Sampling with Stein Discrepancies

Chris. J. Oates

Potsdam, September 2023

Stein Discrepancy (informal)

A *Stein discrepancy* is a statistical divergence

$$\mathrm{D}_P(\pi) \geq 0 \qquad \text{with equality if and only if} \qquad \pi = P$$

which can be computed without the normalisation constant of P.

Stein discrepancies are useful addition to the statistical and computational toolkit:

Posterior Approximation

$$\boxed{\arg\min_{\pi} \mathrm{D}_P(\pi)}$$

- thinning Markov chain Monte Carlo (MCMC) output [Riabiz et al., 2022]
- importance sampling [Liu and Lee, 2017, Hodgkinson et al., 2020]
- variational inference [Ranganath et al., 2016, Fisher et al., 2021]
- ...

Intractable Likelihood

$$\boxed{\arg\min_{\theta} \mathrm{D}_{P_\theta}(P_n)}$$

- goodness-of-fit testing [Liu et al., 2016, Chwialkowski et al., 2016]
- parameter estimation [Barp et al., 2019, Matsubara et al., 2022]
- ...

Stein Discrepancy (informal)

A *Stein discrepancy* is a statistical divergence

$$\mathrm{D}_P(\pi) \geq 0 \qquad \text{with equality if and only if} \qquad \pi = P$$

which can be computed without the normalisation constant of P.

Stein discrepancies are useful addition to the statistical and computational toolkit:

Posterior Approximation

$$\boxed{\arg\min_{\pi} \mathrm{D}_P(\pi)}$$

- thinning MCMC output [Riabiz et al., 2022]
- importance sampling [Liu and Lee, 2017, Hodgkinson et al., 2020]
- variational inference [Ranganath et al., 2016, Fisher et al., 2021]
- ...

Intractable Likelihood

$$\boxed{\arg\min_{\theta} \mathrm{D}_{P_\theta}(P_n)}$$

- goodness-of-fit testing [Liu et al., 2016, Chwialkowski et al., 2016]
- parameter estimation [Barp et al., 2019, Matsubara et al., 2022]
- ...

Stein Discrepancy (informal)

A *Stein discrepancy* is a statistical divergence

$$\mathrm{D}_P(\pi) \geq 0 \qquad \text{with equality if and only if} \qquad \pi = P$$

which can be computed without the normalisation constant of P.

Stein discrepancies are useful addition to the statistical and computational toolkit:

Posterior Approximation

$$\boxed{\arg\min_{\pi} \mathrm{D}_P(\pi)}$$

- thinning MCMC output [Riabiz et al., 2022]
- importance sampling [Liu and Lee, 2017, Hodgkinson et al., 2020]
- variational inference [Ranganath et al., 2016, Fisher et al., 2021]
- ...

Intractable Likelihood

$$\boxed{\arg\min_{\theta} \mathrm{D}_{P_\theta}(P_n)}$$

- goodness-of-fit testing [Liu et al., 2016, Chwialkowski et al., 2016]
- parameter estimation [Barp et al., 2019, Matsubara et al., 2022]
- ...

Case Study: Stein Importance Sampling

Stein Importance Sampling

1. Generate $(x_1, \ldots, x_n) \sim \mathbb{P}$.
2. Compute optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_P \left(\textstyle\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \ w^\top 1 = 1 \right\}.$$

3. Return the approximation $P_n^\star = \sum_{i=1}^n w_i^\star \delta(x_i)$.

Properties:

- Consistency $\mathrm{D}_P(P_n^\star) \xrightarrow{L^2(\mathbb{P})} 0$ [Hodgkinson et al., 2020] and strong consistency $\mathrm{D}_P(P_n^\star) \xrightarrow{\text{as}} 0$ [Riabiz et al., 2022] when $\mathbb{P}$ is Π-invariant MCMC with $\Pi \approx P$.
- Remarkable empirical performance on sufficiently nice P (see next slide).

Questions:

- How to select $\mathbb{P}$?
- I cannot access gradients of P, is this a problem?

Case Study: Stein Importance Sampling

Stein Importance Sampling

1. Generate $(x_1, \ldots, x_n) \sim \mathbb{P}$.
2. Compute optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_P \left(\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \ w^\top 1 = 1 \right\}.$$

3. Return the approximation $P_n^\star = \sum_{i=1}^n w_i^\star \delta(x_i)$.

Properties:

- Consistency $\mathrm{D}_P(P_n^\star) \overset{L^2(\mathbb{P})}{\to} 0$ [Hodgkinson et al., 2020] and strong consistency $\mathrm{D}_P(P_n^\star) \overset{\text{as}}{\to} 0$ [Riabiz et al., 2022] when $\mathbb{P}$ is Π-invariant MCMC with $\Pi \approx P$.
- Remarkable empirical performance on sufficiently nice P (see next slide).

Questions:

- How to select $\mathbb{P}$?
- I cannot access gradients of P, is this a problem?

Case Study: Stein Importance Sampling

Stein Importance Sampling

1. Generate $(x_1, \dots, x_n) \sim \mathbb{P}$.
2. Compute optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_P \left(\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \; w^\top 1 = 1 \right\}.$$

3. Return the approximation $P_n^\star = \sum_{i=1}^n w_i^\star \delta(x_i)$.

Properties:

- Consistency $\mathrm{D}_P(P_n^\star) \stackrel{L^2(\mathbb{P})}{\to} 0$ [Hodgkinson et al., 2020] and strong consistency $\mathrm{D}_P(P_n^\star) \stackrel{\text{as}}{\to} 0$ [Riabiz et al., 2022] when $\mathbb{P}$ is Π-invariant MCMC with $\Pi \approx P$.
- Remarkable empirical performance on sufficiently nice P (see next slide).

Questions:

- How to select $\mathbb{P}$?
- I cannot access gradients of P, is this a problem?

Case Study: Stein Importance Sampling

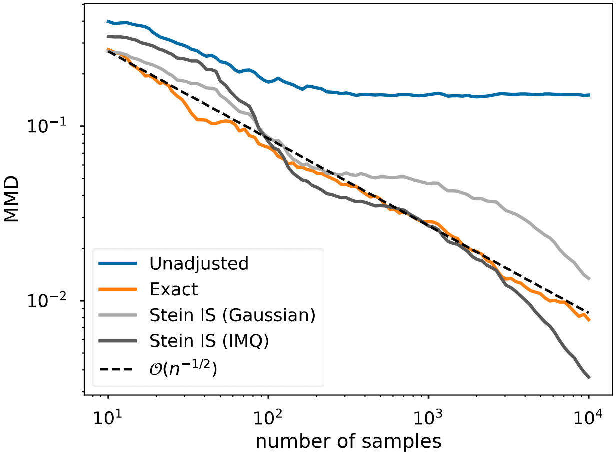
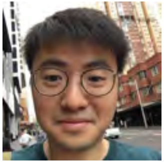


Figure: A 20-dimensional Gaussian target, with (biased) samples generated from the tamed unadjusted Langevin algorithm (TULA). Reproduced from Hodgkinson et al. [2020].

Stein Π-Importance Sampling


Congye Wang	Wilson Chen	Heishiro Kanagawa
Newcastle University	University of Sydney	Newcastle University

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$
$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$
$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$

$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$
$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\begin{aligned} \mu : \mathcal{P}_k(\mathbb{R}^d) &\to \mathcal{H}(k) \\ P &\mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x) \end{aligned}$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$
$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)} = \sup \left\{ \int h \, \mathrm{d}Q : \|h\|_{\mathcal{H}(k_P)} \le 1 \right\}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$.

