dimensionality)

— Traditional methods (χ^2, Kolmogorov-Smirnov etc)

Today: Evaluating KSD.

A "likelihood-free" approach

with guaranteed statistical significance.

(i). Preliminaries, definition of KSD

(ii). Properties of KSD

(iii). Original SVGD approach.

(iv). Open problems.

(i) >>

• Stein's method. (1970s)

for distributional approximation.

"Probability metrics", "MMD, ..."

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{h\in\mathcal{H}} \left| \int h\,d\mu - \int h\,d\nu \right|$$

Test function family $\mathcal{H}$: rich enough?

$\mathbb{1}_{\{\cdot \le x\}}$ / Lip / $\mathbb{1}_{\{\cdot \in \text{Borel } A\}}$ ...

d_K d_W d_{TV}

Q. Given $\{X_i\}_{i=1}^n$ independent variables,

How to bound $d_{\mathcal{H}}\left(\frac{1}{\sqrt{n}}\sum X_i, N(0,1)\right)$?

(CLT: important cvg rate.)

Main idea: Replace the characteristic function typically used to show cvg in distribution with a "characteristic operator".

Define $\mathcal{A}: C'(\mathbb{R}) \longrightarrow C(\mathbb{R})$ by

$$\mathcal{A}f = f' - xf$$

Let Φ be the cdf of $N(0,1)$, then

$\exists!$ f_x solves $f' - xf = \mathbb{1}_{\{\cdot \le x\}} - \Phi(x)$

for $\forall$ fixed $x \in \mathbb{R}$. (ODE result)

$\Rightarrow$ [Stein lemma]

A r.v. $W \sim N(0,1)$ $\Leftrightarrow$

$\forall$ a.c. f with $E|f'(W)| < \infty$, $E\mathcal{A}f(W) = 0$

$\Rightarrow$ (Cor 1) For $\forall$ r.v. W,

$$|P(W \le x) - \Phi(x)| = |E(f_x'(W) - W f_x(W))|.$$

(dk)

General Setup:

$$d_{\mathcal{H}}(\mu,\nu) = \sup_{h \in \mathcal{H}} \left| \int h\,d\mu - \int h\,d\nu \right|$$

For given $h \in \mathcal{H}$, let f_h solves

$f_h'(w) - w f_h(w) = h(w) - \Phi(h)$, where

$\Phi(h) = E(h(Z))$ for $Z \sim N(0,1)$.

Then $d_{\mathcal{H}}(\mu,\nu) = \sup_{h \in \mathcal{H}} |E f_h'(W) - W f_h(W)|$

for $W \sim d\mu$.

[Stein's Normal Approximation]

Once we solve and restrict our discussion on $\mathcal{F} = \{f \mid \|f\|_\infty, \|f'\|_\infty \le 2, \|f'\|_\infty \le \sqrt{2/\pi}\}$,

we have

$$d_W(W, Z) \le \sup_{f \in \mathcal{F}} |E f'(W) - W f(W)|$$

(Cor 2) For 0-mean independent $\{X_i\}_{i=1}^n$, $W = \Sigma X_i/\sqrt{n}$, we have

$$d_W(W, Z) \le \frac{1}{n^{3/2}} \Sigma E|X_i|^3 + \sqrt{\frac{2}{\pi n^2} \Sigma E|X_i|^4}$$

(N. Ross)

General Setup.

What if we are approximating a probability distribution with smooth density q? Replace $-x$ by q'/q.

On $\mathcal{X} \subseteq \mathbb{R}^d$, 2 smooth densities

p, q are "identical" $\Leftrightarrow$

$$E_p(A_q(f)) = E_p(s_q(x) f(x) + \nabla_x f(x)) = 0.$$

Here $s_q = \nabla_x \ln q(x)$ is the "Stein score function".

linear operator A_q is called the "Stein's operator", acts on the "Stein class of q", i.e. $f \in C^1(\mathcal{X})$

$$\int_{\mathcal{X}} \nabla_x (f(x) p(x)) \, dx = 0$$

For $f = [f_1, \cdots f_{d'}]$, $A_p f \in \mathbb{R}^{d \times d'}$.

