Sampling and Stein's Method

Chris. J. Oates
Newcastle University
Alan Turing Institute

April 2022
Advances in Stein's method and its applications in Machine Learning and Optimization

Sampling and Stein's Method*

Chris. J. Oates
Newcastle University
Alan Turing Institute

April 2022
Advances in Stein's method and its applications in Machine Learning and Optimization

*except Stein variational gradient descent (SVGD)!

Recap: The Sampling Problem in Bayesian Statistics

Computation for the Bayesian Framework

The goal is to obtain an approximation to the posterior in a Bayesian context:

$$P : \pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}$$

where $\theta \in \Theta$ are the unknown parameters of the model, $\pi(\theta)$ is an appropriate prior density and y denotes the dataset.

This raises technical challenges as the normalisation constant

$$\pi(y) = \int_{\Theta} \pi(y|\theta)\pi(\theta)\mathrm{d}\theta$$

is an intractable d-dimensional integral.

Sampling from P via Markov chain Monte Carlo (MCMC) is a popular approach which requires only evaluation of the un-normalised form

$$p(\theta) := \pi(y|\theta)\pi(\theta),$$

but it is not a silver bullet.

Computation for the Bayesian Framework

The goal is to obtain an approximation to the posterior in a Bayesian context:

$$P : \pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}$$

where $\theta \in \Theta$ are the unknown parameters of the model, $\pi(\theta)$ is an appropriate prior density and y denotes the dataset.

This raises technical challenges as the normalisation constant

$$\pi(y) = \int_{\Theta} \pi(y|\theta)\pi(\theta)\mathrm{d}\theta$$

is an intractable d-dimensional integral.

Sampling from P via Markov chain Monte Carlo (MCMC) is a popular approach which requires only evaluation of the un-normalised form

$$p(\theta) := \pi(y|\theta)\pi(\theta),$$

but it is not a silver bullet.

Computation for the Bayesian Framework

The goal is to obtain an approximation to the posterior in a Bayesian context:

$$P \; : \; \pi(\theta|y) \quad = \quad \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}$$

where $\theta \in \Theta$ are the unknown parameters of the model, $\pi(\theta)$ is an appropriate prior density and y denotes the dataset.

This raises technical challenges as the normalisation constant

$$\pi(y) \quad = \quad \int_{\Theta} \pi(y|\theta)\pi(\theta)\mathrm{d}\theta$$

is an intractable d-dimensional integral.

Sampling from P via Markov chain Monte Carlo (MCMC) is a popular approach which requires only evaluation of the un-normalised form

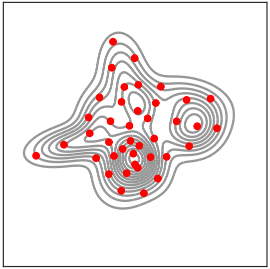
$$p(\theta) := \pi(y|\theta)\pi(\theta),$$

but it is not a silver bullet.

Optimal Quantisation

"Pick a collection of parameters that best represents P"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right)$$



[For now we focus on optimisation in Θ^m, but later we will discuss optimisation over $\mathcal{P}(\Theta)$.]

Remarks:

- "Nice idea, but we don't have access to P."
- "High-dimensional optimisation is hard."

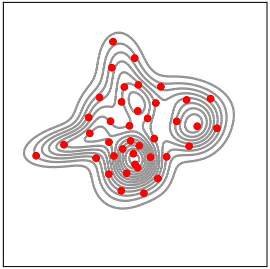
This tutorial will explain how **Stein's Method** can be used to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$, and to review methodology for optimisation of $(*)$.

Optimal Quantisation

"Pick a collection of parameters that best represents P"

Idea:

$$\underset{\theta_1,\dots,\theta_m \in \Theta}{\arg\min} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right)$$



[For now we focus on optimisation in Θ^m, but later we will discuss optimisation over $\mathcal{P}(\Theta)$.]

Remarks:

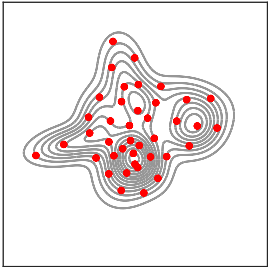
- "Nice idea, but we don't have access to P."
- "High-dimensional optimisation is hard."

This tutorial will explain how **Stein's Method** can be used to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$, and to review methodology for optimisation of $(*)$.

Optimal Quantisation

"Pick a collection of parameters that best represents P"

Idea: $$\operatorname*{arg\,min}_{\theta_1,\ldots,\theta_m \in \Theta} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right)$$



[For now we focus on optimisation in Θ^m, but later we will discuss optimisation over $\mathcal{P}(\Theta)$.]

Remarks:

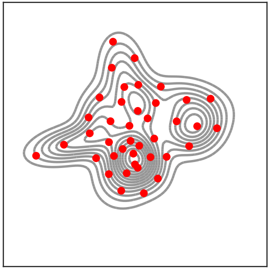
- "Nice idea, but we don't have access to P."
- "High-dimensional optimisation is hard."

