Kernlized Stein Discrepancy

Xing Liu
February 13, 2023
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*

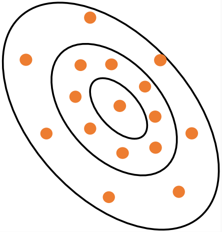
Table of Contents

Motivation

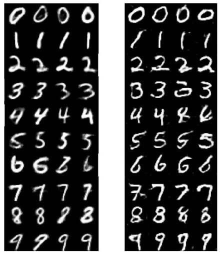
Let Q, P be probability measures on $\mathcal{X} \subset \mathbb{R}^d$.

- P admits a density $p = p^*/Z$, where Z is an unknown normalising constant.
- Samples are observed from Q only.

Problem of interest: How to quantify the discrepancy between P and another probability measure Q ?



P: target distribution
Q: MCMC samples



P: a generative model
Q: true images

Integral Probability Metrics (IPM)[1]

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

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

- 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}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

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

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

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

[1] [Müller, 1997]

Integral Probability Metrics (IPM)[1]

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

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

- 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}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

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

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

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

[1] [Müller, 1997]

Integral Probability Metrics (IPM)[1]

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

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

- 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}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

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

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

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

[1] [Müller, 1997]

Motivation — Quantifying Discrepancy

Integral Probability Metrics (IPM)[1]

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

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

- 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}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

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

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

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

[1] [Müller, 1997]

Motivation — Quantifying Discrepancy

Integral Probability Metrics (IPM)[1]

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

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

- 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}:
  $\mathcal{H}_{bw} = \{h \in \mathcal{H}_W : h \text{ is bounded}\}$

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

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

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

[1] [Müller, 1997]

Kernelized Stein Discrepancy

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

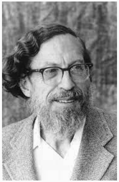
Stein's Identity

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

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

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)



Charles Stein

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

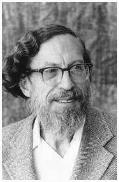
Stein's Identity

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

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

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)



Charles Stein

A Discrepancy based on Stein's Method

Setup: P, Q two probability measures. P has unnormalised density p that is continuously differentiable.

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

~~Kernelized~~ Stein Discrepancy

Given a Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G}$, the Stein discrepancy is:

$$\mathbb{S}(Q,P,\mathcal{G}) = \sup_{g\in\{\mathcal{T}g:\ g\in\mathcal{G}\}} \|\mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)]\|_2.$$

Ideally, we want

- Separation: $\mathbb{S}(Q,P,\mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q,P,\mathcal{G})$ can be efficiently computed even when the normalising constant of p is unknown and sampling from P is infeasible.

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 $\mathcal{G}$? Reproducing Kernel Hilbert Spaces (RKHS)!

A Discrepancy based on Stein's Method

Setup: P, Q two probability measures. P has unnormalised density p that is continuously differentiable.

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

~~Kernelized~~ Stein Discrepancy

Given a Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G}$, the Stein discrepancy is:

$$\mathbb{S}(Q,P,\mathcal{G}) = \sup_{g \in \{\mathcal{T}g:\ g\in\mathcal{G}\}} \|\mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)]\|_2.$$

Ideally, we want

- Separation: $\mathbb{S}(Q,P,\mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q,P,\mathcal{G})$ can be efficiently computed even when the normalising constant of p is unknown and sampling from P is infeasible.

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 $\mathcal{G}$? Reproducing Kernel Hilbert Spaces (RKHS)!

A Discrepancy based on Stein's Method

Setup: P, Q two probability measures. P has unnormalised density p that is continuously differentiable.

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

~~Kernelized~~ Stein Discrepancy

Given a Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G}$, the Stein discrepancy is:

$$\mathbb{S}(Q, P, \mathcal{G}) = \sup_{g \in \{\mathcal{T}g:\ g \in \mathcal{G}\}} \|\mathbb{E}_{X \sim Q}[(\mathcal{T}g)(X)]\|_2.$$

Ideally, we want

- Separation: $\mathbb{S}(Q, P, \mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q, P, \mathcal{G})$ can be efficiently computed even when the normalising constant of p is unknown and sampling from P is infeasible.

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 $\mathcal{G}$? Reproducing Kernel Hilbert Spaces (RKHS)!

A Discrepancy based on Stein's Method

Setup: P, Q two probability measures. P has unnormalised density p that is continuously differentiable.

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

~~Kernelized~~ Stein Discrepancy

Given a Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G}$, the Stein discrepancy is:

$$\mathbb{S}(Q, P, \mathcal{G}) = \sup_{g \in \{\mathcal{T} g: \, g \in \mathcal{G}\}} \|\mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)]\|_2.$$

Ideally, we want

- Separation: $\mathbb{S}(Q, P, \mathcal{G}) = 0 \iff Q = P$
- Computability: $\mathbb{S}(Q, P, \mathcal{G})$ can be efficiently computed even when the normalising constant of p is unknown and sampling from P is infeasible.