Kernel Stein Discrepancies

For a symmetric positive definite function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, called a *kernel*, denote the associated reproducing kernel Hilbert space as $\mathcal{H}(k)$.

(e.g. the *inverse multi-quadric* kernel $k(x, y) = (1 + \|x - y\|^2)^{-1/2}$)

(e.g. $\sum_{i=1}^n w_i k(\cdot, x_i) \in \mathcal{H}(k)$)

Let $\mathcal{P}_k(\mathbb{R}^d)$ be the set of $P \in \mathcal{P}(\mathbb{R}^d)$ for which $\mathcal{H}(k) \subset L^1(P)$.

The *kernel mean embedding* is the map

$$\mu : \mathcal{P}_k(\mathbb{R}^d) \to \mathcal{H}(k)$$
$$P \mapsto \mu_P(\cdot) := \int k(\cdot, x) \, \mathrm{d}P(x)$$

A kernel is called a *Stein (reproducing) kernel* for P if $\mu_P = 0$, and write $k \equiv k_P$ to emphasise that.

Definition (Kernel Stein Discrepancy)

Let k_P be a Stein kernel for $P \in \mathcal{P}(\mathbb{R}^d)$. The associated kernel Stein discrepancy (KSD) is

$$\mathrm{D}_P(Q) = \|\mu_P(Q)\|_{\mathcal{H}(k_P)} = \sup\left\{\int h \, \mathrm{d}Q : \|h\|_{\mathcal{H}(k_P)} \leq 1\right\} = \left(\iint k_P(x, y) \, \mathrm{d}Q(x)\mathrm{d}Q(y)\right)^{1/2}$$

for $Q \in \mathcal{P}_{k_P}(\mathbb{R}^d)$. Computationally convenient.

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^{n} w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^{n} w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^n w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^n w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^{n} w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^{n} w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^{n} w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^{n} w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^n w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^n w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^n w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^n w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^n w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^n w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Problem: The components of $w^\star$ are strongly inter-dependent.

Solution: Consider weights that are *near-optimal* and whose components are only weakly dependent.

Self-normalised importance sampling (SNIS) is the approximation

$$P_n = \sum_{i=1}^n w_i \delta(x_i), \qquad w_i \propto \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i), \qquad x_1, \dots, x_n \overset{\text{IID}}{\sim} \Pi$$

satisfies $w \geq 0$ and $1^\top w = 1$, so that $\mathrm{D}_P(P_n^\star) \leq \mathrm{D}_P(P_n)$.

The asymptotic behaviour of SNIS can be characterised:

$$\mathrm{D}_P(P_n) = \left\| \frac{\xi_n}{\sqrt{n}} \right\|_{\mathcal{H}(k)}, \qquad \xi_n := \sqrt{n} \sum_{i=1}^n w_i k_P(\cdot, x_i) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i) k_P(\cdot, x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{C}_\Pi)$$

where $\mathcal{C}_\Pi : \mathcal{H}(k_P) \to \mathcal{H}(k_P)$ is the covariance operator defined via

$$\langle f, \mathcal{C}_\Pi g \rangle_{\mathcal{H}(k_P)} = \int \left\langle f, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \left\langle g, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x) k_P(\cdot, x) \right\rangle_{\mathcal{H}(k_P)} \mathrm{d}\Pi(x).$$

A Novel Approach to Selecting Π

Idea: Select Π such that $\mathrm{tr}(\mathcal{C}_\Pi)$ is minimised.

The variational problem

$$\arg\min_{\Pi} \mathrm{tr}(\mathcal{C}_\Pi), \qquad \mathrm{tr}(\mathcal{C}_\Pi) = \int \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)^2 k_P(x,x)\, \mathrm{d}\Pi(x)$$

has solution $(\mathrm{d}\Pi/\mathrm{d}P)(x) \propto \sqrt{k_P(x,x)}$.

Π can also be sampled using MCMC

Thus Π is adapted to the Stein kernel / KSD:

Figure: Illustrating our choice of Π in 2D.

A Novel Approach to Selecting Π

Idea: Select Π such that $\mathrm{tr}(\mathcal{C}_\Pi)$ is minimised.

The variational problem

$$\arg\min_{\Pi} \mathrm{tr}(\mathcal{C}_\Pi), \qquad \mathrm{tr}(\mathcal{C}_\Pi) = \int \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)^2 k_P(x,x) \, \mathrm{d}\Pi(x)$$

has solution $(\mathrm{d}\Pi/\mathrm{d}P)(x) \propto \sqrt{k_P(x,x)}$.

Π can also be sampled using MCMC

Thus Π is adapted to the Stein kernel / KSD:

Figure: Illustrating our choice of Π in 2D.

A Novel Approach to Selecting Π

Idea: Select Π such that $\mathrm{tr}(\mathcal{C}_\Pi)$ is minimised.

The variational problem

$$\arg\min_{\Pi} \mathrm{tr}(\mathcal{C}_\Pi), \qquad \mathrm{tr}(\mathcal{C}_\Pi) = \int \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)^2 k_P(x,x) \, \mathrm{d}\Pi(x)$$

has solution $(\mathrm{d}\Pi/\mathrm{d}P)(x) \propto \sqrt{k_P(x,x)}$.

Π can also be sampled using MCMC

Thus Π is adapted to the Stein kernel / KSD:

Figure: Illustrating our choice of Π in 2D.

A Novel Approach to Selecting Π

Idea: Select Π such that $\mathrm{tr}(\mathcal{C}_\Pi)$ is minimised.

The variational problem

$$\arg\min_{\Pi} \mathrm{tr}(\mathcal{C}_\Pi), \qquad \mathrm{tr}(\mathcal{C}_\Pi) = \int \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)^2 k_P(x,x) \, \mathrm{d}\Pi(x)$$

has solution $(\mathrm{d}\Pi/\mathrm{d}P)(x) \propto \sqrt{k_P(x,x)}$.

Π can also be sampled using MCMC

Thus Π is adapted to the Stein kernel / KSD:

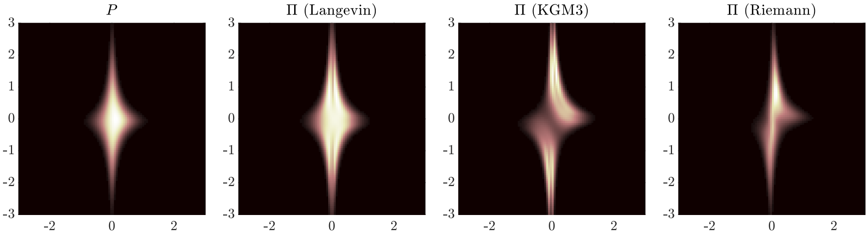


Figure: Illustrating our choice of Π in 2D.

A Novel Approach to Selecting Π

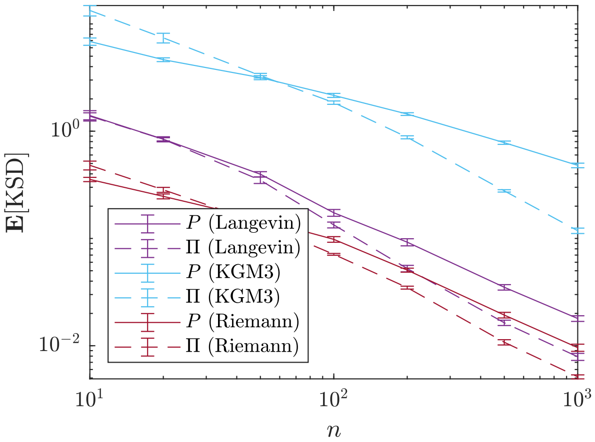


Figure: The mean kernel Stein discrepancy (KSD) for computation performed using the Langevin–Stein kernel (purple), the KGM3–Stein kernel (blue), and the Riemann–Stein kernel (red); in each case, KSD was computed using the same Stein kernel used to construct Π.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS\text{-}MALA}$)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)}\, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_x^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS\text{-}MALA}$)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)} \, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_X^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS}$-MALA)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)} \, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_x^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS}$-MALA)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)} \, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_x^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \stackrel{\text{a.s.}}{\to} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS}$-MALA)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)}\, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_x^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS}$-MALA)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)} \, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_x^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS\text{-}MALA}$)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)}\, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_X^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \overset{\text{a.s.}}{\to} 0$$

as $n \to \infty$.