This tutorial will explain how **Stein's Method** can be used to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$, and to review methodology for optimisation of $(*)$.

Optimal Quantisation

"Pick a collection of parameters that best represents P"

Idea: $$\underset{\theta_1,\dots,\theta_m \in \Theta}{\arg\min} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right)$$



[For now we focus on optimisation in Θ^m, but later we will discuss optimisation over $\mathcal{P}(\Theta)$.]

Remarks:

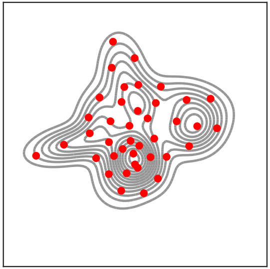
- "Nice idea, but we don't have access to P."
- "High-dimensional optimisation is hard."

This tutorial will explain how **Stein's Method** can be used to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$, and to review methodology for optimisation of $(*)$.

Optimal Quantisation

"Pick a collection of parameters that best represents P"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right)$$



[For now we focus on optimisation in Θ^m, but later we will discuss optimisation over $\mathcal{P}(\Theta)$.]

Remarks:

- "Nice idea, but we don't have access to P."
- "High-dimensional optimisation is hard."

This tutorial will explain how **Stein's Method** can be used to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$, and to review methodology for optimisation of $(*)$.

Sampling and Stein's Method

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot)\rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. **(Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\|\cdot\|_{\mathcal{H}(k)}$:

$$\begin{aligned} \mathrm{diff}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) &:= \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \frac{1}{m}\sum_{i=1}^m f(\theta_i) - \mathbb{E}_{\vartheta \sim P}[f(\vartheta)] \right| \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^m\right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^m)^2 = \left\| \frac{1}{m}\sum_{i=1}^m k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \mathrm{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(\theta_i) - \mathbb{E}_{\vartheta \sim P}[f(\vartheta)] \right| \\ &=: D_{\mathcal{H}(k), P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^m)^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. **(Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$**)**

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \mathrm{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} \langle f, k(\theta_i, \cdot) \rangle_{\mathcal{H}(k)} - \mathbb{E}_{\vartheta \sim P} [\langle f, k(\vartheta, \cdot) \rangle_{\mathcal{H}(k)}] \right| \\ &=: D_{\mathcal{H}(k), P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k), P}(\{\theta_i\}_{i=1}^{m})^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot)\rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\|\cdot\|_{\mathcal{H}(k)}$:

$$\begin{aligned} \text{diff}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) &:= \sup_{\|f\|_{\mathcal{H}(k)}\leq 1} \left| \left\langle f, \frac{1}{m}\sum_{i=1}^m k(\theta_i, \cdot) - \mathbb{E}_{\vartheta\sim P}[k(\vartheta, \cdot)] \right\rangle_{\mathcal{H}(k)} \right| \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^m\right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^m)^2 = \left\| \frac{1}{m}\sum_{i=1}^m k(\theta_i, \cdot) - \int k(\theta, \cdot)\mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. **(Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\mathrm{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) := \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} =: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right)$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \operatorname{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. **(Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$**)**

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \mathrm{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) \quad &:= \quad \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\ &=: \quad D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 \quad = \quad \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)}^2$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \operatorname{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta), \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\rangle_{\mathcal{H}(k)}$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \text{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$\begin{aligned} D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 &= \frac{1}{m^2} \sum_{i,j=1}^{m} \langle k(\theta_i, \cdot), k(\theta_j, \cdot) \rangle_{\mathcal{H}(k)} - \frac{2}{m} \sum_{i=1}^{m} \int \langle k(\theta, \cdot), k(\theta_i, \cdot) \rangle_{\mathcal{H}(k)} \mathrm{d}P(\theta) \\ &\quad - \int \int \langle k(\theta, \cdot), k(\theta', \cdot) \rangle_{\mathcal{H}(k)} \mathrm{d}P(\theta) \mathrm{d}P(\theta') \end{aligned}$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$
\begin{aligned}
\mathrm{diff}\left(\frac{1}{m}\sum_{i=1}^{m}\delta(\theta_i), P\right) &:= \left\| \frac{1}{m}\sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot)\mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\
&=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m}\right)
\end{aligned}
$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$
D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} k(\theta_i, \theta_j) - \frac{2}{m}\sum_{i=1}^{m}\int k(\theta, \theta_i)\mathrm{d}P(\theta) + \iint k(\theta, \vartheta)\mathrm{d}P(\theta)\mathrm{d}P(\vartheta)
$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Background: Quantisation via Reproducing Kernel Hilbert Spaces (RKHS)

Let $k : \Theta \times \Theta \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{H}(k)$ of functions from Θ to $\mathbb{R}$; i.e $\forall \theta \in \Theta$, $k(\theta, \cdot) \in \mathcal{H}(k)$ and $f(\theta) = \langle f, k(\theta, \cdot) \rangle_{\mathcal{H}(k)}$ whenever $f \in \mathcal{H}(k)$. (**Intuition:** $f(\theta) = \sum_i c_i k(\theta, \theta_i)$)

