Stein's Method in Statistics

Xing Liu

April 25, 2022

Imperial College London

Main Reference for the Talk

A. Anastasiou, A. Barp, F.-X. Briol, et al. (2021) *Stein's Method Meets Statistics: A Review of Some Recent Developments*

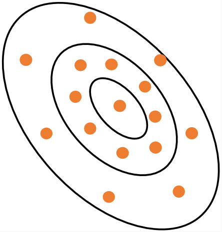
Table of Contents

Ingredients of Stein's Method

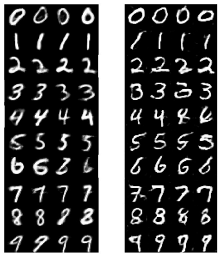
Motivation — Quantifying Discrepancy

Let $\mathcal{X} \subset \mathbb{R}^d$ and P a probability measure on $\mathcal{X}$.

Problem of interest: Given another probability measure Q on $\mathcal{X}$, how to quantify the discrepancy from Q to P ?



P: target distribution
Q: MCMC samples



P: generative models
Q: true images

Motivation — Quantifying Discrepancy

Integral Probability Metrics (IPM)

Given a family $\mathcal{H} \subset L^1(P) \cap L^1(Q)$ of read-valued functions, the IPM[1] is the distance metric

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)]|.$$

- **Total Variation distance**: $\mathcal{H} = \{h : \mathcal{X} \to \mathbb{R} : \sup_{x\in\mathcal{X}} |h(x)| \leq 1\}$
- L^1**-Wasserstein distance**, d_W:
  $\mathcal{H}_W = \{h : \mathcal{X} \to \mathbb{R} : |h(x) - h(y)| \leq \|x - y\|_2, \forall x, y\}$
- **Bounded Wasserstein distance/Dudley metric**, d_{bW}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

---

[1] Müller [1997]

Motivation — Quantifying Discrepancy

Integral Probability Metrics (IPM)

Given a family $\mathcal{H} \subset L^1(P) \cap L^1(Q)$ of read-valued functions, the IPM is the distance metric

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{X \sim P}[h(X)]|.$$

Problem: $d_{\mathcal{H}}(Q, P)$ requires integrating over P, so it cannot be computed!

Solution: Find $\mathcal{H}$ so that $\forall h \in \mathcal{H}$, $\mathbb{E}_{X \sim P}[h(X)] = 0$. Then

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim Q}[h(X)] - \cancel{\mathbb{E}_{X \sim P}[h(X)]}\,|.$$

How to choose $\mathcal{H}$ for a generic P ? — Use *Stein's method* !

Motivation — Quantifying Discrepancy

Integral Probability Metrics (IPM)

Given a family $\mathcal{H} \subset L^1(P) \cap L^1(Q)$ of read-valued functions, the IPM is the distance metric

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{X \sim P}[h(X)]|.$$

Problem: $d_{\mathcal{H}}(Q, P)$ requires integrating over P, so it cannot be computed!

Solution: Find $\mathcal{H}$ so that $\forall h \in \mathcal{H}$, $\mathbb{E}_{X \sim P}[h(X)] = 0$. Then

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim Q}[h(X)] - \cancel{\mathbb{E}_{X \sim P}[h(X)]}\,|.$$

How to choose $\mathcal{H}$ for a generic P ? — Use *Stein's method* !

Given a probability measure P on $\mathcal{X}$, we are interested in finding a linear operator $\mathcal{T}$ acting on some set $\mathcal{G}(\mathcal{T})$ of functions on $\mathcal{X}$ such that

For all probability measure Q on $\mathcal{X}$,

$$Q = P \iff \mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] = 0, \text{ for all } g \in \mathcal{G}(\mathcal{T}). \tag{1}$$

Glossary:

- **Stein operator**: $\mathcal{T}$
- **Stein class**: $\mathcal{G}(\mathcal{T})$ for which $\mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] = 0$ for all $g \in \mathcal{G}(\mathcal{T})$
- **Stein set**: Any $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$
- **Stein characterisation**: The equivalence (1)