Theoretical Guarantees

Question: Is Stein Π-Importance Sampling consistent?

Idea: Leverage the analysis of $\mathrm{S\Pi IS}$ in Riabiz et al. [2022] and the explicit conditions for ergodicity of MALA in Durmus and Moulines [2022].

Theorem (Strong consistency of $\mathrm{S\Pi IS}$-MALA)

Assume that

1. $\nabla \log p \in C^2(\mathbb{R}^d)$ *with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\| < \infty$ *(bounded second derivative)*
2. $-\nabla^2 \log p(x) \succeq b_1 I$ *for all* $\|x\| \geq B_1$ *(sub-Gaussian tail)*
3. $\inf_x k_P(x,x) > 0$, $\int \sqrt{k_P(x,x)} \, \mathrm{d}P(x) < \infty$, $k_P \in C^2(\mathbb{R}^d)$ *(embeddability)*
4. $\nabla_X^2 k_P(x,x) \preceq b_2 I$ *for all* $\|x\| \geq B_2$ *(sub-quadratic growth of Stein kernel)*

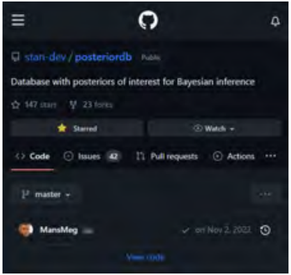
Then there exists $\epsilon_0 > 0$ *such that, for all step sizes* $\epsilon \in (0, \epsilon_0)$ *and all initial states* $x_0 \in \mathbb{R}^d$

$$\mathrm{D}_P(P_n^\star) \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$.

Performance Assessment with PosteriorDB and BridgeStan

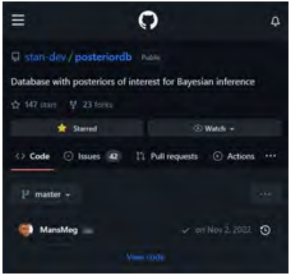
Task	d	Langevin Kernel Stein Discrepancy MALA	SIS-MALA	SΠIS-MALA	KGM3 Kernel Stein Discrepancy MALA	SIS-MALA	SΠIS-MALA
earnings-earn_height	3	1.41	0.0674	**0.0332**	5.33	0.656	**0.181**
gp_pois_regr-gp_regr	3	0.298	0.0436	**0.0373**	1.22	0.385	**0.223**
kidiq-kidscore_momhs	3	1.04	0.109	**0.0941**	4.66	0.848	**0.476**
kidiq-kidscore_momiq	3	5.03	0.516	**0.358**	25.3	4.86	**1.55**
mesquite-logmesquite_logvolume	3	1.10	0.179	**0.156**	4.97	1.70	**0.844**
arma-arma11	4	4.47	1.09	**1.01**	26.0	8.91	**6.03**
earnings-logearn_logheight_male	4	9.46	1.96	**1.59**	53.9	15.4	**8.65**
garch-garch11	4	0.543	0.159	**0.130**	4.70	1.16	**1.01**
kidiq-kidscore_momhsiq	4	5.21	0.982	**0.897**	29.3	7.25	**5.05**
earnings-logearn_interaction_z	5	3.09	1.36	**1.33**	19.3	10.4	**8.94**
kidiq-kidscore_interaction	5	7.74	**1.65**	1.79	47.8	13.2	**10.1**
kidiq_with_mom_work-kidscore_interaction_c	5	1.35	**0.659**	0.711	7.92	**4.05**	4.17
kidiq_with_mom_work-kidscore_interaction_c2	5	1.38	**0.689**	0.699	8.09	**4.24**	4.25
kidiq_with_mom_work-kidscore_interaction_z	5	1.11	0.500	**0.499**	6.62	**2.63**	3.25
kidiq_with_mom_work-kidscore_mom_work	5	1.07	**0.507**	0.545	6.70	**2.63**	3.04
low_dim_gauss_mix-low_dim_gauss_mix	5	5.51	1.87	**1.76**	37.5	14.7	**11.3**
mesquite-logmesquite_logva	5	1.83	0.821	**0.818**	12.6	5.73	**5.59**
hmm_example-hmm_example	6	1.99	0.578	**0.523**	11.6	4.13	**3.40**
sblrc-blr	6	479	154	**134**	3300	1100	**854**
sblri-blr	6	201	66.7	**60.3**	1340	**514**	595
arK-arK	7	6.87	3.39	**3.16**	60.4	26.4	**23.0**
mesquite-logmesquite_logvash	7	1.89	**1.18**	1.23	15.5	**8.88**	10.1
bball_drive_event_0-hmm_drive_0	8	1.15	**0.679**	0.698	8.55	4.72	**3.99**
bball_drive_event_1-hmm_drive_1	8	42.9	**11.9**	12.4	285	85.6	**67.8**
hudson_lynx_hare-lotka_volterra	8	4.62	2.29	**2.15**	47.4	**18.8**	18.9
mesquite-logmesquite	8	1.46	**1.00**	1.06	13.3	**8.28**	9.14
mesquite-logmesquite_logvas	8	2.02	**1.31**	1.35	19.2	**10.8**	12.2
mesquite-mesquite	8	0.429	0.268	**0.235**	3.71	**2.17**	2.42
eight_schools-eight_schools_centered	10	0.526	**0.100**	0.182	7.53	**2.15**	215
eight_schools-eight_schools_noncentered	10	0.210	0.137	**0.137**	43.6	28.7	**27.5**
nes1972-nes	10	6.16	3.89	**3.45**	72.9	36.2	**34.4**
nes1976-nes	10	6.67	3.86	**3.53**	77.5	35.5	**34.4**
nes1980-nes	10	4.34	2.68	**2.57**	49.8	**25.4**	25.7
nes1984-nes	10	6.18	3.75	**3.43**	71.3	34.9	**33.6**
nes1988-nes	10	7.40	3.70	**3.27**	81.4	34.6	**32.4**
nes1992-nes	10	7.52	4.32	**3.84**	89.1	39.7	**37.3**
nes1996-nes	10	6.44	3.87	**3.53**	74.1	36.4	**34.3**
nes2000-nes	10	3.35	2.22	**2.20**	38.6	**21.3**	22.8
diamonds-diamonds	26	196	157	**143**	5120	2990	**2620**
mcycle_gp-accel_gp	66	11.3	**8.25**	9.79	960	**623**	815