Consider an **integral probability pseudo-metric** based on $\| \cdot \|_{\mathcal{H}(k)}$:

$$\begin{aligned} \mathrm{diff}\left(\frac{1}{m} \sum_{i=1}^{m} \delta(\theta_i), P \right) &:= \left\| \frac{1}{m} \sum_{i=1}^{m} k(\theta_i, \cdot) - \int k(\theta, \cdot) \mathrm{d}P(\theta) \right\|_{\mathcal{H}(k)} \\ &=: D_{\mathcal{H}(k),P}\left(\{\theta_i\}_{i=1}^{m} \right) \end{aligned}$$

which is sometimes called the *maximum mean discrepancy*, or the *worst-case integration error* for the RKHS $\mathcal{H}(k)$.

Let's try to compute this:

$$D_{\mathcal{H}(k),P}(\{\theta_i\}_{i=1}^{m})^2 = \frac{1}{m^2} \sum_{i,j=1}^{m} k(\theta_i, \theta_j) - \frac{2}{m} \sum_{i=1}^{m} \int k(\theta, \theta_i) \mathrm{d}P(\theta) + \iint k(\theta, \vartheta) \mathrm{d}P(\theta) \mathrm{d}P(\vartheta)$$

Problem: We need to choose k carefully, so that the integrals can be evaluated. How?

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A}f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Sketch (easy direction, $d = 1$)

$$\mathbb{E}_{\vartheta \sim P}[\mathcal{A}f(\vartheta)] = \int \frac{(fp)'}{p} \mathrm{d}P = \int (fp)' \mathrm{d}x = f(\infty)p(\infty) - f(-\infty)p(-\infty) = 0$$

Stein Characterisation

Definition (Stein Characterisation)

A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Set $\mathcal{F}$, if it holds that

$$\vartheta \sim P \quad \text{iff} \quad \mathbb{E}[\mathcal{A} f(\vartheta)] = 0 \quad \forall f \in \mathcal{F}.$$

Proposition (Chwialkowski, Strathmann, and Gretton [2016])

Suppose that κ is a reproducing kernel on $\Theta = \mathbb{R}^d$ such that κ and its first-order mixed derivatives are bounded, that κ is C_0-universal, and that $\mathbb{E}_{\vartheta \sim P}[\|\nabla \log p(\vartheta)\|^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A} f = \frac{\nabla \cdot (fp)}{p}, \qquad \mathcal{F} = \left\{ f \in \mathcal{H}(\kappa)^d : \sum_{i=1}^{d} \|f_i\|^2_{\mathcal{H}(\kappa)} \leq 1 \right\}.$$

Proposition (CJO, Girolami, and Chopin [2017])

The above functions $\mathcal{A}f$ constitute the unit ball in a Stein RKHS $\mathcal{H}(k_P) := \mathcal{A}\mathcal{H}(\kappa)$ with kernel

$$k_P(\theta, \theta') = \nabla_\theta \cdot \nabla_{\theta'} \kappa(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \cdot \nabla_{\theta'} \kappa(\theta, \theta') + \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} \cdot \nabla_\theta \kappa(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \cdot \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} \kappa(\theta, \theta').$$

In particular, $\int k_P(\theta, \cdot) \mathrm{d}P(\theta) = 0$ and $\iint k_P(\theta, \vartheta) \mathrm{d}P(\theta) \mathrm{d}P(\vartheta) = 0$.

Kernel Stein Discrepancy

The kernel Stein discrepancy [KSD; Chwialkowski et al., 2016, Liu et al., 2016] is defined as the worst-case integration error for the Stein RKHS $\mathcal{H}(k_P)$:

$$D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) = \sqrt{\frac{1}{m^2}\sum_{i,j=1}^m k_P(\theta_i,\theta_j) - \frac{2}{m}\sum_{i=1}^m \cancel{\int k_P(\theta,\theta_i)\mathrm{d}P(\theta)} + \cancel{\iint k_P(\theta,\vartheta)\mathrm{d}P(\theta)\mathrm{d}P(\vartheta)}}$$

Computation of the KSD does not require knowledge of the normalisation constant $\pi(y)$ and so it can be explicitly computed.

Gorham and Mackey [2017] established that (for suitable P)

$$\begin{array}{ccccc} d_{\mathrm{Dud}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) & & D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) & & d_{\mathrm{Wass}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD has $\kappa(\theta,\theta') = (\sigma^2 + \|\theta-\theta'\|^2)^{-\beta}$ being the inverse-multiquadric kernel. (d_{Dud} is the Dudley metric and metrises weak convergence. d_{Wass} is the Wasserstein metric, popular from optimal transport.)