Given a probability measure P on $\mathcal{X}$, we are interested in finding a linear operator $\mathcal{T}$ acting on some set $\mathcal{G}(\mathcal{T})$ of functions on $\mathcal{X}$ such that

For all probability measure Q on $\mathcal{X}$,

$$Q = P \iff \mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)] = 0, \text{ for all } g \in \mathcal{G}(\mathcal{T}). \tag{1}$$

Glossary:

- **Stein operator**: $\mathcal{T}$
- **Stein class**: $\mathcal{G}(\mathcal{T})$ for which $\mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)] = 0$ for all $g \in \mathcal{G}(\mathcal{T})$
- **Stein set**: Any $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$
- **Stein characterisation**: The equivalence (1)

Why Stein's Method?

Stein's method is useful in many areas:

- **Theoretical stats:**
  - Deriving explicit (non-asymptotic) bounds on the distance between distributions. [Reinert, 1998, Mijoule et al., 2021]
- **Computational stats/machine learning:**
  - Quantifying the discrepancy between distributions (Stein Discrepancy) [Gorham and Mackey, 2015, Liu et al., 2016, Chwialkowski et al., 2016].
  - Sampling from unnormalised densities (Stein Variational Gradient Descent). [Liu and Wang, 2016, Gong et al., 2021, Liu et al., 2022]
  - Training generative models [Grathwohl et al., 2020].
  - Variance reduction [Mira et al., 2013, Oates et al., 2017]

Constructing the Stein Operator $\mathcal{T}$

TL; DR: So long as P is *sufficiently regular*, a Stein operator $\mathcal{T}$ (and Stein class $\mathcal{G}(\mathcal{T})$) can be constructed in a schematic approach.

Approaches:

- Generator approach
- Density approach
- Couplings, orthogonal polynomials, ODEs...

Constructing the Stein Operator $\mathcal{T}$

TL; DR: So long as P is *sufficiently regular*, a Stein operator $\mathcal{T}$ (and Stein class $\mathcal{G}(\mathcal{T})$) can be constructed in a schematic approach.

Approaches:

- Generator approach
- Density approach
- Couplings, orthogonal polynomials, ODEs...

Generator Approach

If a Markov process $(Z_t)_{t\geq 0}$ with invariant measure P is sufficiently regular (i.e. a *Feller process*) (e.g. when P has a density function $p : \mathcal{X} \to \mathbb{R}^+$ w.r.t. some dominating measure), then it has an *infinitesimal generator* $\mathcal{T}$ that satisfies

$$\mathbb{E}_{Z\sim P}[(\mathcal{T}u)(Z)] = 0 \text{ for all } u : \mathbb{R}^d \to \mathbb{R} \text{ in the domain of } \mathcal{T}.$$

Generator Approach — Examples

E.g.1 Standard multivariate Normal

$P = \mathcal{N}(0, I_d)$ is an invariant measure of the process $Z_{t,x} = e^{-t}x + \sqrt{1 - e^{-2t}}Z$, where $Z \sim \mathcal{N}(0, I_d)$. The Stein operator is

$$(\mathcal{T}g)(x) = \nabla^\intercal \nabla g(x) - x^\intercal g(x),$$

for twice differentiable $u : \mathbb{R}^d \to \mathbb{R}$.

E.g.2 Langevin Stein Operator (Popular in ML!)

Let P have density p supported on $\mathcal{X}$. Assume $\mathbb{E}_{X\sim P}[\|\nabla \log p(x)\|_2] < \infty$. P is an invariant measure of the *Langevin diffusion* $dZ_{t,x} = \frac{1}{2p(x)}\langle \nabla, p(x)\rangle dt + dW_t$, where $(W_t)_{t\geq 0}$ is a Brownian motion. This leads to the *Langevin Stein operator*

$$(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle.$$

Generator Approach — Examples

E.g.1 Standard multivariate Normal

$P = \mathcal{N}(0, I_d)$ is an invariant measure of the process $Z_{t,x} = e^{-t}x + \sqrt{1-e^{-2t}}Z$, where $Z \sim \mathcal{N}(0, I_d)$. The Stein operator is

$$(\mathcal{T}g)(x) = \nabla^\intercal \nabla g(x) - x^\intercal g(x),$$

for twice differentiable $u : \mathbb{R}^d \to \mathbb{R}$.