Improvement on ≈ 70% of tasks in PosteriorDB

Performance Assessment with PosteriorDB and BridgeStan

Task	d	Langevin Kernel Stein Discrepancy MALA	SIS-MALA	SΠIS-MALA	KGM3 Kernel Stein Discrepancy MALA	SIS-MALA	SΠIS-MALA
earnings-earn_height	3	1.41	0.0674	**0.0332**	5.33	0.656	**0.181**
gp_pois_regr-gp_regr	3	0.298	0.0436	**0.0373**	1.22	0.385	**0.223**
kidiq-kidscore_momhs	3	1.04	0.109	**0.0941**	4.66	0.848	**0.476**
kidiq-kidscore_momiq	3	5.03	0.516	**0.358**	25.3	4.86	**1.55**
mesquite-logmesquite_logvolume	3	1.10	0.179	**0.156**	4.97	1.70	**0.844**
arma-arma11	4	4.47	1.09	**1.01**	26.0	8.91	**6.03**
earnings-logearn_logheight_male	4	9.46	1.96	**1.59**	53.9	15.4	**8.65**
garch-garch11	4	0.543	0.159	**0.130**	4.70	1.16	**1.01**
kidiq-kidscore_momhsiq	4	5.21	0.982	**0.897**	29.3	7.25	**5.05**
earnings-logearn_interaction_z	5	3.09	1.36	**1.33**	19.3	10.4	**8.94**
kidiq-kidscore_interaction	5	7.74	**1.65**	1.79	47.8	13.2	**10.1**
kidiq_with_mom_work-kidscore_interaction_c	5	1.35	**0.659**	0.711	7.92	**4.05**	4.17
kidiq_with_mom_work-kidscore_interaction_c2	5	1.38	**0.689**	0.699	8.09	**4.24**	4.25
kidiq_with_mom_work-kidscore_interaction_z	5	1.11	0.500	**0.499**	6.62	**2.63**	3.25
kidiq_with_mom_work-kidscore_mom_work	5	1.07	**0.507**	0.545	6.70	**2.63**	3.04
low_dim_gauss_mix-low_dim_gauss_mix	5	5.51	1.87	**1.76**	37.5	14.7	**11.3**
mesquite-logmesquite_logva	5	1.83	0.821	**0.818**	12.6	5.73	**5.59**
hmm_example-hmm_example	6	1.99	0.578	**0.523**	11.6	4.13	**3.40**
sblrc-blr	6	479	154	**134**	3300	1100	**854**
sblri-blr	6	201	66.7	**60.3**	1340	**514**	595
arK-arK	7	6.87	3.39	**3.16**	60.4	26.4	**23.0**
mesquite-logmesquite_logvash	7	1.89	**1.18**	1.23	15.5	**8.88**	10.1
bball_drive_event_0-hmm_drive_0	8	1.15	**0.679**	0.698	8.55	4.72	**3.99**
bball_drive_event_1-hmm_drive_1	8	42.9	**11.9**	12.4	285	85.6	**67.8**
hudson_lynx_hare-lotka_volterra	8	4.62	2.29	**2.15**	47.4	**18.8**	18.9
mesquite-logmesquite	8	1.46	**1.00**	1.06	13.3	**8.28**	9.14
mesquite-logmesquite_logvas	8	2.02	**1.31**	1.35	19.2	**10.8**	12.2
mesquite-mesquite	8	0.429	0.268	**0.235**	3.71	**2.17**	2.42
eight_schools-eight_schools_centered	10	0.526	**0.100**	0.182	7.53	**2.15**	215
eight_schools-eight_schools_noncentered	10	0.210	0.137	**0.137**	43.6	28.7	**27.5**
nes1972-nes	10	6.16	3.89	**3.45**	72.9	36.2	**34.4**
nes1976-nes	10	6.67	3.86	**3.53**	77.5	35.5	**34.4**
nes1980-nes	10	4.34	2.68	**2.57**	49.8	**25.4**	25.7
nes1984-nes	10	6.18	3.75	**3.43**	71.3	34.9	**33.6**
nes1988-nes	10	7.40	3.70	**3.27**	81.4	34.6	**32.4**
nes1992-nes	10	7.52	4.32	**3.84**	89.1	39.7	**37.3**
nes1996-nes	10	6.44	3.87	**3.53**	74.1	36.4	**34.3**
nes2000-nes	10	3.35	2.22	**2.20**	38.6	**21.3**	22.8
diamonds-diamonds	26	196	157	**143**	5120	2990	**2620**
mcycle_gp-accel_gp	66	11.3	**8.25**	9.79	960	**623**	815



Improvement on ≈ 70% of tasks in PosteriorDB

Gradient-Free Kernel Stein Discrepancy

Matthew Fisher
Newcastle University

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P \mathcal{H}(k), \qquad S_P h := \frac{1}{p} \nabla \cdot (p \nabla h).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{d} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P \mathcal{H}(k), \qquad S_P h := \frac{1}{p} \nabla \cdot (p \nabla h).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{\mathrm{d}} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad S_P\mathbf{h} := \frac{1}{p}\nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{\text{d}} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad S_P\mathbf{h} := \frac{1}{p}\nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{d} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad S_P \mathbf{h} := \frac{1}{p} \nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{d} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad \mathcal{S}_P \mathbf{h} := \frac{1}{p} \nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{\text{d}} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad \mathcal{S}_P \mathbf{h} := \frac{1}{p} \nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{\text{d}} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Question: How to construct a Stein kernel?

The *Langevin*–Stein kernel k_P is defined as

$$\mathcal{H}(k_P) = \mathcal{S}_P[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)], \qquad S_P\mathbf{h} := \frac{1}{p}\nabla \cdot (p\mathbf{h}).$$

It is a popular choice since it

- does not require the normalisation constant of P
- has *weak convergence control*: $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{d} P$ [Gorham and Mackey, 2017]

However, all existing Stein kernels require that the gradient $\nabla \log p$

- exists, and
- can be efficiently computed.

Question: Can we construct a Stein kernel without taking a gradient?

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\qquad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad$ $(\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q \mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Our starting point is a gradient-free Stein operator, introduced in Han and Liu [2018] in the context of Stein variational gradient descent [Liu and Wang, 2016]:

Definition (Gradient-Free Stein Operator)

For $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$ and $\nabla \log q$ well-defined, the *gradient-free Stein operator* is defined as

$$\mathcal{S}_{P,Q}\mathbf{h} := \frac{q}{p}\left(\nabla \cdot \mathbf{h} + \mathbf{h} \cdot \nabla \log q\right),$$

acting on differentiable functions $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$.

Remarks:

- $\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = 0$ for suitably 'nice' $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ $\quad (\int \mathcal{S}_{P,Q}\mathbf{h} \, \mathrm{d}P = \int \mathcal{S}_Q\mathbf{h} \, \mathrm{d}Q)$
- if $Q \neq P$, the dependence on the derivatives of p is removed
- Q is an additional degree of freedom - this can be good and bad
- the canonical (or *Langevin*) Stein operator is recovered when $P = Q$

Now to create a discrepancy ...

Gradient-Free Kernel Stein Discrepancy

Definition (Gradient-Free Kernel Stein Discrepancy)

For $\pi \in \mathcal{P}(\mathbb{R}^d)$, the *gradient-free kernel Stein discrepancy* is defined as

$$\mathrm{D}_{P,Q}(\pi) = \left(\iint k_{P,Q}(x, y) \, \mathrm{d}\pi(x)\mathrm{d}\pi(y) \right)^{1/2}$$

where the *gradient-free Stein kernel* $k_{P,Q}$ is defined as $\mathcal{H}(k_{P,Q}) = \mathcal{S}_{P,Q}[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)]$.

This is well-defined if there is an $\alpha > 1$ such that

- $\int (q/p)^{\alpha} \, \mathrm{d}\pi < \infty$ and
- $\int \|\nabla \log q\|^{\alpha/(\alpha-1)} \, \mathrm{d}\pi < \infty$,

which are quite trivial when π is finitely supported. Call these "weak regularity conditions" (WRC).

GF-KSD is computable up to proportionality when p has an intractable normalising constant (like KSD).

But is this a useful discrepancy?