Kernel Stein Discrepancy

The kernel Stein discrepancy [KSD; Chwialkowski et al., 2016, Liu et al., 2016] is defined as the worst-case integration error for the Stein RKHS $\mathcal{H}(k_P)$:

$$D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) = \sqrt{\frac{1}{m^2}\sum_{i,j=1}^m k_P(\theta_i,\theta_j)}$$

Computation of the KSD does not require knowledge of the normalisation constant $\pi(y)$ and so it can be explicitly computed.

Gorham and Mackey [2017] established that (for suitable P)

$$\begin{array}{ccccc} d_{\text{Dud}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) & & D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) & & d_{\text{Wass}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD has $\kappa(\theta,\theta') = (\sigma^2 + \|\theta-\theta'\|^2)^{-\beta}$ being the inverse-multiquadric kernel. (d_{Dud} is the Dudley metric and metrises weak convergence. d_{Wass} is the Wasserstein metric, popular from optimal transport.)

Kernel Stein Discrepancy

The kernel Stein discrepancy [KSD; Chwialkowski et al., 2016, Liu et al., 2016] is defined as the worst-case integration error for the Stein RKHS $\mathcal{H}(k_P)$:

$$D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) = \sqrt{\frac{1}{m^2}\sum_{i,j=1}^m k_P(\theta_i,\theta_j)}$$

Computation of the KSD does not require knowledge of the normalisation constant $\pi(y)$ and so it can be explicitly computed.

Gorham and Mackey [2017] established that (for suitable P)

$$\begin{array}{ccccc} d_{\text{Dud}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) & & D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) & & d_{\text{Wass}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD has $\kappa(\theta,\theta') = (\sigma^2 + \|\theta - \theta'\|^2)^{-\beta}$ being the inverse-multiquadric kernel. (d_{Dud} is the Dudley metric and metrises weak convergence. d_{Wass} is the Wasserstein metric, popular from optimal transport.)

Kernel Stein Discrepancy

The kernel Stein discrepancy [KSD; Chwialkowski et al., 2016, Liu et al., 2016] is defined as the worst-case integration error for the Stein RKHS $\mathcal{H}(k_P)$:

$$D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) = \sqrt{\frac{1}{m^2}\sum_{i,j=1}^m k_P(\theta_i,\theta_j)}$$

Computation of the KSD does not require knowledge of the normalisation constant $\pi(y)$ and so it can be explicitly computed.

Gorham and Mackey [2017] established that (for suitable P)

$$\begin{array}{ccccc} d_{\text{Dud}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) & & D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i=1}^m\right) & & d_{\text{Wass}}\left(\frac{1}{m}\sum_{i=1}^m \delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD has $\kappa(\theta,\theta') = (\sigma^2 + \|\theta-\theta'\|^2)^{-\beta}$ being the inverse-multiquadric kernel. (d_{Dud} is the Dudley metric and metrises weak convergence. d_{Wass} is the Wasserstein metric, popular from optimal transport.)

Sampling and Stein's Method

"Pick a sample that minimises KSD"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; D_{\mathcal{H}(k_P),P}(\{\theta_i\}_{i=1}^m)$$

Sampling is now an optimisation problem, and we can design optimisation methodology:

- Sequential grid search over Θ [Chen et al., 2018]
- Sequential stochastic search over Θ [Chen et al., 2019]
- Sequential search over a Markov chain sample path [Riabiz et al., 2022]

Sampling and Stein's Method

"Pick a sample that minimises KSD"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; D_{\mathcal{H}(k_P),P}(\{\theta_i\}_{i=1}^m)$$

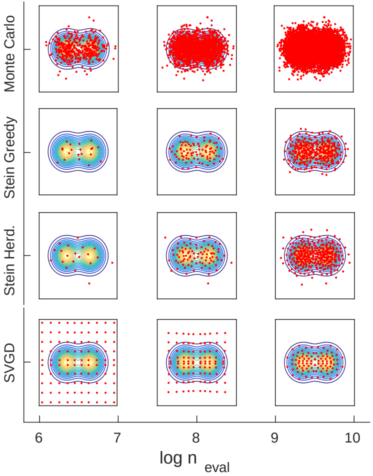
Sampling is now an optimisation problem, and we can design optimisation methodology:

- Sequential grid search over Θ [Chen et al., 2018]
- Sequential stochastic search over Θ [Chen et al., 2019]
- Sequential search over a Markov chain sample path [Riabiz et al., 2022]

Sampling and Stein's Method

"Pick a sample that minimises KSD"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; D_{\mathcal{H}(k_P),P}(\{\theta_i\}_{i=1}^m)$$



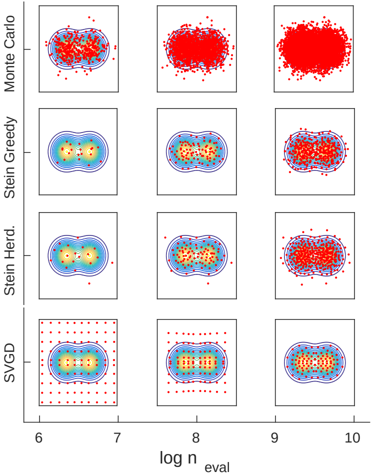
Sampling is now an optimisation problem, and we can design optimisation methodology:

- Sequential grid search over Θ [Chen et al., 2018]
- Sequential stochastic search over Θ [Chen et al., 2019]
- Sequential search over a Markov chain sample path [Riabiz et al., 2022]

Sampling and Stein's Method

"Pick a sample that minimises KSD"

Idea: $$\underset{\theta_1,\ldots,\theta_m \in \Theta}{\arg\min} \; D_{\mathcal{H}(k_P),P}(\{\theta_i\}_{i=1}^m)$$

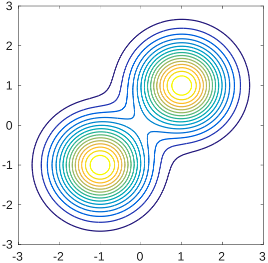


Sampling is now an optimisation problem, and we can design optimisation methodology:

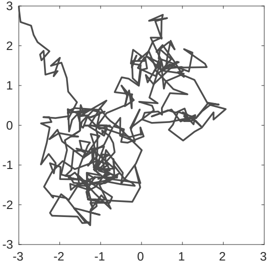
- Sequential grid search over Θ [Chen et al., 2018]
- Sequential stochastic search over Θ [Chen et al., 2019]
- **Sequential search over a Markov chain sample path** [Riabiz et al., 2022]

Optimal Thinning of MCMC Output

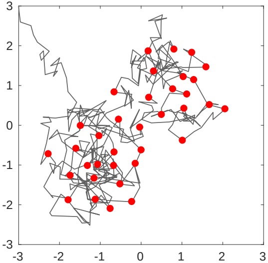
In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior P:



P



MCMC output
$(\theta_i)_{i=1}^n$



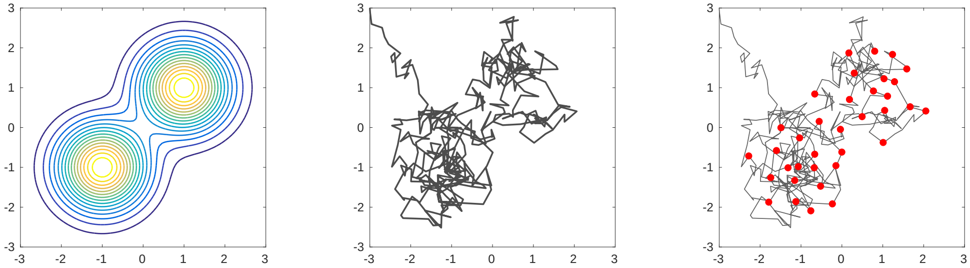
Representative Subset
$(\theta_i)_{i \in S}$

Desiderata:

- Fix problems with MCMC (*automatic identification of burn-in; mitigation of poor mixing; number of points proportional to the probability mass in a region; etc.*)
- Compressed representation of the posterior, to reduce any downstream computational load.

Optimal Thinning of MCMC Output

In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior P:



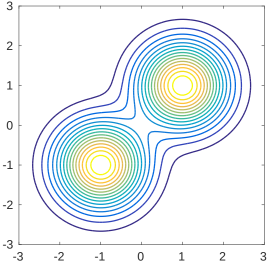
P | MCMC output $(\theta_i)_{i=1}^n$ | Representative Subset $(\theta_i)_{i \in S}$

Desiderata:

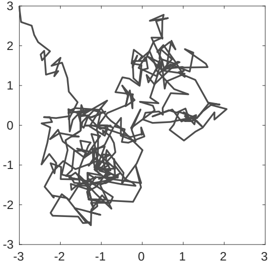
- Fix problems with MCMC (*automatic identification of burn-in; mitigation of poor mixing; number of points proportional to the probability mass in a region; etc.*)
- Compressed representation of the posterior, to reduce any downstream computational load.

Optimal Thinning of MCMC Output

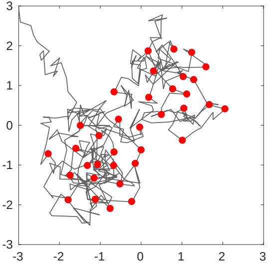
In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior P:



P



MCMC output
$(\theta_i)_{i=1}^n$



Representative Subset
$(\theta_i)_{i \in S}$

Desiderata:

- Fix problems with MCMC (*automatic identification of burn-in; mitigation of poor mixing; number of points proportional to the probability mass in a region; etc.*)
- Compressed representation of the posterior, to reduce any downstream computational load.