E.g.2 Langevin Stein Operator (Popular in ML!)

Let P have density p supported on $\mathcal{X}$. Assume $\mathbb{E}_{X\sim P}[\|\nabla \log p(x)\|_2] < \infty$. P is an invariant measure of the *Langevin diffusion* $dZ_{t,x} = \frac{1}{2p(x)}\langle \nabla, p(x)\rangle dt + dW_t$, where $(W_t)_{t\geq 0}$ is a Brownian motion. This leads to the *Langevin Stein operator*

$$(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle.$$

Applications in Theoretical Statistics

Stein Equation

Let $\mathcal{T}$ be a Stein operator and $\mathcal{G}(\mathcal{T})$ a Stein class. For any $g \in \mathcal{G}(\mathcal{T})$, we can find h so that

$$(\mathcal{T}g)(\cdot) = h(\cdot) - \mathbb{E}_{X\sim P}[h(X)]. \tag{2}$$

"Reversed" question: Given $h \in \mathcal{H} \subset L^1(P)$, when does a solution $g = g_h$ to (2) exist?

- Why bother? Studying the properties of g_h can help us to bound differences of the form

$$\mathbb{E}_{W_n}[h(W_n)] - \mathbb{E}_{X\sim P}[h(X)] = \mathbb{E}_{W_n}[(\mathcal{T}g)(W_n)],$$

where W_n is a sum of independent terms.

Answer:

- Existence of g_h guaranteed with many $\mathcal{T}$ and $\mathcal{G}(\mathcal{T})$.
- Regularity on g_h can be shown assuming regularity on h.

Stein Equation

Let $\mathcal{T}$ be a Stein operator and $\mathcal{G}(\mathcal{T})$ a Stein class. For any $g \in \mathcal{G}(\mathcal{T})$, we can find h so that

$$(\mathcal{T}g)(\cdot) = h(\cdot) - \mathbb{E}_{X \sim P}[h(X)]. \tag{2}$$

"Reversed" question: Given $h \in \mathcal{H} \subset L^1(P)$, when does a solution $g = g_h$ to (2) exist?

- Why bother? Studying the properties of g_h can help us to bound differences of the form

$$\mathbb{E}_{W_n}[h(W_n)] - \mathbb{E}_{X \sim P}[h(X)] = \mathbb{E}_{W_n}[(\mathcal{T}g)(W_n)],$$

where W_n is a sum of independent terms.

Answer:

- Existence of g_h guaranteed with many $\mathcal{T}$ and $\mathcal{G}(\mathcal{T})$.
- Regularity on g_h can be shown assuming regularity on h.

Stein Equation

Let $\mathcal{T}$ be a Stein operator and $\mathcal{G}(\mathcal{T})$ a Stein class. For any $g \in \mathcal{G}(\mathcal{T})$, we can find h so that

$$(\mathcal{T}g)(\cdot) = h(\cdot) - \mathbb{E}_{X\sim P}[h(X)]. \quad (2)$$

"Reversed" question: Given $h \in \mathcal{H} \subset L^1(P)$, when does a solution $g = g_h$ to (2) exist?

- Why bother? Studying the properties of g_h can help us to bound differences of the form

$$\mathbb{E}_{W_n}[h(W_n)] - \mathbb{E}_{X\sim P}[h(X)] = \mathbb{E}_{W_n}[(\mathcal{T}g)(W_n)],$$

  where W_n is a sum of independent terms.

Answer:

- Existence of g_h guaranteed with many $\mathcal{T}$ and $\mathcal{G}(\mathcal{T})$.
- Regularity on g_h can be shown assuming regularity on h.