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 $\mathcal{G}$? Reproducing Kernel Hilbert Spaces (RKHS)!

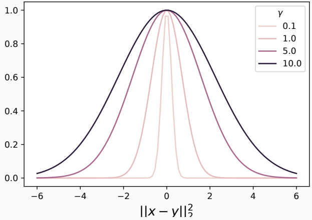
Reproducing Kernel Hilbert Spaces (RKHS)

Reproducing kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

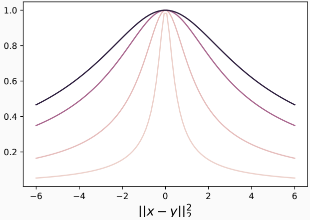
- Symmetric: $k(x, y) = k(y, x)$.
- Positive definite: For any $n \in \mathbb{Z}_+$, $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$, $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$.

RKHS: A Hilbert space $\mathcal{H}_k$ is a RKHS associated with k if

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.
- Reproducing property: $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.



Radial basis function (RBF):
$k(x, y) = \exp\left(-\frac{1}{\gamma}\|x - y\|_2^2\right)$



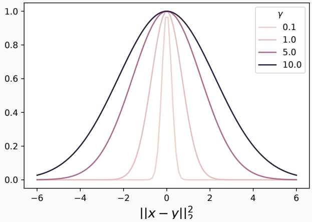
Inverse multi-quadric (IMQ):
$k(x, y) = \left(1 + \frac{1}{\gamma}\|x - y\|_2^2\right)^{-1/2}$

Reproducing kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

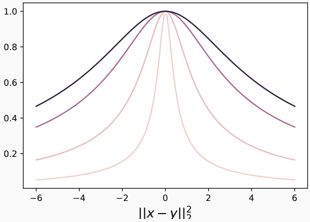
- Symmetric: $k(x, y) = k(y, x)$.
- Positive definite: For any $n \in \mathbb{Z}_+$, $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$, $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$.

RKHS: A Hilbert space $\mathcal{H}_k$ is a RKHS associated with k if

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.
- Reproducing property: $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.



Radial basis function (RBF):
$k(x, y) = \exp\left(-\frac{1}{\gamma}\|x - y\|_2^2\right)$



Inverse multi-quadric (IMQ):
$k(x, y) = \left(1 + \frac{1}{\gamma}\|x - y\|_2^2\right)^{-1/2}$

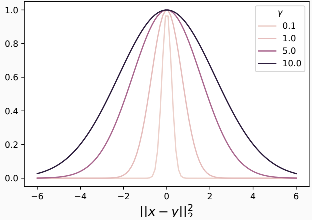
Reproducing Kernel Hilbert Spaces (RKHS)

Reproducing kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

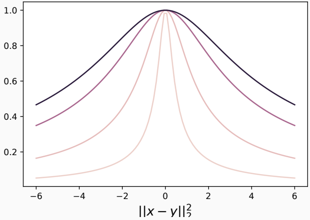
- Symmetric: $k(x, y) = k(y, x)$.
- Positive definite: For any $n \in \mathbb{Z}_+$, $x_1, \dots, x_n \in \mathcal{X}$ and $c_1, \dots, c_n \in \mathbb{R}$, $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$.

RKHS: A Hilbert space $\mathcal{H}_k$ is a RKHS associated with k if

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.
- Reproducing property: $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.



Radial basis function (RBF):
$$k(x, y) = \exp\left(-\frac{1}{\gamma}\|x - y\|_2^2\right)$$



Inverse multi-quadric (IMQ):
$$k(x, y) = \left(1 + \frac{1}{\gamma}\|x - y\|_2^2\right)^{-1/2}$$

Kernelized Stein Discrepancy

(Langevin) Kernelized Stein Discrepancy (KSD)[2]

Choosing $\mathcal{G}_k^d := \times_{j=1}^d \mathcal{G}_k$ for $\mathcal{G}_k :=$ unit-ball in a RKHS $\mathcal{H}_k$, the KSD is

$$\mathbb{D}(Q,P) := \mathbb{S}^2(Q,P,\mathcal{G}_k^d) = \mathbb{E}_{X,X'\sim Q}[k_P(X,X')],$$

where

$$\begin{aligned} k_P(x,x') &:= k(x,x')\langle s_p(x), s_p(x')\rangle + \langle \nabla_x k(x,x'), s_p(x')\rangle \\ &\quad + \langle \nabla_{x'} k(x,x'), s_p(x)\rangle + \langle \nabla_x, \nabla_{x'} k(x,x')\rangle, \end{aligned}$$

and $s_p(x) := \nabla_x \log p(x)$.

