Controlling Moments with Kernel Stein Discrepancies

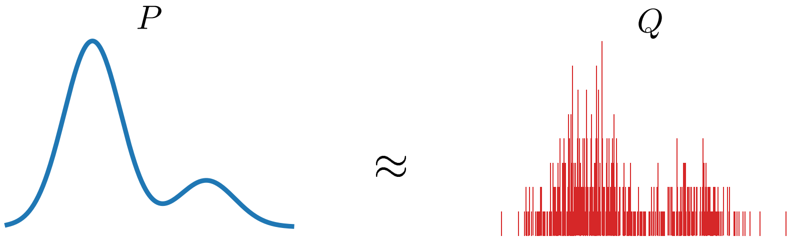
Heishiro Kanagawa

heishiro.kanagawa@gmail.com

Newcastle University, UK

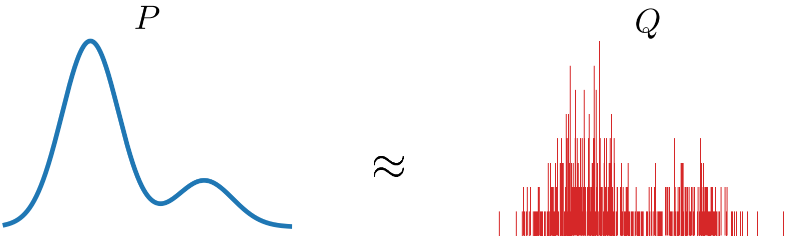
28 June 2023

Intro: evaluating sample quality



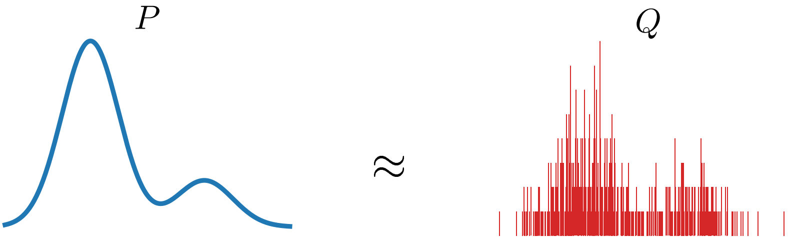
- Task: Approximate $\mathbb{E}_{P}[f]$ with $\mathbb{E}_{Q}[f] = \sum_i w_i f(x_i)$
- Example (Bayesian inference):
  - P : Posterior ($p(x) = \tilde{p}(x)/Z$)
  - Q : Markov chain Monte Carlo sampler
- Question: how good is approximation Q?

Intro: evaluating sample quality



- Task: Approximate $\mathbb{E}_P[f]$ with $\mathbb{E}_Q[f] = \sum_i w_i f(x_i)$
- Example (Bayesian inference):
  - P : Posterior ($p(x) = \tilde{p}(x)/Z$)
  - Q : Markov chain Monte Carlo sampler
- Question: how good is approximation Q?

Intro: evaluating sample quality



- Task: Approximate $\mathbb{E}_P[f]$ with $\mathbb{E}_Q[f] = \sum_i w_i f(x_i)$
- Example (Bayesian inference):
  - P : Posterior $(p(x) = \tilde{p}(x)/Z)$
  - Q : Markov chain Monte Carlo sampler
- Question: how good is approximation Q?

One approach: kernel Stein discrepancy

Kernel Stein discrepancy:

$$\mathrm{KSD}_P(Q) = \sqrt{\mathbb{E}_{X,Y \sim Q \otimes Q}[k_p(X, Y)]}$$

- Computable discrepancy measure
- Just a moment...what can we read off $\mathrm{KSD}_P(Q)$?
  - Does smaller $\mathrm{KSD}_P(Q)$ mean $\mathbb{E}_Q[f]$ is closer to $\mathbb{E}_P[f]$?
  - Does $\mathrm{KSD}_P(Q_n) \to 0$ mean $\mathbb{E}_{Q_n}[f] \to \mathbb{E}_P[f]$?
- This talk: f is of polynomial growth (e.g., $f(x) = x^2$)