Example 1: Central Limit Theorem

E.g.1 Central Limit Theorem

Let univariate $X_1, \ldots, X_n$ be independent, zero-mean with unit variance, and $\mathbb{E}[|X_i^3|] < \infty$. Put $W_n = n^{-1/2} \sum_{i=1}^n X_i$, and let Q_n denote the measure of W_n. Then

$$d_W(Q_n, \mathcal{N}(0,1)) \leq \frac{1}{\sqrt{n}} \left(2 + \frac{1}{n} \sum_i \mathbb{E}[|X_i^3|]\right).$$

Idea of proof: Fix h 1-Lipschitz with derivative h'.

$$\begin{aligned}
&\mathbb{E}[h(W_n)] - \mathbb{E}[h(Z)] \\
&= \mathbb{E}[h(W_n) - \mathbb{E}[h(Z)]] \\
&= \mathbb{E}[g_h''(W_n) - W_n g_h'(W_n)] \text{ for some } g_h \text{ with } \|g_h^{(3)}\|_\infty \leq 2\|h\|_\infty. \\
&\leq \cdots \\
&\leq \frac{\|h'\|_\infty}{\sqrt{n}} \left(2 + \frac{1}{n} \sum_i \mathbb{E}[|X_i^3|]\right).
\end{aligned}$$

Example 2: Explicit bound on normality of MLE

Let $X_1, \ldots, X_n$ be i.i.d. from a single-parameter distribution P_{θ_0} with parameter space Θ. Under regularity conditions, as $n \to \infty$,

- Asymptotic normality of MLE:

$$W_n \coloneqq \sqrt{ni(\theta_0)}(\hat{\theta}_n(X) - \theta_0) \to_d \mathcal{N}(0,1).$$

- Anastasiou and Reinert [2017]: For ϵ with $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset \Theta$,

$$\begin{aligned}
&d_{bW}(W_n, \mathcal{N}(0,1)) \\
&\leq \tfrac{1}{n}\left(2 + \tfrac{1}{[i(\theta_0)]^{3/2}}\mathbb{E}\left[\left|\tfrac{d}{d\theta}\log f(X_1|\theta_0)\right|^3\right]\right) \\
&+ \tfrac{1}{\sqrt{i(\theta_0)}}\sqrt{\mathrm{Var}\left(\tfrac{d^2}{d\theta^2}\log f(X_1|\theta_0)\right)}\sqrt{\mathbb{E}[(\hat{\theta}_n(X) - \theta_0)^2]} \\
&+ \tfrac{2}{\epsilon^2}\mathbb{E}[(\hat{\theta}_n(X) - \theta_0)^2] \\
&+ \tfrac{1}{2\sqrt{ni(\theta_0)}}\left[\mathbb{E}\left[\left(\textstyle\sum_i M(X_i)\right)^2 \middle| |\hat{\theta}_n(X) - \theta_0| < \epsilon\right]\right]^{1/2}\left[\mathbb{E}[(\hat{\theta}_n(X) - \theta_0)^4]\right]^{1/2}
\end{aligned}$$

Each term on the RHS can be computed *explicitly* for simple P_θ!

Example 2: Explicit bound on normality of MLE

E.g. Exponential distribution

Let $P_{\theta_0} = \text{Exponential}(\theta_0)$. Then, for $\epsilon = \theta_0/2 > 0$,

$$d_{bW}(W_n, P_{\theta_0}) \leq \frac{4.41456}{\sqrt{n}} + \frac{8(n+2)(1+\sqrt{n})}{(n-1)(n-2)}.$$

Applications in Machine Learning

A Discrepancy based on Stein's Method

Recall: The IPM is $d_{\mathcal{H}}(Q,P) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)]|$.

Stein Discrepancy

Given a valid Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$, choosing $\mathcal{H} = \{\mathcal{T}g : g \in \mathcal{G}\}$ in IPM defines a discrepancy, called the *Stein discrepancy* [2]: $\mathbb{S}(Q,P,\mathcal{G}) = \sup_{g\in\mathcal{G}} \|\mathbb{E}_{X\sim G}[(\mathcal{T}g)(X)]\|_2$.

How to choose $\mathcal{T}$? Langevin Stein operator

$$(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle.$$

How to choose $\mathcal{G}$? Ideally, want

- Discriminability: $\mathbb{S}(Q,P,\mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q,P,\mathcal{G})$ can be efficiently computed.

[2] [Gorham and Mackey, 2015]

A Discrepancy based on Stein's Method

Recall: The IPM is $d_{\mathcal{H}}(Q,P) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)]|$.