- k_P: Stein reproducing kernel.
- $\mathbb{D}(Q,P) \geq 0$ and $\mathbb{D}(Q,P) = 0 \iff Q = P$.
- k_P is computable even if p is only known up to a normalisation: $s_p(x) = \nabla_x \log p(x) = \nabla_x \log(p^*(x)/Z) = \nabla_x \log p^*(x) - \cancel{\nabla_x Z}$.
- Estimation: given i.i.d. $\{X_i\}_{i=1}^n \sim Q$,

$$\mathbb{D}_n := \sum_{1\leq i\neq j}^n k_P(X_i, X_j)$$

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

Kernelized Stein Discrepancy

(Langevin) Kernelized Stein Discrepancy (KSD)[2]

Choosing $\mathcal{G}_k^d := \times_{j=1}^d \mathcal{G}_k$ for $\mathcal{G}_k :=$ unit-ball in a RKHS $\mathcal{H}_k$, the KSD is

$$\mathbb{D}(Q, P) := \mathbb{S}^2(Q, P, \mathcal{G}_k^d) = \mathbb{E}_{X,X' \sim Q}[k_P(X, X')],$$

where

$$\begin{aligned} k_P(x, x') &:= k(x, x')\langle s_p(x), s_p(x')\rangle + \langle \nabla_x k(x, x'), s_p(x')\rangle \\ &\quad + \langle \nabla_{x'} k(x, x'), s_p(x)\rangle + \langle \nabla_x, \nabla_{x'} k(x, x')\rangle, \end{aligned}$$

and $s_p(x) := \nabla_x \log p(x)$.

- k_P: Stein reproducing kernel.
- $\mathbb{D}(Q, P) \geq 0$ and $\mathbb{D}(Q, P) = 0 \iff Q = P$.
- k_P is computable even if p is only known up to a normalisation: $s_p(x) = \nabla_x \log p(x) = \nabla_x \log(p^*(x)/Z) = \nabla_x \log p^*(x) - \cancel{\nabla_x Z}$.
- Estimation: given i.i.d. $\{X_i\}_{i=1}^n \sim Q$,

$$\mathbb{D}_n := \sum_{1 \leq i \neq j}^n k_P(X_i, X_j)$$

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

(Langevin) Kernelized Stein Discrepancy (KSD)[2]

Choosing $\mathcal{G}_k^d := \times_{j=1}^d \mathcal{G}_k$ for $\mathcal{G}_k :=$ unit-ball in a RKHS $\mathcal{H}_k$, the KSD is

$$\mathbb{D}(Q, P) := \mathbb{S}^2(Q, P, \mathcal{G}_k^d) = \mathbb{E}_{X,X' \sim Q}[k_P(X, X')],$$

where

$$\begin{aligned} k_P(x, x') &:= k(x, x')\langle s_p(x), s_p(x')\rangle + \langle \nabla_x k(x, x'), s_p(x')\rangle \\ &\quad + \langle \nabla_{x'} k(x, x'), s_p(x)\rangle + \langle \nabla_x, \nabla_{x'} k(x, x')\rangle, \end{aligned}$$

and $s_p(x) := \nabla_x \log p(x)$.

- k_P: Stein reproducing kernel.
- $\mathbb{D}(Q, P) \geq 0$ and $\mathbb{D}(Q, P) = 0 \iff Q = P$.
- k_P is computable even if p is only known up to a normalisation: $s_p(x) = \nabla_x \log p(x) = \nabla_x \log(p^*(x)/Z) = \nabla_x \log p^*(x) - \cancel{\nabla_x Z}$.
- Estimation: given i.i.d. $\{X_i\}_{i=1}^n \sim Q$,

$$\mathbb{D}_n := \sum_{1 \leq i \neq j}^n k_P(X_i, X_j)$$

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

Kernelized Stein Discrepancy

(Langevin) Kernelized Stein Discrepancy (KSD)[2]

Choosing $\mathcal{G}_k^d := \times_{j=1}^d \mathcal{G}_k$ for $\mathcal{G}_k :=$ unit-ball in a RKHS $\mathcal{H}_k$, the KSD is

$$\mathbb{D}(Q, P) := \mathbb{S}^2(Q, P, \mathcal{G}_k^d) = \mathbb{E}_{X,X' \sim Q}[k_P(X, X')],$$

where

$$\begin{aligned} k_P(x, x') &:= k(x, x')\langle s_p(x), s_p(x')\rangle + \langle \nabla_x k(x, x'), s_p(x')\rangle \\ &\quad + \langle \nabla_{x'} k(x, x'), s_p(x)\rangle + \langle \nabla_x, \nabla_{x'} k(x, x')\rangle, \end{aligned}$$

and $s_p(x) := \nabla_x \log p(x)$.