Gradient-Free Kernel Stein Discrepancy

Definition (Gradient-Free Kernel Stein Discrepancy)

For $\pi \in \mathcal{P}(\mathbb{R}^d)$, the *gradient-free kernel Stein discrepancy* is defined as

$$\mathrm{D}_{P,Q}(\pi) = \left(\iint k_{P,Q}(x,y) \, \mathrm{d}\pi(x) \mathrm{d}\pi(y) \right)^{1/2}$$

where the *gradient-free Stein kernel* $k_{P,Q}$ is defined as $\mathcal{H}(k_{P,Q}) = \mathcal{S}_{P,Q}[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)]$.

This is well-defined if there is an $\alpha > 1$ such that

- $\int (q/p)^\alpha \, \mathrm{d}\pi < \infty$ and
- $\int \|\nabla \log q\|^{\alpha/(\alpha-1)} \, \mathrm{d}\pi < \infty$,

which are quite trivial when π is finitely supported. Call these "weak regularity conditions" (WRC).

GF-KSD is computable up to proportionality when p has an intractable normalising constant (like KSD).

But is this a useful discrepancy?

Definition (Gradient-Free Kernel Stein Discrepancy)

For $\pi \in \mathcal{P}(\mathbb{R}^d)$, the *gradient-free kernel Stein discrepancy* is defined as

$$\mathrm{D}_{P,Q}(\pi) = \left(\iint k_{P,Q}(x,y) \, \mathrm{d}\pi(x) \mathrm{d}\pi(y) \right)^{1/2}$$

where the *gradient-free Stein kernel* $k_{P,Q}$ is defined as $\mathcal{H}(k_{P,Q}) = \mathcal{S}_{P,Q}[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)]$.

This is well-defined if there is an $\alpha > 1$ such that

- $\int (q/p)^\alpha \, \mathrm{d}\pi < \infty$ and
- $\int \|\nabla \log q\|^{\alpha/(\alpha-1)} \, \mathrm{d}\pi < \infty$,

which are quite trivial when π is finitely supported. Call these "weak regularity conditions" (WRC).

GF-KSD is computable up to proportionality when p has an intractable normalising constant (like KSD).

But is this a useful discrepancy?

Gradient-Free Kernel Stein Discrepancy

Definition (Gradient-Free Kernel Stein Discrepancy)

For $\pi \in \mathcal{P}(\mathbb{R}^d)$, the *gradient-free kernel Stein discrepancy* is defined as

$$\mathrm{D}_{P,Q}(\pi) = \left(\iint k_{P,Q}(x,y) \, \mathrm{d}\pi(x)\mathrm{d}\pi(y) \right)^{1/2}$$

where the *gradient-free Stein kernel* $k_{P,Q}$ is defined as $\mathcal{H}(k_{P,Q}) = \mathcal{S}_{P,Q}[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)]$.

This is well-defined if there is an $\alpha > 1$ such that

- $\int (q/p)^\alpha \, \mathrm{d}\pi < \infty$ and
- $\int \|\nabla \log q\|^{\alpha/(\alpha-1)} \, \mathrm{d}\pi < \infty$,

which are quite trivial when π is finitely supported. Call these "weak regularity conditions" (WRC).

GF-KSD is computable up to proportionality when p has an intractable normalising constant (like KSD).

But is this a useful discrepancy?

Gradient-Free Kernel Stein Discrepancy

Definition (Gradient-Free Kernel Stein Discrepancy)

For $\pi \in \mathcal{P}(\mathbb{R}^d)$, the *gradient-free kernel Stein discrepancy* is defined as

$$\mathrm{D}_{P,Q}(\pi) = \left(\iint k_{P,Q}(x,y) \, \mathrm{d}\pi(x)\mathrm{d}\pi(y) \right)^{1/2}$$

where the *gradient-free Stein kernel* $k_{P,Q}$ is defined as $\mathcal{H}(k_{P,Q}) = \mathcal{S}_{P,Q}[\mathcal{H}(k) \times \cdots \times \mathcal{H}(k)]$.

This is well-defined if there is an $\alpha > 1$ such that

- $\int (q/p)^\alpha \, \mathrm{d}\pi < \infty$ and
- $\int \|\nabla \log q\|^{\alpha/(\alpha-1)} \, \mathrm{d}\pi < \infty$,

which are quite trivial when π is finitely supported. Call these "weak regularity conditions" (WRC).

GF-KSD is computable up to proportionality when p has an intractable normalising constant (like KSD).

But is this a useful discrepancy?

Theoretical Justification for GF-KSD

For measurable $g : \mathbb{R}^d \to \mathbb{R}$, we follow Huggins and Mackey [2018] and denote the *tilted* Wasserstein distance as

$$\mathrm{W}_1(\pi, P; g) := \sup_{\mathrm{Lip}(f)\leq 1} \left| \int fg \, \mathrm{d}\pi - \int fg \, \mathrm{d}P \right|$$

whenever this expression is well-defined.

Theorem (GF-KSD Detects Convergence)

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$, $\nabla \log q$ Lipschitz and $\int \|\nabla \log q\|^2 \, \mathrm{d}Q < \infty$.
Assume the sequence $(\pi_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.
Then

$$\mathrm{W}_1(\pi_n, P; q/p) \to 0 \quad \Rightarrow \quad \mathrm{D}_{P,Q}(\pi_n) \to 0.$$

Theoretical Justification for GF-KSD

For measurable $g : \mathbb{R}^d \to \mathbb{R}$, we follow Huggins and Mackey [2018] and denote the *tilted* Wasserstein distance as

$$\mathrm{W}_1(\pi, P; g) := \sup_{\mathrm{Lip}(f) \leq 1} \left| \int fg \, \mathrm{d}\pi - \int fg \, \mathrm{d}P \right|$$

whenever this expression is well-defined.

Theorem (GF-KSD Detects Convergence)

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$, $\nabla \log q$ Lipschitz and $\int \|\nabla \log q\|^2 \, \mathrm{d}Q < \infty$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{W}_1(\pi_n, P; q/p) \to 0 \quad \Rightarrow \quad \mathrm{D}_{P,Q}(\pi_n) \to 0.$$

Theoretical Justification for GF-KSD

For measurable $g : \mathbb{R}^d \to \mathbb{R}$, we follow Huggins and Mackey [2018] and denote the *tilted* Wasserstein distance as

$$\mathrm{W}_1(\pi, P; g) := \sup_{\mathrm{Lip}(f) \leq 1} \left| \int fg \, \mathrm{d}\pi - \int fg \, \mathrm{d}P \right|$$

whenever this expression is well-defined.

Theorem (GF-KSD Detects Convergence)

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$, $\nabla \log q$ Lipschitz and $\int \|\nabla \log q\|^2 \, \mathrm{d}Q < \infty$.
Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.
Then

$$\mathrm{W}_1(\pi_n, P; q/p) \to 0 \quad \Rightarrow \quad \mathrm{D}_{P,Q}(\pi_n) \to 0.$$

Theoretical Justification for GF-KSD

For measurable $g : \mathbb{R}^d \to \mathbb{R}$, we follow Huggins and Mackey [2018] and denote the *tilted* Wasserstein distance as

$$\mathrm{W}_1(\pi, P; g) := \sup_{\mathrm{Lip}(f) \leq 1} \left| \int fg \, \mathrm{d}\pi - \int fg \, \mathrm{d}P \right|$$

whenever this expression is well-defined.