One approach: kernel Stein discrepancy

Kernel Stein discrepancy:

$$\mathrm{KSD}_P(Q) = \sqrt{\mathbb{E}_{X,Y \sim Q \otimes Q}[k_p(X, Y)]}$$

- Computable discrepancy measure
- Just a moment...what can we read off $\mathrm{KSD}_P(Q)$?
  - Does smaller $\mathrm{KSD}_P(Q)$ mean $\mathbb{E}_Q[f]$ is closer to $\mathbb{E}_P[f]$?
  - Does $\mathrm{KSD}_P(Q_n) \to 0$ mean $\mathbb{E}_{Q_n}[f] \to \mathbb{E}_P[f]$?
- This talk: f is of polynomial growth (e.g., $f(x) = x^2$)

This talk: closeness in moments

Established result:

$$\mathrm{KSD}_P(Q_n) \to 0 \text{ implies } \sup_{f \in \mathcal{F}_{\mathrm{poly}}} \left| \mathbb{E}_{Q_n}[f] - \mathbb{E}_P[f] \right| \to 0$$

- KSD controls worst-case error w.r.t. $\mathcal{F}_{\mathrm{poly}}$
- The rest of the talk clarifies ambiguities ($\mathcal{F}_{\mathrm{poly}}$, RKHS kernel)

Outline

Controlling Moments
with Kernel Stein Discrepancies

- Introduction to KSD
- Part 1: Stein equation – how KSD is related to a particular function
- Part 2: Clarifying conditions on the RKHS

Prep: Idea of Stein discrepancy

Suppose we have function h_P

$$\mathbb{E}_P[h_P] = 0$$

Then

$$\mathbb{E}_Q[h_P] \neq 0 \Rightarrow Q \neq P$$

P-mean-zero function can quantify $Q \neq P$

Prep: Idea of Stein discrepancy

Suppose we have function h_P

$$\mathbb{E}_P[h_P] = 0$$

Then

$$\mathbb{E}_Q[h_P] \neq 0 \Rightarrow Q \neq P$$

P-mean-zero function can quantify $Q \neq P$

Prep: Idea of Stein discrepancy

Suppose we have function h_P

$$\mathbb{E}_P[h_P] = 0$$

Then

$$\mathbb{E}_Q[h_P] \neq 0 \Rightarrow Q \neq P$$

P-mean-zero function can quantify $Q \neq P$

Prep: Idea of Stein discrepancy

Suppose a family $\mathcal{H}_P$ of P-mean-zero functions,

$$\mathbb{E}_P[h_P] = 0,\ h_P \in \mathcal{H}_P$$

Stein discrepancy = worst-case error

$$\sup_{h \in \mathcal{H}} |\mathbb{E}_Q[h_P]|$$

Non-zero Stein discrepancy $\Rightarrow$ $Q \neq P$

How to prepare such $\mathcal{H}_P$? $\rightarrow$ Stein operator + RKHS

Prep: Idea of Stein discrepancy

Suppose a family $\mathcal{H}_P$ of P-mean-zero functions,

$$\mathbb{E}_P[h_P] = 0, \;\; h_P \in \mathcal{H}_P$$

Stein discrepancy = worst-case error

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_Q[h_P] \right|$$

Non-zero Stein discrepancy $\Rightarrow$ $Q \neq P$

How to prepare such $\mathcal{H}_P$? $\rightarrow$ Stein operator + RKHS

Prep: Idea of Stein discrepancy

Suppose a family $\mathcal{H}_P$ of P-mean-zero functions,

$$\mathbb{E}_P[h_P] = 0, \ h_P \in \mathcal{H}_P$$

Stein discrepancy = worst-case error

$$\sup_{h \in \mathcal{H}} |\mathbb{E}_Q[h_P]|$$

Non-zero Stein discrepancy $\Rightarrow$ $Q \neq P$

How to prepare such $\mathcal{H}_P$? $\rightarrow$ Stein operator + RKHS

Prep: Diffusion Stein operator

Diffusion Stein operator

$$\mathcal{T}_P v(x) = \frac{\langle \nabla, p(x) m(x) v(x) \rangle}{p(x)}$$

where

$$v : \mathbb{R}^d \to \mathbb{R}^d, \;\; m : \mathbb{R}^d \to \mathbb{R}^{d \times d}.$$

Properties:

1. Normalisation constant of p not required
2. Zero mean: if mv is P-integrable (by the divergence theorem),

$$\mathbb{E}_P[\mathcal{T}_P v] = 0$$

Prep: Diffusion Stein operator

Why diffusion? $\to$ associated diffusion

$$\mathrm{d}Z_t^x = b(Z_t^x)\mathrm{d}t + \sigma(Z_t^x)\mathrm{d}B_t \text{ with } Z_0^x = x$$

where

- Drift $b(x) = \langle \nabla, p(x)m(x)\rangle / \{2p(x)\}$
- Diffusion matrix $m(x) = \sigma(x)\sigma(x)^\top$

Then

$$\mathcal{T}_P v(x) = 2\langle b(x), v(x)\rangle + \langle m(x), \nabla v(x)\rangle$$

Prep: Diffusion kernel Stein discrepancy

(Diffusion) Kernel Stein discrepancy:

$$\mathrm{KSD}_{P}(Q) = \sup_{\|v\|_{\mathcal{H}_K} \leq 1} \left| \mathbb{E}_{Q}[\mathcal{T}_{P} v] \right|$$

where $\mathcal{H}_K$ is vector-valued RKHS defined by matrix-valued K

Prep: Diffusion kernel Stein discrepancy

If each $\mathcal{T}_P v$ is Q-integrable,

$$\mathrm{KSD}_P(Q)^2 = \mathbb{E}_{X,Y\sim Q\otimes Q}[k_p(X,Y)]$$

where

$$k_p(x,y) = \frac{1}{p(x)p(y)}\left\langle \nabla_y, \langle \nabla_x, (p(x)\underbrace{m(x)K(x,y)m(y)^\top}_{\text{Langevin KSD}=k\mathrm{Id},\ m\equiv\mathrm{Id}} p(y)\rangle \right\rangle$$

KSD is possible to compute in cloed form

Prep: Diffusion kernel Stein discrepancy

If each $\mathcal{T}_P v$ is Q-integrable,

$$\mathrm{KSD}_P(Q)^2 = \mathbb{E}_{X,Y\sim Q\otimes Q}[k_p(X,Y)]$$

where

$$k_p(x,y) = \frac{1}{p(x)p(y)}\left\langle \nabla_y, \langle \nabla_x, (p(x)\underbrace{m(x)K(x,y)m(y)^\top}_{\text{Langevin KSD}=k\mathrm{Id},\ m\equiv\mathrm{Id}} p(y)\rangle \right\rangle$$

KSD is possible to compute in cloed form

Stein equation

KSD and Stein equation (1)

Q. How is $\mathrm{KSD}_P(Q)$ *related to* $|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|$?

KSD and Stein equation (1)

$$Q.\ \textit{How is } \mathrm{KSD}_P(Q) \textit{ related to } |\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|?$$

Suppose we have a solution v_f to *Stein equation*:

$$\mathcal{T}_P v = f - \mathbb{E}_P[f]$$

KSD and Stein equation (1)

$$Q.\ \textit{How is}\ \mathrm{KSD}_P(Q)\ \textit{related to}\ |\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|?$$

Suppose we have a solution v_f to *Stein equation*:

$$\mathcal{T}_P v = f - \mathbb{E}_P[f]$$

Step 1: rewrite $|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|$ as

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| = \left|\mathbb{E}_Q\left[\mathcal{T}_P v_f\right]\right|$$

KSD and Stein equation (1)

Q. How is $\mathrm{KSD}_P(Q)$ *related to* $|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|$?

Suppose we have a solution v_f to *Stein equation*:

$$\mathcal{T}_P v = f - \mathbb{E}_P[f]$$

Step 2: approximate $\mathcal{T}_P v_f$ using an RKHS function

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| = \left|\mathbb{E}_Q\left[\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}} + \mathcal{T}_P v_{\mathrm{RKHS}}\right]\right|$$

KSD and Stein equation (1)

Q. How is $\mathrm{KSD}_P(Q)$ *related to* $|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]|$?

Suppose we have a solution v_f to *Stein equation*:

$$\mathcal{T}_P v = f - \mathbb{E}_P[f]$$

Step 2: approximate $\mathcal{T}_P v_f$ using an RKHS function

$$\begin{aligned}
|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| &= \left|\mathbb{E}_Q\left[\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}} + \mathcal{T}_P v_{\mathrm{RKHS}}\right]\right| \\
&\le \left|\mathbb{E}_Q\left[\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}\right]\right| + \left|\mathbb{E}_Q\left[\mathcal{T}_P v_{\mathrm{RKHS}}\right]\right| \\
&\le \underbrace{\mathbb{E}_Q\left[\left|\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}\right|\right]}_{\text{Approximation error}} + \underbrace{\|v_{\mathrm{RKHS}}\|_{\mathcal{H}_K} \mathrm{KSD}_P(Q)}_{\text{Stein discrepancy (and norm)}}
\end{aligned}$$