- k_P: Stein reproducing kernel.
- $\mathbb{D}(Q, P) \geq 0$ and $\mathbb{D}(Q, P) = 0 \iff Q = P$.
- k_P is computable even if p is only known up to a normalisation: $s_p(x) = \nabla_x \log p(x) = \nabla_x \log(p^*(x)/Z) = \nabla_x \log p^*(x) - \cancel{\nabla_x Z}$.
- Estimation: given i.i.d. $\{X_i\}_{i=1}^n \sim Q$,

$$\mathbb{D}_n := \sum_{1 \leq i \neq j}^n k_P(X_i, X_j)$$

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

Kernelized Stein Discrepancy

(Langevin) Kernelized Stein Discrepancy (KSD)[2]

Choosing $\mathcal{G}_k^d := \times_{j=1}^d \mathcal{G}_k$ for $\mathcal{G}_k :=$ unit-ball in a RKHS $\mathcal{H}_k$, the KSD is

$$\mathbb{D}(Q,P) := \mathbb{S}^2(Q,P,\mathcal{G}_k^d) = \mathbb{E}_{X,X'\sim Q}[k_P(X,X')],$$

where

$$\begin{aligned} k_P(x,x') &:= k(x,x')\langle s_p(x), s_p(x')\rangle + \langle \nabla_x k(x,x'), s_p(x')\rangle \\ &\quad + \langle \nabla_{x'} k(x,x'), s_p(x)\rangle + \langle \nabla_x, \nabla_{x'} k(x,x')\rangle, \end{aligned}$$

and $s_p(x) := \nabla_x \log p(x)$.

- k_P: Stein reproducing kernel.
- $\mathbb{D}(Q,P) \geq 0$ and $\mathbb{D}(Q,P) = 0 \iff Q = P$.
- k_P is computable even if p is only known up to a normalisation: $s_p(x) = \nabla_x \log p(x) = \nabla_x \log(p^*(x)/Z) = \nabla_x \log p^*(x) - \cancel{\nabla_x Z}$.
- Estimation: given i.i.d. $\{X_i\}_{i=1}^n \sim Q$,

$$\mathbb{D}_n := \sum_{1\leq i\neq j}^n k_P(X_i, X_j)$$

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

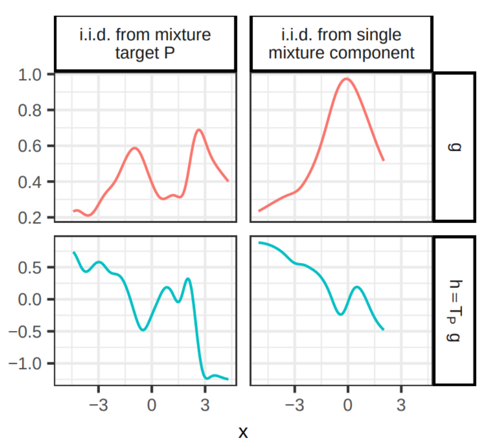


Figure credit: [Gorham and Mackey, 2017]

Convergence Determination

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure.

Questions:

1. Does $Q_n \to_d P$ imply $\mathbb{D}(Q_n, P) \to \mathbb{D}(P, P) = 0$?
2. Does $\mathbb{D}(Q_n, P) \to 0$ 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 $d_W(Q_n, P) \to 0 \implies \mathbb{D}(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is distantly dissipative (a relaxation of log-concavity). If either an IMQ kernel is used or $(Q_n)_{n\geq 1}$ is uniformly tight (a tail condition). Then $\mathbb{D}(Q_n, P) \to 0 \implies Q_n \to_d P$.

General conditions under which $Q_n \to_d P \iff \mathbb{D}(Q_n, P) \to 0$: [Hodgkinson et al., 2020, Barp et al., 2022]

Convergence Determination

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure.

Questions:

1. Does $Q_n \to_d P$ imply $\mathbb{D}(Q_n, P) \to \mathbb{D}(P, P) = 0$?
2. Does $\mathbb{D}(Q_n, P) \to 0$ 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 $d_W(Q_n, P) \to 0 \implies \mathbb{D}(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is distantly dissipative (a relaxation of log-concavity). If either an IMQ kernel is used or $(Q_n)_{n\geq 1}$ is uniformly tight (a tail condition). Then $\mathbb{D}(Q_n, P) \to 0 \implies Q_n \to_d P$.

General conditions under which $Q_n \to_d P \iff \mathbb{D}(Q_n, P) \to 0$: [Hodgkinson et al., 2020, Barp et al., 2022]

Convergence Determination

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure.