Theorem (GF-KSD Detects Convergence)

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $Q \ll P$, $\nabla \log q$ Lipschitz and $\int \|\nabla \log q\|^2 \, \mathrm{d}Q < \infty$.
Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.
Then

$$\mathrm{W}_1(\pi_n, P; q/p) \to 0 \quad \Rightarrow \quad \mathrm{D}_{P,Q}(\pi_n) \to 0.$$

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Theoretical Justification for GF-KSD

Main condition on q: Let $\mathcal{Q}(\mathbb{R}^d)$ denote the set of probability distributions with positive density function $q : \mathbb{R}^d \to (0, \infty)$ for which $\nabla \log q$ is Lipschitz and q is strongly log-concave outside (and on the boundary of) a compact set.

(implies Q-invariant overdamped Langevin mixes fast)

Theorem (GF-KSD Controls Convergence)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$ be such that p is continuous and $\inf_{x \in \mathbb{R}^d} q(x)/p(x) > 0$.

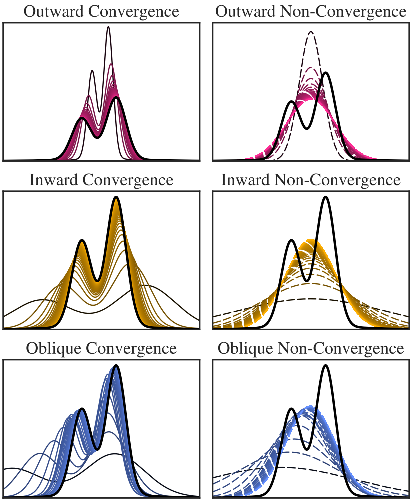
Assume the sequence $(\pi_n)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfies WRC.

Then

$$\mathrm{D}_{P,Q}(\pi_n) \to 0 \quad \Rightarrow \quad \pi_n \xrightarrow{d} P.$$

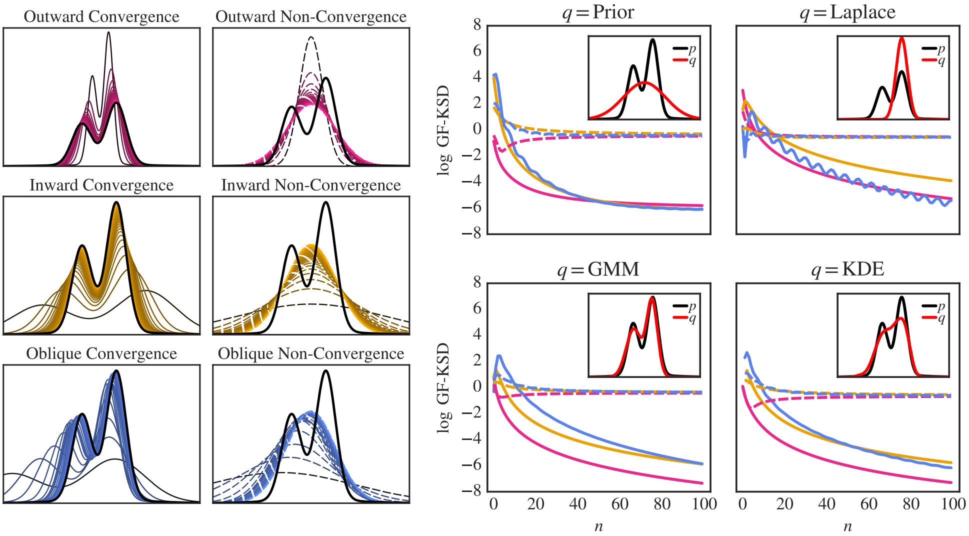
The proof is based on re-casting GF-KSD as standard KSD between Q and a transformed distribution $\bar{\pi}$, then appealing to the analysis of Gorham and Mackey [2017].

Selecting q in GF-KSD



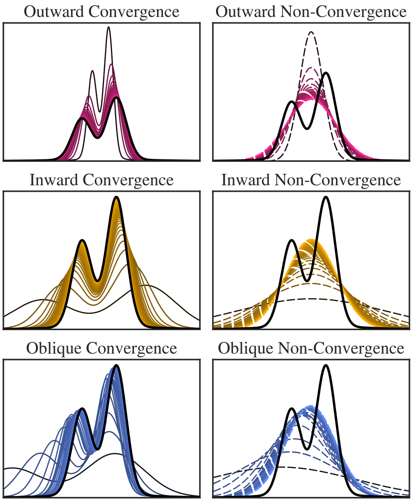
Selecting q will ultimately be task-specific; we start with Laplace and then go beyond...

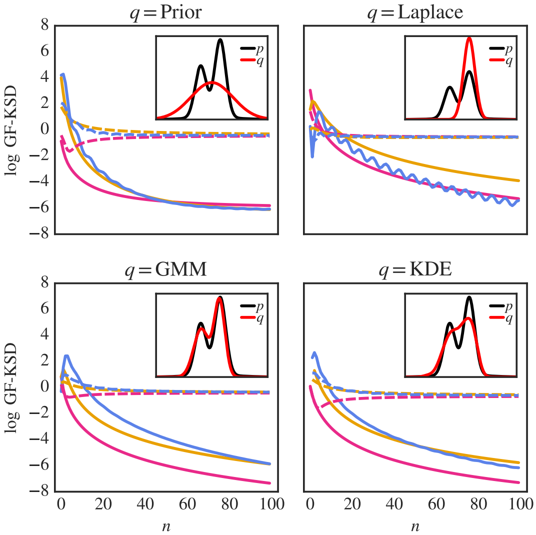
Selecting q in GF-KSD



Selecting q will ultimately be task-specific; we start with Laplace and then go beyond...

Selecting q in GF-KSD





Selecting q will ultimately be task-specific; we start with Laplace and then go beyond...

Application #1: Gradient-Free Stein Importance Sampling

To date, applications of Stein importance sampling have been limited to instances where the statistical model p can be differentiated; our contribution is to remove this requirement.

Theorem (Gradient-Free Stein Importance Sampling)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$, p continuous, $\inf q/p > 0$, and $\int \exp\{\gamma \|\nabla \log q\|^2\} \, \mathrm{d}Q < \infty$.
Let $(x_n)_{n \in \mathbb{N}}$ be independent samples from Q.
To the sample, assign optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_{P,Q} \left(\textstyle\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \; w^\top 1 = 1 \right\}.$$

Then

$$\sum_{i=1}^n w_i^\star \delta(x_i) \xrightarrow{\mathrm{d}} P \qquad a.s. \;\; as \; n \to \infty.$$

Application #1: Gradient-Free Stein Importance Sampling

To date, applications of Stein importance sampling have been limited to instances where the statistical model p can be differentiated; our contribution is to remove this requirement.

Theorem (Gradient-Free Stein Importance Sampling)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$, p *continuous*, $\inf q/p > 0$, *and* $\int \exp\{\gamma \|\nabla \log q\|^2\} \, \mathrm{d}Q < \infty$.

Let $(x_n)_{n \in \mathbb{N}}$ *be independent samples from* Q.

To the sample, assign optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_{P,Q}\left(\textstyle\sum_{i=1}^n w_i \delta(x_i)\right) : 0 \leq w, \; w^\top 1 = 1 \right\}.$$

Then

$$\sum_{i=1}^n w_i^\star \delta(x_i) \xrightarrow{\mathrm{d}} P \qquad a.s. \;\; as \; n \to \infty.$$

Application #1: Gradient-Free Stein Importance Sampling

To date, applications of Stein importance sampling have been limited to instances where the statistical model p can be differentiated; our contribution is to remove this requirement.