KSD and Stein equation (2)

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq \underbrace{\mathbb{E}_Q\left[\left|\mathcal{T}_P v_f - \mathcal{T}_P v_{\text{RKHS}}\right|\right]}_{\text{Approximation error}} + \underbrace{\|v_{\text{RKHS}}\|_{\mathcal{H}_K} \text{KSD}_P(Q)}_{\text{Stein discrepancy (and norm)}}$$

Comments:

- Key idea: bounding the error yields an estimate of $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$
- A result: $\text{KSD}_P(Q_n) \to 0$ implies $|\mathbb{E}_P[f] - \mathbb{E}_{Q_n}[f]| \to 0$ if well approximated

Two questions:

1. Stein equation and solution:
   - Do we have a solution to $\mathcal{T}_P v_f = f - \mathbb{E}_P[f]$?
   - What properties does v_f have?
2. RKHS – what conditions required to achieve approximation?

KSD and Stein equation (2)

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq \underbrace{\mathbb{E}_Q\left[\left|\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}\right|\right]}_{\text{Approximation error}} + \underbrace{\|v_{\mathrm{RKHS}}\|_{\mathcal{H}_K} \mathrm{KSD}_P(Q)}_{\text{Stein discrepancy (and norm)}}$$

Comments:

- Key idea: bounding the error yields an estimate of $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$
- A result: $\mathrm{KSD}_P(Q_n) \to 0$ implies $|\mathbb{E}_P[f] - \mathbb{E}_{Q_n}[f]| \to 0$ if well approximated

Two questions:

1. Stein equation and solution:
   - Do we have a solution to $\mathcal{T}_P v_f = f - \mathbb{E}_P[f]$?
   - What properties does v_f have?
2. RKHS – what conditions required to achieve approximation?

Preparation: Pseudo-Lipschitz functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is pseudo-Lipschitz of order $q - 1$ if

$$\frac{|f(x) - f(y)|}{\|x - y\|_2} \leq C(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}) \text{ for all } x, y \in \mathbb{R}^d,$$

Some comments:

- C is a constant; we use $C = 1$ (and some other conditions)
- $q = 1$ recovers the usual Lipschitz-ness
- f (and its derivatives) are allowed to grow like deg-q polynomials

Preparation: Pseudo-Lipschitz functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is pseudo-Lipschitz of order $q - 1$ if

$$\frac{|f(x) - f(y)|}{\|x - y\|_2} \le C(1 + \|x\|_2^{q-1} + \|y\|_2^{q-1}) \text{ for all } x, y \in \mathbb{R}^d,$$

Some comments:

- C is a constant; we use $C = 1$ (and some other conditions)
- $q = 1$ recovers the usual Lipschitz-ness
- f (and its derivatives) are allowed to grow like deg-q polynomials

Stein equation and solution

$$\mathcal{T}_P v_f = f - \mathbb{E}_P[f]$$

- Existence of solution depends on P and f
- Solution is often implicit but can be characterised as follows:

Theorem (Erdogdu, Mackey, and Shamir, Neurips 2018)

If $f \in C^3$ is pseudo-Lipschitz of order $q-1$, under appropriate conditions on P,

$$\|\nabla^i v_f(x)\|_{\text{op}} \leq \zeta_i(P,f)\left(1 + \|x\|_2^{q-1}\right), \textit{ for } i \in \{0,1,2\};$$

i.e., the growth of v_f (and derivatives) is of $O(\|x\|_2^{q-1})$

An appropriate subset $\mathcal{F}$ of pLip functions makes $\zeta_i(P,f)$ independent of specific f

Stein equation and solution

$$\mathcal{T}_P v_f = f - \mathbb{E}_P[f]$$

- Existence of solution depends on P and f
- Solution is often implicit but can be characterised as follows:

Theorem (Erdogdu, Mackey, and Shamir, Neurips 2018)