Questions:

1. Does $Q_n \to_d P$ imply $\mathbb{D}(Q_n, P) \to \mathbb{D}(P, P) = 0$?
2. Does $\mathbb{D}(Q_n, P) \to 0$ 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 $d_W(Q_n, P) \to 0 \implies \mathbb{D}(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is distantly dissipative (a relaxation of log-concavity). If either an IMQ kernel is used or $(Q_n)_{n\geq 1}$ is uniformly tight (a tail condition). Then $\mathbb{D}(Q_n, P) \to 0 \implies Q_n \to_d P$.

General conditions under which $Q_n \to_d P \iff \mathbb{D}(Q_n, P) \to 0$: [Hodgkinson et al., 2020, Barp et al., 2022]

Convergence Determination

Setup: P same as before, and $\{Q_n\}_{n\geq 1}$ is a sequence of empirical measure.

Questions:

1. Does $Q_n \to_d P$ imply $\mathbb{D}(Q_n, P) \to \mathbb{D}(P, P) = 0$?
2. Does $\mathbb{D}(Q_n, P) \to 0$ 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 $d_W(Q_n, P) \to 0 \implies \mathbb{D}(Q_n, P) \to 0$.
2. Assume $\nabla \log p$ is distantly dissipative (a relaxation of log-concavity). If either an IMQ kernel is used or $(Q_n)_{n\geq 1}$ is uniformly tight (a tail condition). Then $\mathbb{D}(Q_n, P) \to 0 \implies Q_n \to_d P$.

General conditions under which $Q_n \to_d P \iff \mathbb{D}(Q_n, P) \to 0$: [Hodgkinson et al., 2020, Barp et al., 2022]

Application 1: Goodness-of-Fit Testing

Application 1: Goodness-of-Fit Testing

Goodness-of-Fit Testing

Given sample $\{X_i\}_{i=1}^n$ drawn independently from Q, test

$$H_0 : Q = P \text{ vs. } H_1 : Q \neq P \ .$$

$$\iff H_0 : \mathbb{D}(Q, P) = 0 \text{ vs. } H_1 : \mathbb{D}(Q, P) \neq 0 \ .$$

KSD test[3]: Compute test statistic $\mathbb{D}_n$ using $\{X_i\}_{i=1}^n$, and reject for large values.

Given significance level $\alpha \in (0, 1)$, the rejection threshold $\hat{q}_{1-\alpha}$ should satisfy

$$\text{Type-I error } := \mathbb{P}_P(\hat{\mathbb{D}}_n \geq \hat{q}_{1-\alpha}) \leq \alpha \ .$$

To compute $\hat{q}_{1-\alpha}$, we need to know the distribution of $\mathbb{D}_n$ under H_0.

- Intractable, but can be approximated using bootstraping.

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

Goodness-of-Fit Testing

Given sample $\{X_i\}_{i=1}^n$ drawn independently from Q, test

$$H_0 : Q = P \text{ vs. } H_1 : Q \neq P \,.$$

$$\iff H_0 : \mathbb{D}(Q, P) = 0 \text{ vs. } H_1 : \mathbb{D}(Q, P) \neq 0 \,.$$

KSD test[3]: Compute test statistic $\mathbb{D}_n$ using $\{X_i\}_{i=1}^n$, and reject for large values.

Given significance level $\alpha \in (0, 1)$, the rejection threshold $\hat{q}_{1-\alpha}$ should satisfy

$$\text{Type-I error } := \mathbb{P}_P(\hat{\mathbb{D}}_n \geq \hat{q}_{1-\alpha}) \leq \alpha \,.$$

To compute $\hat{q}_{1-\alpha}$, we need to know the distribution of $\mathbb{D}_n$ under H_0.

- Intractable, but can be approximated using bootstraping.

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

Goodness-of-Fit Testing

Given sample $\{X_i\}_{i=1}^n$ drawn independently from Q, test

$$H_0 : Q = P \text{ vs. } H_1 : Q \neq P \ .$$

$$\iff H_0 : \mathbb{D}(Q, P) = 0 \text{ vs. } H_1 : \mathbb{D}(Q, P) \neq 0 \ .$$

KSD test[3]: Compute test statistic $\mathbb{D}_n$ using $\{X_i\}_{i=1}^n$, and reject for large values.

Given significance level $\alpha \in (0, 1)$, the rejection threshold $\hat{q}_{1-\alpha}$ should satisfy

$$\text{Type-I error } := \mathbb{P}_P(\hat{\mathbb{D}}_n \geq \hat{q}_{1-\alpha}) \leq \alpha \ .$$

To compute $\hat{q}_{1-\alpha}$, we need to know the distribution of $\mathbb{D}_n$ under H_0.