Stein Thinning of MCMC Output

"Greedily pick states θ_i from the MCMC output to minimise KSD"

The "Stein Thinning" algorithm produces a subset $S = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ consisting of:

$$\begin{aligned} i_1 &\in \underset{i \in \{1,\ldots,n\}}{\arg\max}\; p(\theta_i | y) \\ i_m &\in \underset{i \in \{1,\ldots,n\}}{\arg\min}\; D_{\mathcal{H}(k_P),P}\left(\{\theta_{i_j}\}_{j=1}^{m-1} \cup \{\theta_i\} \right), \qquad m \geq 2 \\ &= \underset{i \in \{1,\ldots,n\}}{\arg\min} \sum_{j=1}^{m-1} k_P(\theta_i, \theta_{i_j}) + \frac{k_P(\theta_i, \theta_i)}{2} \end{aligned}$$

This requires searching over a finite set only and can therefore be exactly implemented. The cost of selecting the mth point is $O(mn)$.

Stein Thinning of MCMC Output

"Greedily pick states θ_i from the MCMC output to minimise KSD"

The "Stein Thinning" algorithm produces a subset $S = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ consisting of:

$$
\begin{aligned}
i_1 &\in \operatorname*{arg\,max}_{i \in \{1,\dots,n\}} p(\theta_i | y) \\
i_m &\in \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} D_{\mathcal{H}(k_P),P}\left(\{\theta_{i_j}\}_{j=1}^{m-1} \cup \{\theta_i\}\right), \qquad m \geq 2 \\
&= \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} \sum_{j=1}^{m-1} k_P(\theta_i, \theta_{i_j}) + \frac{k_P(\theta_i, \theta_i)}{2}
\end{aligned}
$$

This requires searching over a finite set only and can therefore be exactly implemented. The cost of selecting the mth point is $O(mn)$.

Convergence and Bias Removal

Stein Thinning does not require MCMC to be P-invariant - as long as the relevant part of the parameter space is explored:

Theorem (Riabiz, Chen, Cockayne, Swietach, Niederer, Mackey, and CJO [2022])

Let $(\theta_i)_{i\in\mathbb{N}}$ be a Q-invariant, time-homogeneous, reversible Markov chain, such that P is absolutely continuous with respect to Q and

- $(\theta_i)_{i\in\mathbb{N}}$ *is V-uniformly ergodic with* $V(\theta) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta)\sqrt{k_P(\theta,\theta)}$
- $\sup_{i\in\mathbb{N}} \mathbb{E}[\frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)\sqrt{k_P(\theta_i,\theta_i)}V(\theta_i)] < \infty$
- $\exists \gamma > 0$ *s.t.* $b := \sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma \max(1, \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)^2) k_P(\theta_i,\theta_i)}] < \infty$.

Then the output of Stein Thinning satisfies

$$P_{\mathrm{ST}} := \frac{1}{m}\sum_{i\in S}\delta(\theta_i) \Rightarrow P$$

almost surely as $n, m \to \infty$ with $m \leq n$ and $\log(n) = O(m^{\beta/2})$ for some $\beta < 1$.

Convergence and Bias Removal

Stein Thinning does not require MCMC to be P-invariant - as long as the relevant part of the parameter space is explored:

Theorem (Riabiz, Chen, Cockayne, Swietach, Niederer, Mackey, and CJO [2022])

Let $(\theta_i)_{i \in \mathbb{N}}$ be a Q-invariant, time-homogeneous, reversible Markov chain, such that P is absolutely continuous with respect to Q and

- $(\theta_i)_{i \in \mathbb{N}}$ *is V-uniformly ergodic with* $V(\theta) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta)\sqrt{k_P(\theta, \theta)}$
- $\sup_{i \in \mathbb{N}} \mathbb{E}[\frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)\sqrt{k_P(\theta_i, \theta_i)}V(\theta_i)] < \infty$
- $\exists \gamma > 0$ *s.t.* $b := \sup_{i \in \mathbb{N}} \mathbb{E}[e^{\gamma \max(1, \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)^2) k_P(\theta_i, \theta_i)}] < \infty$.

Then the output of Stein Thinning satisfies

$$P_{\mathrm{ST}} := \frac{1}{m} \sum_{i \in S} \delta(\theta_i) \Rightarrow P$$

almost surely as $n, m \to \infty$ with $m \leq n$ and $\log(n) = O(m^{\beta/2})$ for some $\beta < 1$.

Convergence and Bias Removal

Stein Thinning does not require MCMC to be P-invariant - as long as the relevant part of the parameter space is explored:

Theorem (Riabiz, Chen, Cockayne, Swietach, Niederer, Mackey, and CJO [2022])

Let $(\theta_i)_{i \in \mathbb{N}}$ be a Q-invariant, time-homogeneous, reversible Markov chain, such that P is absolutely continuous with respect to Q and

- $(\theta_i)_{i \in \mathbb{N}}$ *is V-uniformly ergodic with* $V(\theta) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta)\sqrt{k_P(\theta, \theta)}$
- $\sup_{i \in \mathbb{N}} \mathbb{E}[\frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)\sqrt{k_P(\theta_i, \theta_i)} V(\theta_i)] < \infty$
- $\exists \gamma > 0$ *s.t.* $b := \sup_{i \in \mathbb{N}} \mathbb{E}[e^{\gamma \max(1, \frac{\mathrm{d}P}{\mathrm{d}Q}(\theta_i)^2) k_P(\theta_i, \theta_i)}] < \infty$.