If $f \in \mathcal{C}^3$ is pseudo-Lipschitz of order $q-1$, under appropriate conditions on P,

$$\|\nabla^i v_f(x)\|_{\mathrm{op}} \leq \zeta_i(P, f)\left(1 + \|x\|_2^{q-1}\right), \text{ for } i \in \{0, 1, 2\};$$

i.e., the growth of v_f (and derivatives) is of $O(\|x\|_2^{q-1})$

An appropriate subset $\mathcal{F}$ of pLip functions makes $\zeta_i(P, f)$ independent of specific f

RKHS to control moments

Conditions on RKHS

Recall $\mathcal{T}_P v_f(x) = f - \mathbb{E}_P[f] = O\left(\|x\|_2^q\right)$ and

$$\mathcal{T}_P v(x) = 2\langle b(x, v(x)\rangle + \langle m(x), \nabla v(x)\rangle$$

Evaluate approximation error $\left|\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}\right|$:

$$\left|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)\right| \underbrace{(1\{\|x\|_2 > r\} + 1\{\|x\|_2 \leq r\})}_{=1}$$

$$\leq \underbrace{2\|x\|_2^q 1\{\|x\|_2 > r\}}_{\text{(A):Behaviour at infinity}} + \underbrace{\left|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)\right| 1\{\|x\|_2 \leq r\}}_{\text{(B):Error in bounded region}}$$

Desiderata on RKHS $\mathcal{H}_K$:

- $\mathcal{T}_P(\mathcal{H}_K)$ consists of $O(\|x\|_2^q)$ functions
- $\mathcal{T}_P(\mathcal{H}_K)$ can approximate $x \mapsto \|x\|_2^q 1\{\|x\| > r\}$
- $\mathcal{H}_K$ can approximate any function up to first derivatives

Conditions on RKHS

Recall $\mathcal{T}_P v_f(x) = f - \mathbb{E}_P[f] = O(\|x\|_2^q)$ and

$$\mathcal{T}_P v(x) = 2\langle b(x, v(x)\rangle + \langle m(x), \nabla v(x)\rangle$$

Evaluate approximation error $|\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}|$:

$$|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)| \underbrace{(1\{\|x\|_2 > r\} + 1\{\|x\|_2 \le r\})}_{=1}$$

$$\le \underbrace{2\|x\|_2^q 1\{\|x\|_2 > r\}}_{\text{(A):Behaviour at infinity}} + \underbrace{|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)|\, 1\{\|x\|_2 \le r\}}_{\text{(B):Error in bounded region}}$$

Desiderata on RKHS $\mathcal{H}_K$:

- $\mathcal{T}_P(\mathcal{H}_K)$ consists of $O(\|x\|_2^q)$ functions
- $\mathcal{T}_P(\mathcal{H}_K)$ can approximate $x \mapsto \|x\|_2^q 1\{\|x\| > r\}$
- $\mathcal{H}_K$ can approximate any function up to first derivatives

Conditions on RKHS

Recall $\mathcal{T}_P v_f(x) = f - \mathbb{E}_P[f] = O(\|x\|_2^q)$ and

$$\mathcal{T}_P v(x) = 2\langle b(x, v(x)\rangle + \langle m(x), \nabla v(x)\rangle$$

Evaluate approximation error $|\mathcal{T}_P v_f - \mathcal{T}_P v_{\mathrm{RKHS}}|$:

$$|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)| \underbrace{(1\{\|x\|_2 > r\} + 1\{\|x\|_2 \leq r\})}_{=1}$$

$$\leq \underbrace{2\|x\|_2^q 1\{\|x\|_2 > r\}}_{\text{(A):Behaviour at infinity}} + \underbrace{|\mathcal{T}_P v_f(x) - \mathcal{T}_P v_{\mathrm{RKHS}}(x)|\, 1\{\|x\|_2 \leq r\}}_{\text{(B):Error in bounded region}}$$

Desiderata on RKHS $\mathcal{H}_K$:

- $\mathcal{T}_P(\mathcal{H}_K)$ consists of $O(\|x\|_2^q)$ functions
- $\mathcal{T}_P(\mathcal{H}_K)$ can approximate $x \mapsto \|x\|_2^q 1\{\|x\| > r\}$
- $\mathcal{H}_K$ can approximate any function up to first derivatives

Conditions on RKHS (contd.)