- Intractable, but can be approximated using bootstraping.

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

Goodness-of-Fit Testing

Given sample $\{X_i\}_{i=1}^n$ drawn independently from Q, test

$$H_0 : Q = P \text{ vs. } H_1 : Q \neq P \,.$$

$$\iff H_0 : \mathbb{D}(Q, P) = 0 \text{ vs. } H_1 : \mathbb{D}(Q, P) \neq 0 \,.$$

KSD test[3]: Compute test statistic $\mathbb{D}_n$ using $\{X_i\}_{i=1}^n$, and reject for large values.

Given significance level $\alpha \in (0, 1)$, the rejection threshold $\hat{q}_{1-\alpha}$ should satisfy

$$\text{Type-I error } := \mathbb{P}_P(\hat{\mathbb{D}}_n \geq \hat{q}_{1-\alpha}) \leq \alpha \,.$$

To compute $\hat{q}_{1-\alpha}$, we need to know the distribution of $\mathbb{D}_n$ under H_0.

- Intractable, but can be approximated using bootstraping.

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

Algorithm (KSD Test)

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

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

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

   where $\epsilon_1^b, \ldots, \epsilon_n^b$ are i.i.d. Rademacher r.v. in $\{-1, 1\}$.

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

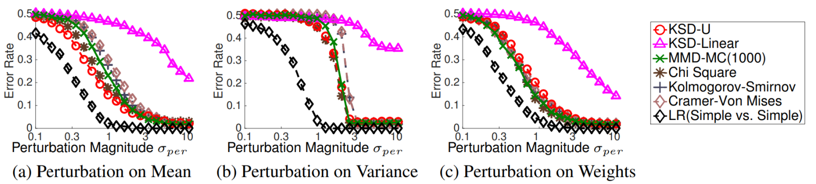


Figure credit: [Liu et al., 2016]

- $P = \sum_{k=1}^{5} w_k \mathcal{N}(\mu_k, \sigma^2)$, where $w_k = \frac{1}{5}$, $\sigma^2 = 1$, and $\mu_k \in [0, 10]$.
- Q = same as P but with Gaussian noise injected into μ_k, σ^2 and $\log w_k$.

Blindness of KSD

$\mathbb{D}(Q, P) \approx 0$ when Q and P are multi-modal distributions with well-separated modes. $\longrightarrow$ KSD test power $\approx \alpha$.

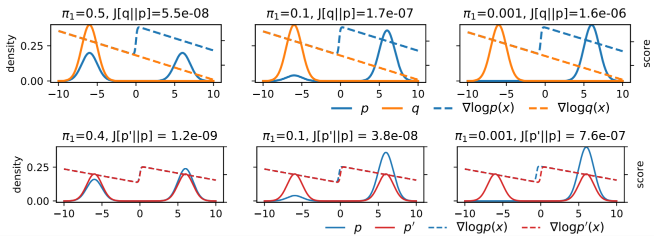


Figure credit: [Wenliang and Kanagawa, 2020]

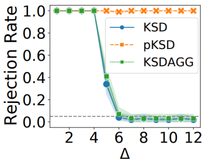


Figure credit: [Liu et al., 2022a]

Application 2: Sample Quality Quantification

Application 2: Sample Quality Quantification

Setup: P same as before, and $\{X_i\}_{i=1}^n$ an i.i.d. sample from some Q.

- E.g., Q is a MCMC sampler targeting P, or a generative model.

Questions: How to quantify how well $\{X_i\}_{i=1}^n$ fits P ?

- Classical diagnostics such as Effective Sample Size and the Gelman-Rubin statistic do **not** account for asymptotic bias.
- KSD is a natural metric due to its convergence-determining property!

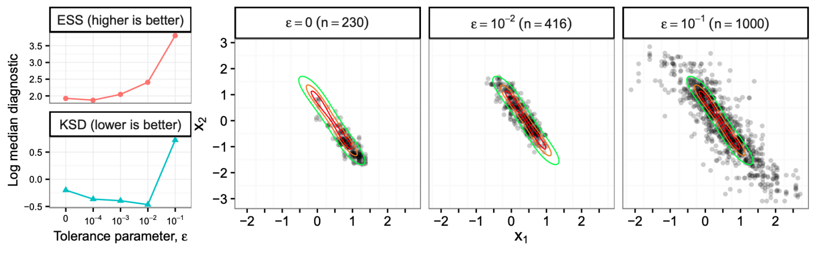
Application 2: Sample Quality Quantification

Setup: P same as before, and $\{X_i\}_{i=1}^n$ an i.i.d. sample from some Q.

- E.g., Q is a MCMC sampler targeting P, or a generative model.