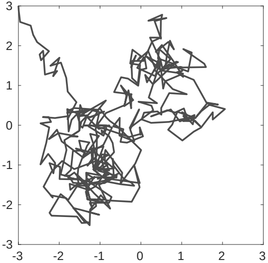
Then the output of Stein Thinning satisfies

$$P_{\mathrm{ST}} := \frac{1}{m} \sum_{i \in S} \delta(\theta_i) \Rightarrow P$$

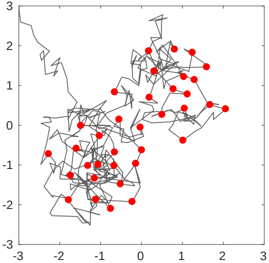
almost surely as $n, m \to \infty$ *with* $m \leq n$ *and* $\log(n) = O(m^{\beta/2})$ *for some* $\beta < 1$.

Stein Thinning of MCMC Output

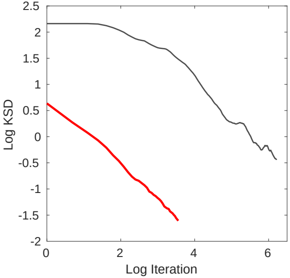
The figures we saw before were actually produced by Stein Thinning!



MCMC output
$(\theta_i)_{i=1}^n$



Representative Subset
$\{\theta_i\}_{i \in S}$



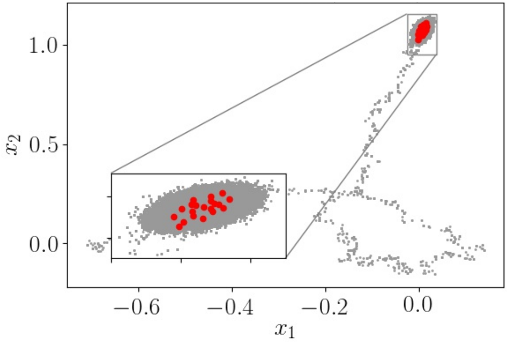
Performance
$m \mapsto D_{\mathcal{H}(k_P),P}\left(\{\theta_i\}_{i \in S}\right)$
(log-scales used)

Full details in:

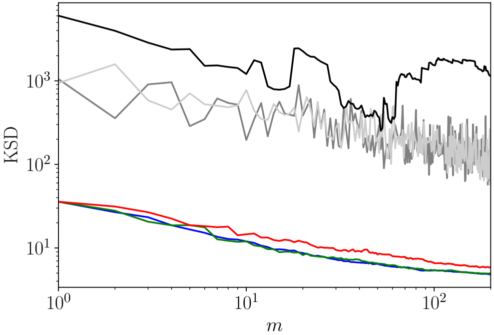
- M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and CJO. Optimal thinning of MCMC output. *JRSSB*, 2022

Illustrative Application to Differential Equation Constrained Inverse Problems

Goodwin oscillator; $d = 4$ parameters to be estimated. (Red dots are Stein Thinning, while gray dots are MCMC.)



Cardiac model; $d = 38$ parameters to be esitmated. (Blue, red, and green are Stein Thinning, while black are MCMC.)



Stein-Thinning.org

Stein Thinning

Optimally thinning of output from a sampling procedure, such as MCMC. Here the red samples are automatically chosen by Stein Thinning to provide a more accurate approximation to the distributional target, compared with the original MCMC output. [Read more]

View the Project on GitHub
wilson-ye-chen/stein_thinning_start

About

Stein Thinning is a tool for post-processing the output of a sampling procedure, such as Markov chain Monte Carlo (MCMC). It aims to minimise a Stein discrepancy, selecting a subsequence of samples that best represent the distributional target.

The user provides two arrays: one containing the samples and another containing the corresponding gradients of the log-target. Stein Thinning returns a vector of indices, indicating which samples were selected.

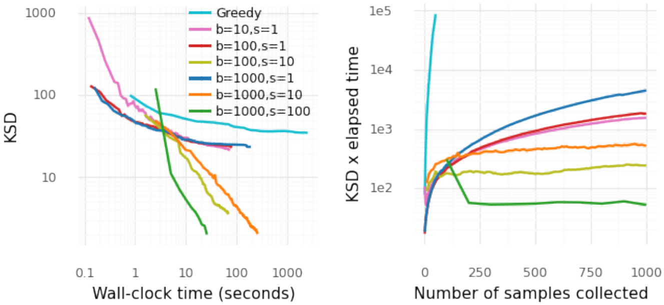
In favourable circumstances, Stein Thinning is able to:

- automatically identify and remove the burn-in period from MCMC,
- perform bias-removal for biased sampling procedures,
- provide improved approximations of the distributional target,
- offer a compressed representation of sample-based output.