Proposition

RKHS defined by kernel $K = k\mathrm{Id}$ with

$$k(x,y) = w(x)w(y)\left(\ell(x,y) + \frac{\tau^2 + \langle x, y\rangle}{\sqrt{\tau^2+\|x\|^2}\sqrt{\tau^2+\|y\|^2}}\right)$$

satisfies the desiderata if

1. ℓ *is translation invariant and* C_0^1*-universal (e.g., Matérn, Gaussian, IMQ kernels)*
2. $w(x) = (\tau^2 + \|x\|_2^2)^{(q-1)/2}$
3. *The* P*-targeting diffusion is dissipative; i.e.,*

$$2\langle b(x), x\rangle + \mathrm{tr}[m(x)] \leq -\alpha\|x\|_2^2 + \beta$$

for $\alpha, \beta > 0$

Conditions on RKHS (contd.)

Proposition

RKHS defined by kernel $K = k\mathrm{Id}$ *with*

$$k(x, y) = w(x)w(y)\left(\ell(x, y) + \frac{\tau^2 + \langle x, y\rangle}{\sqrt{\tau^2 + \|x\|^2}\sqrt{\tau^2 + \|y\|^2}}\right)$$

satisfies the desiderata if

1. ℓ *is translation invariant and* C_0^1*-universal (e.g., Matérn, Gaussian, IMQ kernels)*
2. $w(x) = (\tau^2 + \|x\|_2^2)^{(q-1)/2}$
3. *The* P*-targeting diffusion is dissipative; i.e.,*

$$2\langle b(x), x\rangle + \mathrm{tr}[m(x)] \leq -\alpha\|x\|_2^2 + \beta$$

for $\alpha, \beta > 0$

Main result: KSD bound on pLip metric

Theorem (Informal bound)

For any $\varepsilon > 0$, we have

$$\sup_{f \in \mathcal{F}_q} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq c_{P,d} \left(g(\varepsilon^{-1}) \cdot \mathrm{KSD}_P(Q) \right) + \varepsilon$$

where

- $\mathcal{F}_q \approx \{$*1-pseudo Lipschitz functions of order* $q-1\}$
- $c_{P,d} > 0$
- g*: increasing function*

For sequence of distributions $\{Q_1, Q_2, \ldots\}$,

$$\mathrm{KSD}_P(Q_n) \to 0 \Rightarrow \sup_{f \in \mathcal{F}_q} |\mathbb{E}_P[f] - \mathbb{E}_{Q_n}[f]| \to 0$$

"KSD convergence implies moment convergence"

Main result: KSD bound on pLip metric

Theorem (Informal bound)

For any $\varepsilon > 0$, we have

$$\sup_{f \in \mathcal{F}_q} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq c_{P,d} \left(g(\varepsilon^{-1}) \cdot \mathrm{KSD}_P(Q) \right) + \varepsilon$$

where

- $\mathcal{F}_q \approx \{$*1-pseudo Lipschitz functions of order* $q - 1\}$
- $c_{P,d} > 0$
- g*: increasing function*

For sequence of distributions $\{Q_1, Q_2, \ldots\}$,

$$\mathrm{KSD}_P(Q_n) \to 0 \Rightarrow \sup_{f \in \mathcal{F}_q} |\mathbb{E}_P[f] - \mathbb{E}_{Q_n}[f]| \to 0$$

"KSD convergence implies moment convergence"

Matérn KSD bound on pLip metric

Theorem

There exist $c_{P,d}, c_{\nu,q} > 0$ *such that*

$$\sup_{f \in \mathcal{F}_q} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq c_{P,d} \cdot \mathrm{KSD}_P\,(Q)^{\frac{1}{d+1+c_{\nu,q}}}$$

if ℓ *is chosen as*

$$\ell(x, y) = \frac{2^{1-(d/2+\nu)}}{\Gamma\{(d/2+\nu)\}} \|x-y\|_2^{\nu} K_{-\nu}(\|x-y\|_2),$$

where $K_{-\nu}$ *is the Bessel function of the second kind and* $\nu > 1$

Experiments

Toy experiment 1: variance perturbation

Convergence of contaminated distribution

$$P = \mathcal{N}(0,\ \mathrm{Id}),\ \ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\{0, (n+1)\mathrm{Id}\}$$

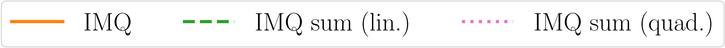
- Q_n converges P in distribution but not in <u>variance</u>
- Check KSD with IMQ kernel $\ell(x, y) = \left(1 + \|x - y\|_2^2\right)^{-1/2}$