Theorem (Gradient-Free Stein Importance Sampling)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$, *$p$ continuous,* $\inf q/p > 0$, *and* $\int \exp\{\gamma \|\nabla \log q\|^2\} \, \mathrm{d}Q < \infty$.
Let $(x_n)_{n \in \mathbb{N}}$ *be independent samples from* Q.
To the sample, assign optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_{P,Q} \left(\textstyle\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \ w^\top 1 = 1 \right\}.$$

Then

$$\sum_{i=1}^n w_i^\star \delta(x_i) \xrightarrow{\mathrm{d}} P \qquad \text{a.s. as } n \to \infty.$$

Application #1: Gradient-Free Stein Importance Sampling

To date, applications of Stein importance sampling have been limited to instances where the statistical model p can be differentiated; our contribution is to remove this requirement.

Theorem (Gradient-Free Stein Importance Sampling)

Let $P \in \mathcal{P}(\mathbb{R}^d)$, $Q \in \mathcal{Q}(\mathbb{R}^d)$, p continuous, $\inf q/p > 0$, *and* $\int \exp\{\gamma \|\nabla \log q\|^2\} \, \mathrm{d}Q < \infty$.
Let $(x_n)_{n \in \mathbb{N}}$ be independent samples from Q.
To the sample, assign optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_{P,Q} \left(\textstyle\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \le w, \; w^\top 1 = 1 \right\}.$$

Then

$$\sum_{i=1}^n w_i^\star \delta(x_i) \xrightarrow{\mathrm{d}} P \qquad \text{a.s. as } n \to \infty.$$

Application #1: Gradient-Free Stein Importance Sampling

To date, applications of Stein importance sampling have been limited to instances where the statistical model p can be differentiated; our contribution is to remove this requirement.

Theorem (Gradient-Free Stein Importance Sampling)

Let $P \in \mathcal{P}(\mathbb{R}^d)$*,* $Q \in \mathcal{Q}(\mathbb{R}^d)$*,* p *continuous,* $\inf q/p > 0$*, and* $\int \exp\{\gamma \|\nabla \log q\|^2\} \, \mathrm{d}Q < \infty$*.*

Let $(x_n)_{n \in \mathbb{N}}$ *be independent samples from* Q*.*

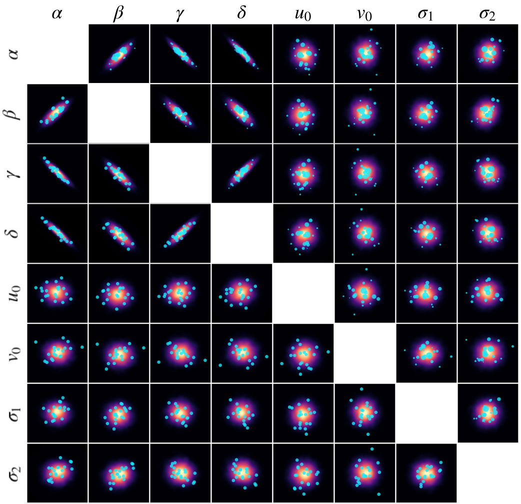
To the sample, assign optimal weights

$$w^\star \in \arg\min \left\{ \mathrm{D}_{P,Q} \left(\textstyle\sum_{i=1}^n w_i \delta(x_i) \right) : 0 \leq w, \; w^\top 1 = 1 \right\}.$$

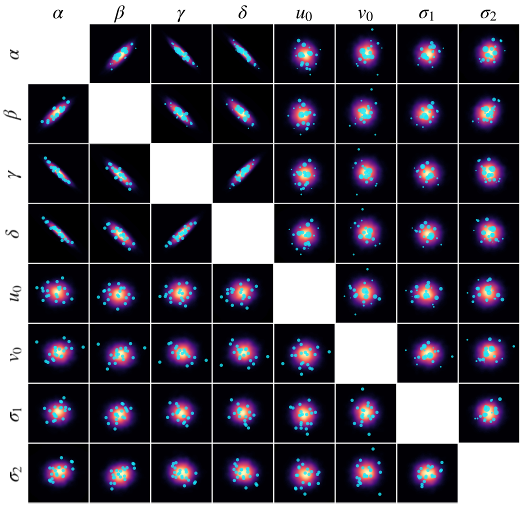
Then

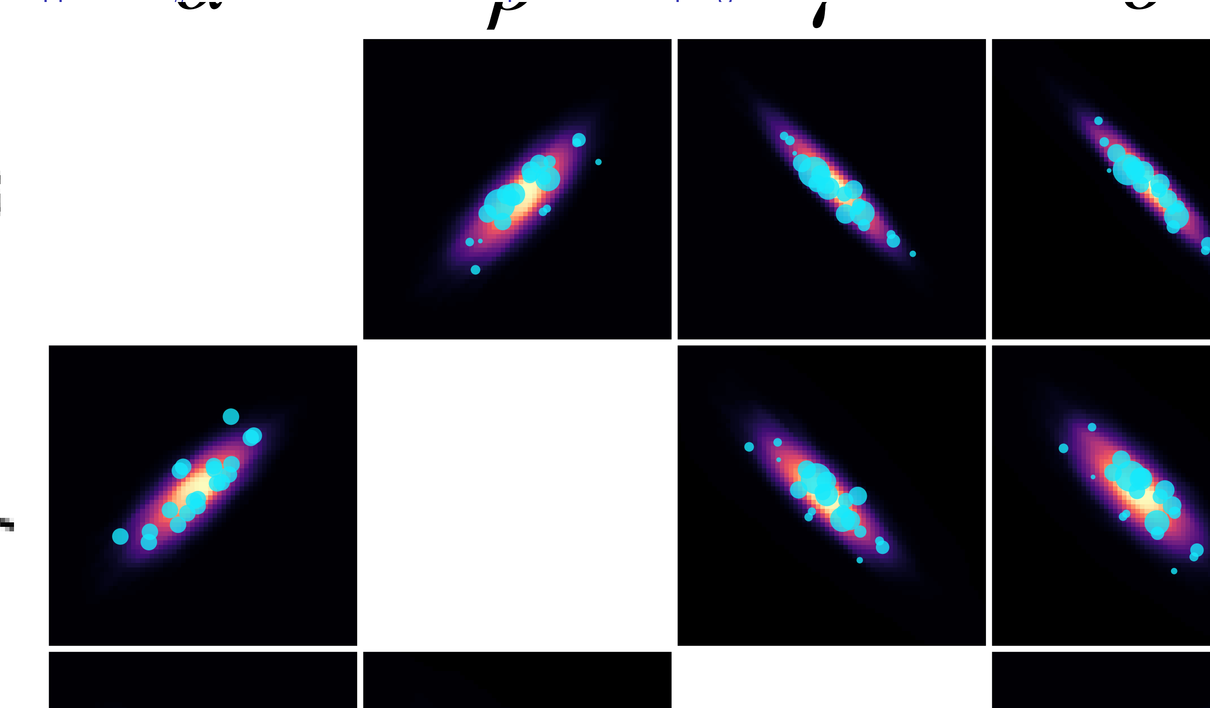
$$\sum_{i=1}^n w_i^\star \delta(x_i) \xrightarrow{\mathrm{d}} P \qquad a.s. \;\; as \; n \to \infty.$$