(Gorham & Mackey 15')

$$S(p,q) = \sup_{f \in \mathcal{F}} \left(E_p(A_q f)\right)^2$$

— requires a difficult variational optimization.

So...

[RKHS Settings]

$\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$, $\mathcal{H} \sim k$.

RKHS for vector-valued functions

k is strictly p.d.

Then in (Liu, Lee, Jordan),

(Definition) For $p, q \in \mathcal{P}(\mathbb{R}^d)$.

$$KSD_k(p,q) := S(p,q)$$

$$= \mathbb{E}_{x,x' \sim p}\left[(s_q - s_p)^T(x)\, k(x,x')\, (s_q - s_p)(x') \right]$$

"Score difference" pd.

$p(s_q - s_p) \in L^2(\mathcal{X})$. $KSD(p,q) = 0 \iff p = q$.

(ii). Characterization & properties.

[Thm 3.6].

$$S(p,q) = \mathbb{E}_{x,x' \sim p}\left(u_q(x,x')\right)$$

with kernel $u_q(x,x') = s_q(x)^T k(x,x') s_q(x') +$

$s_q(x)^T \nabla_{x'} k(x,x') + \nabla_x k(x,x')^T s_q(x') + \mathrm{tr}\left(\nabla_{x,x'} k(x,x')\right)$.

Not symmetric w.r.t. (p,q)!

A "special" kernelized MMD. (later).

[Proof]. Notice $\mathbb{E}_p(A_q f) = \mathbb{E}_p((s_q - s_p) f^T)$.

i.e. $\mathbb{E}_p(\mathrm{tr}(A_q f)) = \mathbb{E}_p((s_q - s_p)^T f)$

Apply it on $k(x,\cdot)$ (~~fixed~~ x), we have

$$S(p,q) = E_{x,x'\sim p}\left[(S_q - S_p)^T A_q k_x(x')\right]$$

Apply it again for ~~fixed~~ x', ✓ □

[Thm 3.8].

For $\beta(x') = E_p(A_q k_{x'}(x))$

$$S(p,q) = \|\beta\|_{\mathcal{H}^d}^2 = \max_{\substack{f\in\mathcal{H}^d \\ \|f\|_{\mathcal{H}^d}\le 1}} \left(E_p \operatorname{tr}(A_q f)\right)^2$$

[Proof].

$$S(p,q) = E_{x,x'\sim p}\left[(S_q(x)-S_p(x))^T k(x,x')(S_q(x')-S_p(x'))\right]$$

$$= E_{x,x'}\left[(S_q(x)-S_p(x))^T \langle k(x,\cdot), k(\cdot,x')\rangle (\cdots)\right]$$

$$= \sum_{l=1}^{d} \langle E_x\left[(S_q^l - S_p^l)k(x,\cdot)\right], E_{x'}\left[k(\cdot,x')(S_q^l - S_p^l)\right]\rangle$$

$= \|\beta\|^2_{\mathcal{H}^d}$, same trick as above ●

While

$$\langle f, \beta \rangle_{\mathcal{H}^d} = \sum_{l=1}^{d} \langle f_l, E_{xyp} A_q^l k(x, \cdot) \rangle_{\mathcal{H}}$$

$$= \sum_{l=1}^{d} \langle f_l, E_{xyp} [S_q^l(x) k(x, \cdot) + \nabla_{x_l} k(x, \cdot)] \rangle_{\mathcal{H}}$$

$$= \sum_{l=1}^{d} E_{xyp} [S_q^l(x) \langle f_l, k(x, \cdot) \rangle_{\mathcal{H}} + \langle f_l, \nabla_{x_l} k(x, \cdot) \rangle_{\mathcal{H}}]$$

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

$$= \sum_{l=1}^{d} E_{xyp} [S_q^l(x) f_l(x) + \nabla_{x_l} f_l(x)]$$

$$= E_p(\mathrm{tr}(A_q f)).$$

which finishes the proof. □

We are in a situation to give an estimation of $S(p,q)$:

From [Thm36], "U-statistics"

$$\hat{S}_u(p,q) = \frac{1}{n(n-1)} \sum_{i \neq j} u_q(x_i, x_j)$$

can be used in application (Bootstrap ....)