Stein Discrepancy

Given a valid Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$, choosing $\mathcal{H} = \{\mathcal{T}g : g \in \mathcal{G}\}$ in IPM defines a discrepancy, called the *Stein discrepancy* [2]: $\mathbb{S}(Q,P,\mathcal{G}) = \sup_{g\in\mathcal{G}} \|\mathbb{E}_{X\sim G}[(\mathcal{T}g)(X)]\|_2$.

How to choose $\mathcal{T}$? Langevin Stein operator

$$(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x) \rangle + \langle \nabla, g(x) \rangle.$$

How to choose $\mathcal{G}$? Ideally, want

- Discriminability: $\mathbb{S}(Q,P,\mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q,P,\mathcal{G})$ can be efficiently computed.

[2] [Gorham and Mackey, 2015]

Kernelized Stein Discrepancy

Let $\mathcal{H}_k$ be a scalar-valued RKHS with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and let $\mathcal{T}$ be the Langevin Stein operator [3].

Langevin Kernlized Stein Discrepancy (KSD)

Choosing $\mathcal{G}_k := \{g = (g_1, \ldots, g_d) : \|v_2\|_2 \leq 1 \text{ for } v_j := \|g_j\|_k\}$ leads to the *Langevin KSD* [4]:

$$\mathrm{KSD}_k(Q, P) := \mathbb{S}(Q, P, \mathcal{G}_k) = \sqrt{\mathbb{E}_{X, X' \sim Q}[k_P(X, X')]},$$

where the *Stein reproducing kernel* is

$$\begin{aligned} k_P(X, X') := {} & \langle \nabla_x, \nabla_{x'} k(x, x') \rangle + \langle \nabla_x k(x, x'), \nabla_{x'} \log p(x') \rangle \\ & + \langle \nabla_{x'} k(x, x'), \nabla_x \log p(x) \rangle + k(x, x') \langle \nabla_x \log p(x), \nabla_{x'} \log p(x) \rangle. \end{aligned}$$

[3] Other choices of $\mathcal{T}$ [Gorham et al., 2019] and $\mathcal{G}$ [Gorham and Mackey, 2015] are possible.

[4] Liu et al. [2016], Chwialkowski et al. [2016]

Application 1: Goodness-of-Fit Test

Setup: Let P have continuously differentiable density $p = p^*/Z$ supported on $\mathcal{X} \subset \mathbb{R}^d$, where Z is a normalising constant (unknown), and p^* *can be evaluated pointwise.*

Goodness-of-fit test

Given $\{x_i\}_{i=1}^n$ drawn from another distribution Q supported on $\mathcal{X}$, is $Q = P$?

Want to test $H_0 : Q = P$ against $H_1 : Q \neq P$.

Equivalently, $H_0 : \mathrm{KSD}_k(Q, P) = 0$ against $H_1 : \mathrm{KSD}_k(Q, P) \neq 0$.

KSD test[5]: Compute $\widehat{\mathrm{KSD}}_k(Q, P)$ a test statistic, and reject for large value of $\widehat{\mathrm{KSD}}_k(Q, P)$.

To compute the rejection threshold (or the p-value), we need to know the distribution of $\widehat{\mathrm{KSD}}_k(Q, P)$ under H_0.

[5] Liu et al. [2016], Chwialkowski et al. [2016]

Application 1: Goodness-of-Fit Test

Setup: Let P have continuously differentiable density $p = p^*/Z$ supported on $\mathcal{X} \subset \mathbb{R}^d$, where Z is a normalising constant (unknown), and p^* *can be evaluated pointwise*.

Goodness-of-fit test

Given $\{x_i\}_{i=1}^n$ drawn from another distribution Q supported on $\mathcal{X}$, is $Q = P$?

Want to test $H_0 : Q = P$ against $H_1 : Q \neq P$.

Equivalently, $H_0 : \mathrm{KSD}_k(Q, P) = 0$ against $H_1 : \mathrm{KSD}_k(Q, P) \neq 0$.

KSD test[5]: Compute $\widehat{\mathrm{KSD}}_k(Q, P)$ a test statistic, and reject for large value of $\widehat{\mathrm{KSD}}_k(Q, P)$.