Questions: How to quantify how well $\{X_i\}_{i=1}^n$ fits P ?

- Classical diagnostics such as Effective Sample Size and the Gelman-Rubin statistic do **not** account for asymptotic bias.
- KSD is a natural metric due to its convergence-determining property!



Using KSD to select hyperparameters of a MCMC sampler, with comparisons against ESS (Effective Sample Size). Figure credit: [Gorham and Mackey, 2017]

Application 3: Sample Approximation

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^\infty(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_\# Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{ -\tfrac{d}{d\epsilon} \mathrm{KL}(T_\# Q \| P)|_{\epsilon=0} \} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[\mathcal{T} g(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T} k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^\infty(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_\# Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{-\tfrac{d}{d\epsilon} \mathrm{KL}(T_\# Q \| P)|_{\epsilon=0}\} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T} k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^\infty(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_\# Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{-\tfrac{d}{d\epsilon} \mathrm{KL}(T_\# Q \| P)|_{\epsilon=0}\} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T} k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^\infty(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_\# Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{ -\tfrac{d}{d\epsilon} \mathrm{KL}(T_\# Q \| P)|_{\epsilon=0} \} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T} k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^{\infty}(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_{\#}Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{-\tfrac{d}{d\epsilon}\mathrm{KL}(T_{\#}Q \| P)|_{\epsilon=0}\} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[(\mathcal{T}g)(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T}k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Objective: Sampling from P with continuously differentiable density p.

Idea:

- Initialise $X \sim Q$
- Iteratively apply a map $T(X) = X + \epsilon g(X)$ so that $T^\infty(X) \sim P$.

Choose g in some function class $\mathcal{H}$ to maximally decrease $\mathrm{KL}(T_\# Q \| P)$:

$$\sup_{g \in \mathcal{H}} \{ -\tfrac{d}{d\epsilon} \mathrm{KL}(T_\# Q \| P)|_{\epsilon=0} \} = \sup_{g \in \mathcal{H}} \mathbb{E}_{X \sim Q}[(\mathcal{T} g)(X)] \qquad (*)$$

[Liu and Wang, 2016]:

- $(*) = \mathbb{S}(Q, P, \mathcal{H})$, the Stein discrepancy objective!
- Hence, the optimal g^* is the maximiser in $(*)$.
- Choosing $\mathcal{H}$ to be a RKHS, g^* has an analytical form:

$$g^*(\cdot) = \mathbb{E}_{X \sim Q}[\mathcal{T} k(\cdot, x)] = \mathbb{E}_{X \sim Q}[\underbrace{\nabla \log p(X) k(X, \cdot)}_{attraction} + \underbrace{\nabla_x k(X, \cdot)}_{repulsion}]$$

- Using g^* in the map $T \longrightarrow$ Stein variational gradient descent (SVGD).

Application 3: Stein Variational Gradient Descent

Stein Variational Gradient Descent

- Given $X_1, \ldots, X_n \sim Q$ i.i.d., and $\epsilon > 0$.
- For $t = 1, 2, \ldots$, set

$$X_i^{(t)} \; = \; X_i^{(t-1)} + \frac{\epsilon}{n} \sum_{j=1}^{n} k(X_i^{(t)}, X_j^{(t)}) \nabla \log p(X_j^{(t)}) + \nabla_X k(X_i^{(t)}, X_j^{(t)}) \; .$$

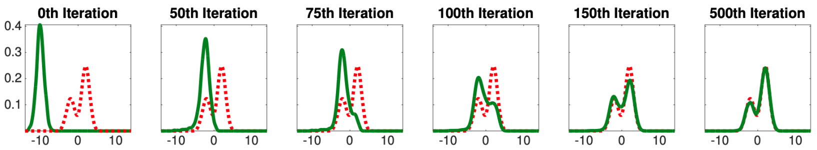


Figure credit: [Liu and Wang, 2016]

- Deterministic interacting particle system.
- Both asymptotic [Liu, 2017] and non-asymptotic theories [Liu and Wang, 2018] are available.

Application 3: Stein Variational Gradient Descent

Stein Variational Gradient Descent

- Given $X_1, \ldots, X_n \sim Q$ i.i.d., and $\epsilon > 0$.
- For $t = 1, 2, \ldots$, set

$$X_i^{(t)} = X_i^{(t-1)} + \frac{\epsilon}{n} \sum_{j=1}^{n} k(X_i^{(t)}, X_j^{(t)}) \nabla \log p(X_j^{(t)}) + \nabla_X k(X_i^{(t)}, X_j^{(t)}) \; .$$

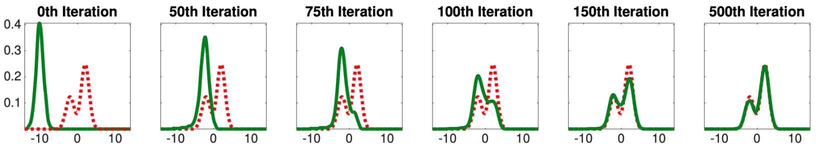


Figure credit: [Liu and Wang, 2016]

- Deterministic interacting particle system.
- Both asymptotic [Liu, 2017] and non-asymptotic theories [Liu and Wang, 2018] are available.