The connection with Fisher divergence & MMD.

1. Fisher $F(p,q) = E_p \| s_p - s_q \|^2$

KSD is a "kernelized" Fisher divergence.

$$\frac{1}{\|s_p - s_q\|_{L^2_p}^2} F^2(p,q) \leq S(p,q) \leq \sqrt{E_{x,x' \sim p} k(x,x')^2}\, F(p,q)$$

$\|\cdot\|_{\mathcal{H}^d}$

$$E_p[\mathrm{tr}(A_q f)] = E_p[(s_q - s_p)^T f] \quad \text{for } \|f\|_{\mathcal{H}^d} \le 1.$$

2. MMD kernelled version:

$$\sup_{f \in \mathcal{H}} \{E_p f - E_q f,\ \|f\|_{\mathcal{H}} \le 1\}$$

$$= E_{\substack{x,x' \sim p \\ y,y' \sim q}} [k(x,x') + k(y,y') - 2k(x,y')].$$

(Gretton 10')

Notice that $\forall x$,

$E_{y,y' \sim q}\, u_q(y,y') - 2u_q(x,y') = 0$.

from Stein identity. Thus KSD is a MMD with asymmetric kernel u_q.

MMD: 2 sample test

KSD: Goodness-of-fit test.

Some asymptotic results.

1

For $p \neq q$, "asymptotically normal"

$$\sqrt{n}\left(\hat{S}_u(p,q) - S(p,q)\right) \xrightarrow{d} N(0, \sigma_u^2)$$

where $\sigma_u^2 = \mathrm{Var}_{x \sim p} E_{x' \sim p} u_q(x, x') > 0$

For $p = q$, $\Rightarrow \sigma_u^2 = 0$

$$n\hat{S}_u(p,q) \xrightarrow{d} \sum_{j=1}^{\infty} q_j (z_j^2 - 1)$$

q_j: eigen. z_j: i.i.d Gaussian

All standard results of U-stat.

2

(Lu & Lu : A Universal Approximation Thm (2020) of DNN for expressing distributions)

$$MMD(\rho,\pi) := \sup_{\|f\|_{\mathcal{H}_\rho}\le 1} |E_\rho f - E_\pi f|$$

$$KSD(\rho,\pi) := \sup_{\|f\|_{\mathcal{H}_d}\le 1} E_\rho(\nabla \ln \pi \cdot f + \nabla \cdot f)$$

[Thm 4.1].

$\{X_i\}_{i=1}^n \sim \pi$, $\rho_n = \frac{1}{n}\sum \delta_{x_i}$. Then there exists realizations of ρ_n s.t. following inequalities hold.

(1). for $\pi \in P_3(\mathbb{R}^d)$.

$$W_1(\rho_n,\pi) \le \begin{cases} \frac{C}{\sqrt{n}} & d=1 \\ \frac{C\log n}{\sqrt{n}} & d=2 \\ C\cdot n^{-\frac{1}{d}} & d\ge 3 \end{cases}$$

(2). for ρ_d, bounded k,

$$MMD(\rho_n,\pi) \le \frac{C}{\sqrt{n}}$$

(3). for k has bdd derivatives &

sub-Gaussian smooth π with $\nabla \ln \pi$ L-Lip,

$$KSD(P_n, \pi) \le C\sqrt{\frac{d}{n}}.$$

(Or you can understand above results as:
"with prob. at least $1-\delta$, ($\le$) holds for $C = C(\delta \cdots)$ independent of n")

[Proof].

(1). (2). Well-known. Sketch of (2):

Here, $(K)MMD(P_n, \pi) = \left\| \int k(x, \cdot)\, d(P_n - \pi) \right\|_{\mathcal{H}}$

$=: \varphi(X_1, \cdots X_n)$.