To compute the rejection threshold (or the p-value), we need to know the distribution of $\widehat{\mathrm{KSD}}_k(Q, P)$ under H_0.

[5] Liu et al. [2016], Chwialkowski et al. [2016]

Goodness-of-Fit Test

Theorem (Asymptotic distributions; informal)

Assume $\mathbb{E}_{X,X'\sim Q}[k_P(X,X')^2] < \infty$. As $n \to \infty$,

- If $Q \neq P$, then

$$\sqrt{n}(\widehat{\mathrm{KSD}}_k(Q,P)^2 - \mathrm{KSD}_k(Q,P)^2) \to_d \mathcal{N}(0, \sigma_k^2),$$

  where $\sigma_k^2 \coloneqq \mathrm{Var}(\mathbb{E}_{X'\sim Q}[k_P(X,X')])$, and $\sigma_k > 0$.

- If $Q = P$, then

$$n\widehat{\mathrm{KSD}}_k(Q,P)^2 \to_d \textstyle\sum_{j=1}^{\infty} c_j(Z_j^2 - 1) =: W_{H_0},$$

  where $Z_j \sim \mathcal{N}(0,1)$ i.i.d., and $\{c_j\}_j$ are the eigenvalues of k_P under Q.

The distribution of W_{H_0} is intractable, but can be approximated using a *wild bootstrap* procedure.

Goodness-of-Fit Tests

KSD Test

Given $\{x_i\}_{i=1}^n \sim Q$ and a test level $\alpha > 0$,

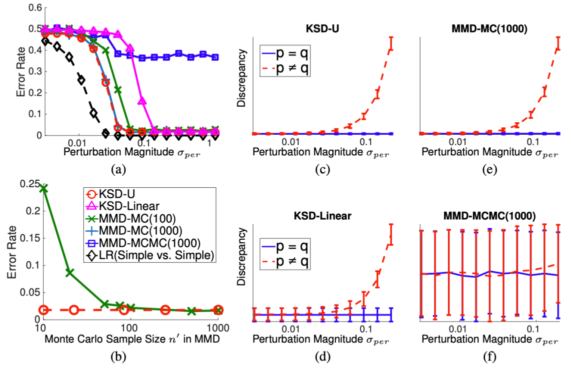
1. For $b = 1, \ldots, B$, compute bootstrap samples

$$\widehat{\mathrm{KSD}}_{k,b}^2 := \frac{1}{n^2} \sum_{1 \leq i \neq j \leq n} (W_i^b - 1)(W_j^b - 1) k_P(x_i, x_j),$$

   where $W^b = (W_1^b, \ldots, W_n^b) \sim \mathrm{Multinom}(n, (1/n, \ldots, 1/n))$.

2. Reject if $\widehat{\mathrm{KSD}}_k^2 \geq \hat{\gamma}_\alpha$, where $\hat{\gamma}_\alpha$ is the $(1-\alpha)$-quantile of $\{\widehat{\mathrm{KSD}}_{k,b}^2\}_{b=1}^B$.

Example — Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)



Target P: $p(x) = \sum_{h \in \{\pm 1\}^{d_h}} p(x, h)$, where

$$p(x, h) \propto \exp\left(\frac{1}{2} x^\intercal B h + b^\intercal x + c^\intercal h - \frac{1}{2} \|x\|_2^2\right).$$

Candidate Q: same as p but with noise injected into the entries of B.

Application 2: Sample Quality Measure

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure $Q_n = n^{-1}\sum_{i=1}^{n}\delta_{x_i}$ based on sample $\{x_i\}_{i=1}^{n}$.

Questions:

1. Does $Q_n \to_d P$ imply $\text{KSD}_k(Q_n, P) \to \text{KSD}_k(P, P) = 0$?
2. Does $\text{KSD}_k(Q_n, P) \to 0$ imply $Q_n \to_d P$?