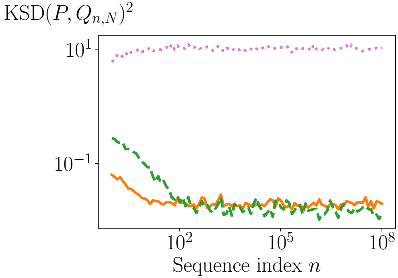
Toy experiment 1: variance perturbation

Convergence of contaminated distribution

$$P = \mathcal{N}(0,\ \mathrm{Id}),\ \ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\{0, (n+1)\mathrm{Id}\}$$

- Q_n converges P in distribution but not in variance
- Check KSD with IMQ kernel $\ell(x, y) = \left(1 + \|x - y\|_2^2\right)^{-1/2}$





Increase n with N fixed:

$$Q_{n,N} = \left(1 - \frac{1}{n+1}\right) \hat{P}_N + \frac{1}{n+1}\hat{\mathcal{N}}_N$$

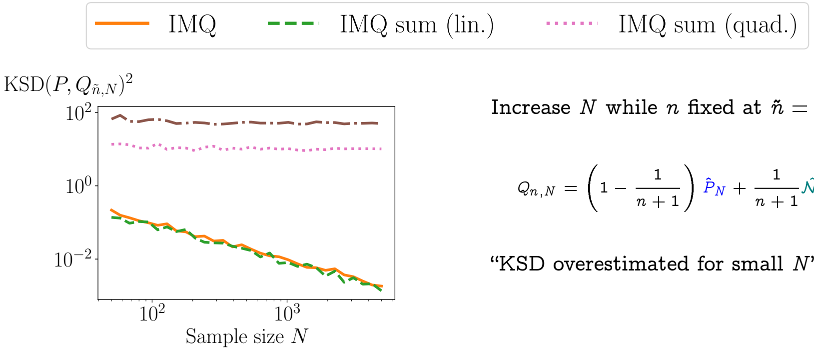
$$\hat{P}_N = N^{-1}\sum_{i=1}^{N} \delta_{X_i},\ \ \hat{\mathcal{N}}_N = N^{-1}\sum_{i=1}^{N} \delta_{\tilde{X}_i}$$

Toy experiment 1: variance perturbation

Convergence of contaminated distribution

$$P = \mathcal{N}(0,\ \mathrm{Id}),\ \ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\{0, (n+1)\mathrm{Id}\}$$

- Q_n converges P in distribution but not in variance
- Check KSD with IMQ kernel $\ell(x, y) = \left(1 + \|x - y\|_2^2\right)^{-1/2}$



Increase N while n fixed at $\tilde{n} = 10^6$

$$Q_{n,N} = \left(1 - \frac{1}{n+1}\right) \hat{P}_N + \frac{1}{n+1}\hat{\mathcal{N}}_N$$

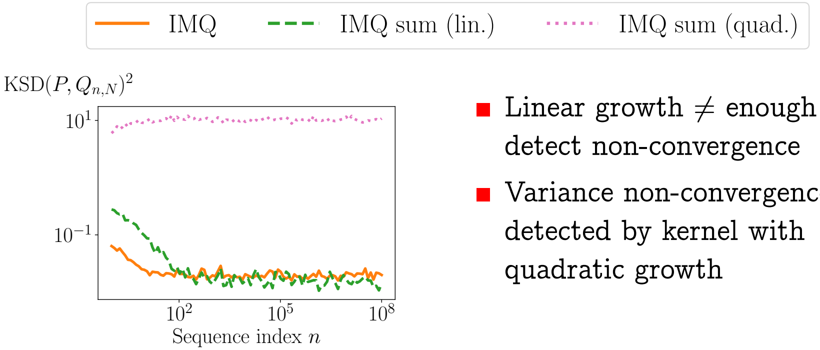
"KSD overestimated for small N"

Toy experiment 1: variance perturbation

Convergence of contaminated distribution

$$P = \mathcal{N}(0,\ \mathrm{Id}),\ \ Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1} \mathcal{N}\{0, (n+1)\mathrm{Id}\}$$

- Q_n converges P in distribution but not in variance
- Check KSD with IMQ kernel $\ell(x, y) = \left(1 + \|x - y\|_2^2\right)^{-1/2}$