(Sriperumbudur'16) or survey

φ satisfies $|\varphi(X_1, \cdots X_i, \cdots X_n) - \varphi(X_1, \cdots X_i', \cdots X_n)|$

$\le \frac{2\sqrt{K_0}}{n}$ ($K_0 = \sup k(x,x)$). Then from

McDiarmid's Ineq, with prob. $1 - e^{-\tau}$ (Hoeffding-type)

$$\left\| \int k(\cdot, x)\, d(P_n - \pi) \right\|_{\mathcal{H}} \le \underline{E(\cdots)} + \sqrt{\frac{2K_0\tau}{n}}.$$

By standard symmetrization argument.

$$\mathbb{E}\left\| \int k(\cdot,x)\, d(P_n - \pi) \right\|_{\mathcal{H}} \le 2 \mathbb{E}_{\{X_i\}} \mathbb{E}_{\{\varepsilon_i\}} \left\| \frac{1}{n} \sum_i \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}_k}$$

"Rademacher averages". i.i.d. $\varepsilon_i \sim \pm 1$

Use McDiarmid again. RHS

$$\le \mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}} + \sqrt{\frac{2 K_0 C}{n}}$$

$$\le \left(\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum \varepsilon_i k(\cdot, X_i) \right\|_{\mathcal{H}}^2 \right)^{\frac{1}{2}} + \searrow \le \frac{C(K_0, C)}{\sqrt{n}} \quad \square$$

(3) Sketch: $KSD^2(P_n, \pi) = \frac{1}{n^2} \sum_{ij} u_\pi(X_i, X_j)$

Use Bernstein-type ineq for V-statistics.

(u_π: Symmetric kernel)

Check this kernel u_π satisfies the conditions of (Thm C.1): degenerate;

$|u_\pi(x,y)| \le g(x) g(y)$: $\mathbb{E}|g(X)^k| \le \xi^2 J^{k-2} k!/2$;

$$\Rightarrow P(|KSD^2| \ge t) \le C \exp\left(-\frac{C_2 n t}{\xi^2 + J t}\right).$$

(iii). "Original" SVGD (Liu, Wang '16 NeurIPS)

- Variational Inference Problem.

Find $q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \{KL(q \| p)\}$, $\mathcal{Q}$ a simpler distribution set.

$q_{[T]}(z)$ be the density of $z = T(x)$,

$$q_{[T]}(z) = q(T^{-1}(z)) \cdot |\det(\nabla_z T^{-1}(z))|.$$

Previous methods: consider T with certain parametric form, then optimize parameters.

In [Thm 3.1], they noticed that if

"Directly applying gradient descent to solve above problem." For $x \sim q$, $q_{[T]}$ defined above, $T(x) = x + \varepsilon\phi(x)$

$$\Rightarrow \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} KL(q_{[T]} \| p) = -E_q(\operatorname{tr}(A_p \phi)) \quad (*)$$

$$\left(\Leftrightarrow \frac{dE}{dt}(q) = \int v(\nabla q + q \nabla \ln p)\, dx\right)$$

$$\partial_t q = -\partial V(qV)$$

Again, they were able to minimize (*) in $\{\|\phi\|_{2t}^2 \le S(q,p)\}$: the "direction of steepest descent" is $V(x) = E_{y\sim q}(k(x,y)\nabla_y \ln p(y) + \nabla_y k(x,y))$.

i.e. for this v, $\frac{\partial}{\partial t} q_t = -\text{grad}_{2t} KL(q_t \| p)$

The "metric" W_{2t} will be discussed later.

Advantages, ... "Stein".

Geometric structure ... (17-20')

[Wang: On the geometry of SVGD. Duncan et al. 19]

Q:

① Under which condition we need to consider this gradient flow (or use

"Stein distance" between measures)? What can we benefit from moving particles smoothly?

② Can we find some alternative "positive" flows to improve SVGD (w.r.t. cvg rate; "outliers")?

③ Understand the bias and variance of SVGD particles, or combine it with traditional MC. [from Liu 17'].

(Nathan Ross) - Fundamentals of

Stein's method.

(I. Borisov) - Approximation of distr. of V-stat. with multidimensional kernels.

Thank you!