Application #1: Gradient-Free Stein Importance Sampling

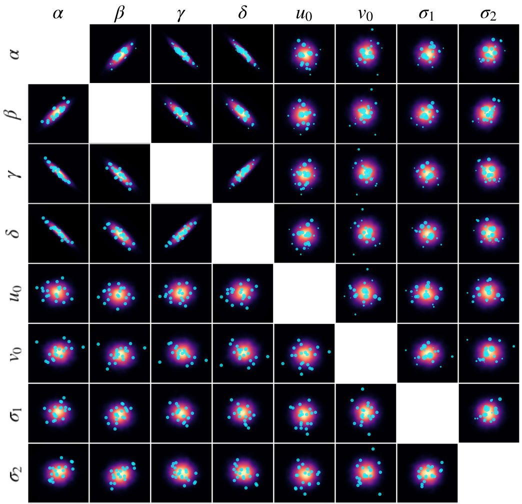


Application #1: Gradient-Free Stein Importance Sampling

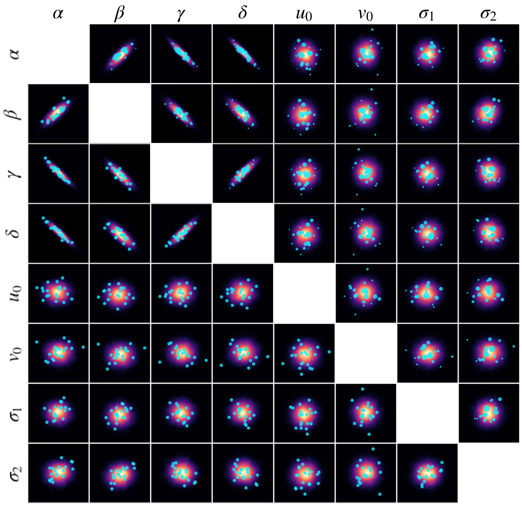


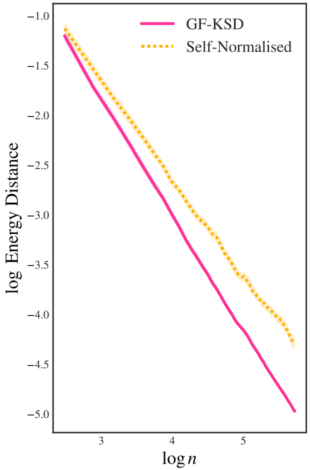


Application #1: Gradient-Free Stein Importance Sampling



Application #1: Gradient-Free Stein Importance Sampling





Application #2: Stein Variational Inference Without Second Order Gradient

KSD is an appealing alternative to KLD for VI because it does not require the variational family to be absolutely continuous with respect to P, unlike KLD.

However, KSD requires second-order gradients of p to be computed [Fisher et al., 2021]; our contribution is to remove this requirement.

An interesting methodological extension is to take $Q = P_{\theta_n}$ to be the 'current approximation' to p along the stochastic optimisation path.

Application #2: Stein Variational Inference Without Second Order Gradient

KSD is an appealing alternative to KLD for VI because it does not require the variational family to be absolutely continuous with respect to P, unlike KLD.

However, KSD requires second-order gradients of p to be computed [Fisher et al., 2021]; our contribution is to remove this requirement.

An interesting methodological extension is to take $Q = P_{\theta_n}$ to be the 'current approximation' to p along the stochastic optimisation path.

Application #2: Stein Variational Inference Without Second Order Gradient

KSD is an appealing alternative to KLD for VI because it does not require the variational family to be absolutely continuous with respect to P, unlike KLD.

However, KSD requires second-order gradients of p to be computed [Fisher et al., 2021]; our contribution is to remove this requirement.

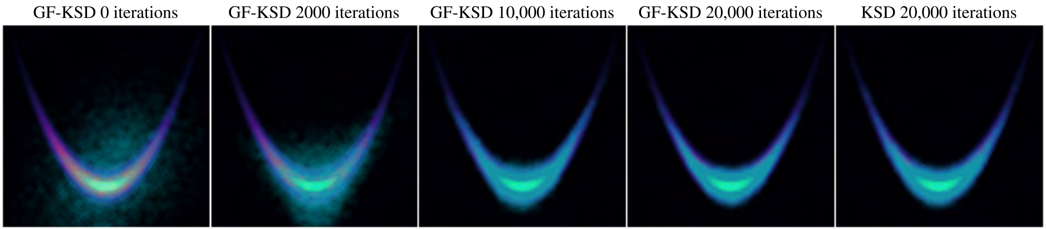
An interesting methodological extension is to take $Q = P_{\theta_n}$ to be the 'current approximation' to p along the stochastic optimisation path.

Application #2: Stein Variational Inference Without Second Order Gradient

KSD is an appealing alternative to KLD for VI because it does not require the variational family to be absolutely continuous with respect to P, unlike KLD.

However, KSD requires second-order gradients of p to be computed [Fisher et al., 2021]; our contribution is to remove this requirement.

An interesting methodological extension is to take $Q = P_{\theta_n}$ to be the 'current approximation' to p along the stochastic optimisation path.



Summary

Stein discrepancies have given rise to a new generation of computational methods!

This raises many interesting research questions:

- explore the interplay between the choice of Stein discrepancy and the sampling method
- identify when one of the failure modes of KSD / GF-KSD has occurred
- extend to spaces other than $\mathbb{R}^d$

Full details are contained in the preprints

Wang C, Chen WY, Kanagawa H, CJO. Stein Π-Importance Sampling, arXiv:2305.10068

Fisher MA and CJO. Gradient-Free Kernel Stein Discrepancy, arXiv:2207.02636

Thank you for your attention!

Summary

Stein discrepancies have given rise to a new generation of computational methods!

This raises many interesting research questions:

- explore the interplay between the choice of Stein discrepancy and the sampling method
- identify when one of the failure modes of KSD / GF-KSD has occurred
- extend to spaces other than $\mathbb{R}^d$

Full details are contained in the preprints

Wang C, Chen WY, Kanagawa H, CJO. Stein Π-Importance Sampling, arXiv:2305.10068

Fisher MA and CJO. Gradient-Free Kernel Stein Discrepancy, arXiv:2207.02636

Thank you for your attention!

Summary

Stein discrepancies have given rise to a new generation of computational methods!

This raises many interesting research questions:

- explore the interplay between the choice of Stein discrepancy and the sampling method
- identify when one of the failure modes of KSD / GF-KSD has occurred
- extend to spaces other than $\mathbb{R}^d$

Full details are contained in the preprints

Wang C, Chen WY, Kanagawa H, CJO. Stein Π-Importance Sampling, arXiv:2305.10068

Fisher MA and CJO. Gradient-Free Kernel Stein Discrepancy, arXiv:2207.02636

Thank you for your attention!

Summary

Stein discrepancies have given rise to a new generation of computational methods!

This raises many interesting research questions:

- explore the interplay between the choice of Stein discrepancy and the sampling method
- identify when one of the failure modes of KSD / GF-KSD has occurred
- extend to spaces other than $\mathbb{R}^d$

Full details are contained in the preprints

Wang C, Chen WY, Kanagawa H, CJO. Stein Π-Importance Sampling, arXiv:2305.10068

Fisher MA and CJO. Gradient-Free Kernel Stein Discrepancy, arXiv:2207.02636

Thank you for your attention!

References I

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, volume 32, pages 12964–12976, 2019.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2606–2615. PMLR, 2016.

A. Durmus and É. Moulines. On the geometric convergence for MALA under verifiable conditions. *arXiv preprint arXiv:2201.01951*, 2022.

M. Fisher, T. Nolan, M. Graham, D. Prangle, and C. J. Oates. Measure transport with kernel Stein discrepancy. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2021. (Here we refer to the error-corrected version arXiv:2010.11779.).

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

J. Han and Q. Liu. Stein variational gradient descent without gradient. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1900–1908. PMLR, 2018.

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

J. H. Huggins and L. Mackey. Random feature Stein discrepancies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1903–1913, 2018.

Q. Liu and J. Lee. Black-box importance sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, 2016.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 276–284. PMLR, 2016.

References II

T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society, Series B*, (84):997–1022, 2022.

R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator variational inference. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, pages 496–504, 2016.

M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society, Series B*, 2022. To appear.

Failure Modes of GF-KSD