Theorem [Gorham and Mackey, 2017]

1. If $\nabla \log p$ is Lipschitz and k is twice continuously differentiable, then $d_W(Q_n, P) \to 0 \implies \text{KSD}_k(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is *distantly dissipative*, and $k(x, y) = \Phi(x - y)$ for some twice continuously differentiable Φ with non-vanishing Fourier transform. If $(Q_n)_{n\geq 1}$ satisfies a tail condition (*uniform tightness*), then $\text{KSD}_k(Q_n, P) \to 0 \implies Q_n \to_d P$.

Application 2: Sample Quality Measure

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ based on sample $\{x_i\}_{i=1}^n$.

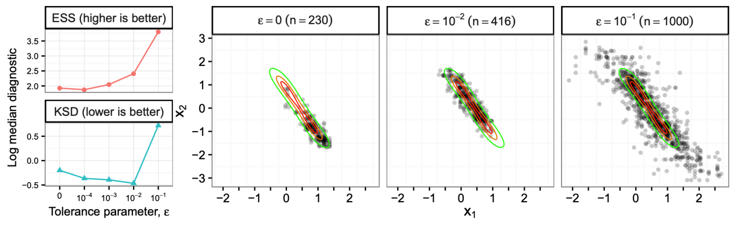
Questions:

1. Does $Q_n \to_d P$ imply $\mathrm{KSD}_k(Q_n, P) \to \mathrm{KSD}_k(P, P) = 0$?
2. Does $\mathrm{KSD}_k(Q_n, P) \to 0$ imply $Q_n \to_d P$?

Theorem [Gorham and Mackey, 2017]

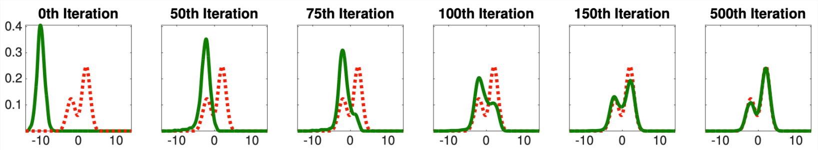
1. If $\nabla \log p$ is Lipschitz and k is twice continuously differentiable, then $d_W(Q_n, P) \to 0 \implies \mathrm{KSD}_k(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is *distantly dissipative*, and $k(x, y) = \Phi(x - y)$ for some twice continuously differentiable Φ with non-vanishing Fourier transform. If $(Q_n)_{n\geq 1}$ satisfies a tail condition (*uniform tightness*), then $\mathrm{KSD}_k(Q_n, P) \to 0 \implies Q_n \to_d P$.

Example — Hyperparamter Selection



Use KSD as a **sample quality measure** to select hyperparameters of a MCMC sampler (Stochastic Graident Fisher Scoring), with comparisons against a classical metric, ESS (Effective Sample Size).

Other Applications



SVGD [6]: Learning a target distribution by iteratively transporting particles drawn from an initial distribution.

And many more! See Anastasiou et al. [2021].

[6] Liu and Wang [2016]

References

A. Anastasiou and G. Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017.

A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, et al. Stein's Method Meets Statistics: A Review of Some Recent Developments. *arXiv preprint arXiv:2105.03481*, 2021.

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/chwialkowski16.html.

W. Gong, Y. Li, and J. M. Hernández-Lobato. Sliced kernelized stein discrepancy. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=t0TaKv0Gx6Z.

J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/698d51a19d8a121ce581499d7b701668-Paper.pdf.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884 – 2928, 2019. doi: 10.1214/19-AAP1467. URL https://doi.org/10.1214/19-AAP1467.

W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, and R. Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pages 3732–3747. PMLR, 2020.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Q. Liu, J. Lee, and M. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/liub16.html.

X. Liu, H. Zhu, J.-F. Ton, G. Wynne, and A. Duncan. Grassmann Stein Variational Gradient Descent. *arXiv preprint arXiv:2202.03297*, 2022.

G. Mijoule, G. Reinert, and Y. Swan. Stein's density method for multivariate continuous distributions. *arXiv preprint arXiv:2101.05079*, 2021.

A. Mira, R. Solgi, and D. Imparato. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23 (5):653–662, 2013.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

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

G. Reinert. Couplings for normal approximations with Stein's method. *DIMACS Ser. Discrete Math. Theoret. Comput. Sci*, 41: 193–207, 1998.