- Linear growth $\neq$ enough to detect non-convergence
- Variance non-convergence detected by kernel with quadratic growth

Toy experiment 2: heavy-tailed target

Standard Student's t-distribution

$$p(x) = \frac{\Gamma\left(\frac{d+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}}\left(1 + \frac{\|x\|_2^2}{\nu}\right)^{-\frac{d+\nu}{2}}$$

- Langevin diffusion $\sigma(x) = \mathrm{Id}$ does not satisfy the required conditions
- Itô diffusion with diffusion coefficient $\sigma(x) = \sqrt{1 + \nu^{-1}\|x\|_2^2}\mathrm{Id}$ does
- Recall: Stein kernel for diffusion KSD

$$k_p(x,y) = \frac{1}{p(x)p(y)}\left\langle \nabla_y, \langle \nabla_x, (p(x)m(x)K(x,y)m(y)^\top p(y)\rangle \right\rangle$$

$\rightarrow$ use $m(x) = \sigma(x)\sigma(x)^\top = (1 + \nu^{-1}\|x\|_2^2)\mathrm{Id}$

Toy experiment 2: heavy-tailed target

Standard Student's t-distribution

$$p(x) = \frac{\Gamma\left(\frac{d+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}}\left(1 + \frac{\|x\|_2^2}{\nu}\right)^{-\frac{d+\nu}{2}}$$

- Langevin diffusion $\sigma(x) = \mathrm{Id}$ does not satisfy the required conditions
- Itô diffusion with diffusion coefficient $\sigma(x) = \sqrt{1 + \nu^{-1}\|x\|_2^2}\mathrm{Id}$ does
- Recall: Stein kernel for diffusion KSD

$$k_p(x, y) = \frac{1}{p(x)p(y)}\left\langle \nabla_y, \langle \nabla_x, (p(x)m(x)K(x, y)m(y)^\top p(y)\rangle \right\rangle$$

$\rightarrow$ use $m(x) = \sigma(x)\sigma(x)^\top = (1 + \nu^{-1}\|x\|_2^2)\mathrm{Id}$

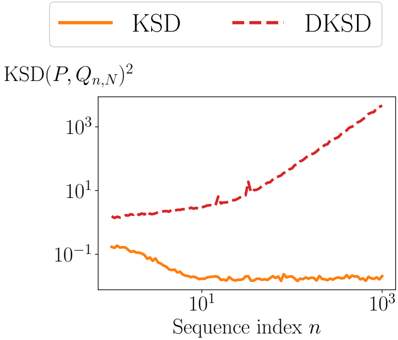
Mean perturbation with heavy-tailed target

Convergence of contaminated distribution

$$p(x) \propto \left(1 + \nu^{-1}\|x\|_2^2\right)^{-(\nu+d)/2}$$

$$Q_n = \left(1 - \frac{1}{n+1}\right) P + \frac{1}{n+1}\mathcal{N}\{(n+1)1, \mathrm{Id}\}$$

(Q_n converges P in distribution but not in mean)



- Both use proposed k with IMQ
- Langevian KSD (IMQ) fails to detect non-convergence
- DKSD detectes mean non-convergence

Summary

1. Kernel Stein discrepancy: computable discrepancy measure
2. Clarified conditions when KSD implies moment convergence
3. Presented a practical kernel

- Reference (to be updated soon, hopefully):
  Controlling Moments with Kernel Stein Discrepancies
  Heishiro Kanagawa, Alessandro Barp, Carl-Johann Simon-Gabriel, Arthur Gretton, Lester Mackey
  https://arxiv.org/abs/2211.05408
- Python code:
  https://github.com/noukoudashisoup/ksd-moment

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Key assumptions on diffusion

Required assumptions:

1. Dissipativity

$$\mathcal{A}_P \|x\|_2^2 \leq -\alpha \|x\|_2^2 + \beta,$$

where $\mathcal{A}_P f(x) = \langle b(x), \nabla f(x) \rangle + \frac{1}{2} \langle \sigma(x)\sigma(x)^\top, \nabla^2 f(x) \rangle$

2. Wasserstein decay (ρ needs to be fast-decaying)

$$\inf_{\text{couplings}(Z_t^x, Z_t^y)} \mathbb{E}[\|Z_t^x - Z_t^y\|_2] \leq \rho(t) \|x - y\|_2 \text{ for } x, y \in \mathbb{R}^D$$