Stein Variational Gradient Descent

- Given $X_1, \ldots, X_n \sim Q$ i.i.d., and $\epsilon > 0$.
- For $t = 1, 2, \ldots$, set

$$X_i^{(t)} \; = \; X_i^{(t-1)} + \frac{\epsilon}{n} \sum_{j=1}^{n} k(X_i^{(t)}, X_j^{(t)}) \nabla \log p(X_j^{(t)}) + \nabla_X k(X_i^{(t)}, X_j^{(t)}) \; .$$

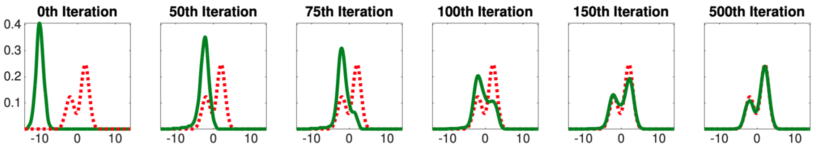


Figure credit: [Liu and Wang, 2016]

- Deterministic interacting particle system.
- Both asymptotic [Liu, 2017] and non-asymptotic theories [Liu and Wang, 2018] are available.

Variance Collapse

Even in moderate dimensions, SVGD particles will collapse onto the modes of P and exhibit no diversity.

- https://github.com/ImperialCollegeLondon/GSVGD/blob/main/imgs/gsvgd_cover.gif

Solutions:

- Work on low-dim projected spaces: [Gong et al., 2021a, Gong et al., 2021b, Liu et al., 2022b].

Yet There are Many More...

- Post-processing of MCMC samples [Riabiz et al., 2020].
- Stein points [Chen et al., 2018, Chen et al., 2019].
- Model training [Barp et al., 2019, Grathwohl et al., 2020].
- ...

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019).
Minimum Stein Discrepancy Estimators.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Barp, A., Simon-Gabriel, C.-J., Girolami, M., and Mackey, L. (2022).
Targeted separation and convergence with kernel discrepancies.
arXiv preprint arXiv:2209.12835.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (2019).
Stein point markov chain monte carlo.
In *International Conference on Machine Learning*, pages 1011–1021. PMLR.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. (2018).
Stein points.
In *International Conference on Machine Learning*, pages 844–853. PMLR.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A Kernel Test of Goodness of Fit.
In Balcan, M. F. and Weinberger, K. Q., 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. PMLR.

Gong, W., Li, Y., and Hernández-Lobato, J. M. (2021a).
Sliced Kernelized Stein Discrepancy.
In *International Conference on Learning Representations*.

Gong, W., Zhang, K., Li, Y., and Hernandez-Lobato, J. M. (2021b).
Active Slices for Sliced Stein Discrepancy.
In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3766–3776. PMLR.

Gorham, J. and Mackey, L. (2017).
Measuring sample quality with kernels.
In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

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

Hodgkinson, L., Salomone, R., and Roosta, F. (2020).
The reproducing Stein kernel approach for post-hoc corrected sampling.
arXiv preprint arXiv:2001.09266.

Liu, Q. (2017).
Stein variational gradient descent as gradient flow.
Advances in neural information processing systems, 30.

Liu, Q., Lee, J., and Jordan, M. (2016).
A Kernelized Stein Discrepancy for Goodness-of-fit Tests.
In Balcan, M. F. and Weinberger, K. Q., 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. PMLR.

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

Liu, Q. and Wang, D. (2018).
Stein variational gradient descent as moment matching.
Advances in Neural Information Processing Systems, 31.

Liu, X., Duncan, A., and Gandy, A. (2022a).
Using Perturbation to Improve Goodness-of-Fit Tests based on Kernelized Stein Discrepancy.
In *NeurIPS 2022 Workshop on Score-Based Methods*.

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

Müller, A. (1997).
Integral Probability Metrics and Their Generating Classes of Functions.
Advances in Applied Probability, 29(2):429–443.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., Oates, C., et al. (2020).
Optimal thinning of MCMC output.
arXiv preprint arXiv:2005.03952.

Wenliang, L. K. and Kanagawa, H. (2020).
Blindness of score-based methods to isolated components and mixing proportions.
arXiv preprint arXiv:2008.10087.