Non-Myopic and Batch Extensions to Stein Thinning

Greedy selection may be sub-optimal. Also, the cost of selecting m points from n using Stein Thinning is high, at $O(m^2 n)$.

- A **non-myopic** algorithm selects s points simultaneously.
- A **mini-batch** algorithm searches over a subset of $b \ll n$ candidates at each step.



Full details in:

- O. Teymur, J. Gorham, M. Riabiz, and CJO. Optimal quantisation of probability measures using maximum mean discrepancy.
  In *AISTATS*, 2021

Sampling and Stein's Method: Broader Context

Broader Context: Optimisation over $\mathcal{P}(\Theta)$

Going beyond optimisation in Θ, we can consider optimisation in $\mathcal{P}(\Theta)$:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{\|f\|_{\mathcal{H}(k_P)} \leq 1} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020], ...

  $$\text{Given } \{\theta_i\}_{i=1}^n, \text{ construct } P_{\text{SIS}} := \sum_{i=1}^n w_i \delta(\theta_i) \text{ where } w \in \operatorname*{arg\,min}_{\substack{w_1,\dots,w_n \geq 0 \\ w_1+\dots+w_n=1}} D_{\mathcal{H}(k_P),P}\left(\sum_{i=1}^n w_i \delta(\theta_i)\right)$$

  Complexity $= O(n^3)$ but $P_{\text{ST}} \to P_{\text{SIS}}$ as $m \to \infty$ for n fixed.
- **Variational Inference:** Ranganath et al. [2016], Hu et al. [2018], Fisher et al. [2021], ...

  $$\min_{Q \in \mathcal{Q}} D_{\mathcal{H}(k_P),P}(Q), \qquad \text{(e.g.) } \mathcal{Q} = \{T_\# Q_0 : T \text{ a neural network}\}$$

  Avoids the requirement in VI that T be a diffeomorphism (i.e. no need for normalising flows!).
- **Gradient Flow:** Korba et al. [2021]

  $$\frac{\partial Q_t}{\partial t} + \text{div}(Q_t v_{Q_t}) = 0, \qquad v_{Q_t} = -\nabla_{W_2} \mathcal{F}(Q_t), \qquad \mathcal{F}(Q) = \frac{1}{2} D_{\mathcal{H}(k_P),P}(Q)^2$$

  *not the same as SVGD [see Liu, 2017].

Broader Context: Optimisation over $\mathcal{P}(\Theta)$

Going beyond optimisation in Θ, we can consider optimisation in $\mathcal{P}(\Theta)$:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{\|f\|_{\mathcal{H}(k_P)} \leq 1} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020], ...

  Given $\{\theta_i\}_{i=1}^n$, construct $P_{\text{SIS}} := \sum_{i=1}^n w_i \delta(\theta_i)$ where $w \in \underset{\substack{w_1,\dots,w_n \geq 0 \\ w_1+\dots+w_n=1}}{\arg\min} D_{\mathcal{H}(k_P),P}\left(\sum_{i=1}^n w_i \delta(\theta_i)\right)$

  Complexity $= O(n^3)$ but $P_{\text{ST}} \to P_{\text{SIS}}$ as $m \to \infty$ for n fixed.

- **Variational Inference:** Ranganath et al. [2016], Hu et al. [2018], Fisher et al. [2021], ...

  $$\min_{Q \in \mathcal{Q}} D_{\mathcal{H}(k_P),P}(Q), \qquad \text{(e.g.) } \mathcal{Q} = \{T_{\#}Q_0 : T \text{ a neural network}\}$$

  Avoids the requirement in VI that T be a diffeomorphism (i.e. no need for normalising flows!).

- **Gradient Flow:** Korba et al. [2021]

  $$\frac{\partial Q_t}{\partial t} + \text{div}(Q_t v_{Q_t}) = 0, \qquad v_{Q_t} = -\nabla_{W_2}\mathcal{F}(Q_t), \qquad \mathcal{F}(Q) = \frac{1}{2} D_{\mathcal{H}(k_P),P}(Q)^2$$

  *not the same as SVGD [see Liu, 2017].

Broader Context: Optimisation over $\mathcal{P}(\Theta)$

Going beyond optimisation in Θ, we can consider optimisation in $\mathcal{P}(\Theta)$:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{\|f\|_{\mathcal{H}(k_P)} \leq 1} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020], ...

  Given $\{\theta_i\}_{i=1}^n$, construct $P_{\text{SIS}} := \sum_{i=1}^n w_i \delta(\theta_i)$ where $w \in \underset{\substack{w_1,\dots,w_n \geq 0 \\ w_1+\dots+w_n=1}}{\arg\min} D_{\mathcal{H}(k_P),P}\left(\sum_{i=1}^n w_i \delta(\theta_i)\right)$

  Complexity $= O(n^3)$ but $P_{\text{ST}} \to P_{\text{SIS}}$ as $m \to \infty$ for n fixed.

- **Variational Inference:** Ranganath et al. [2016], Hu et al. [2018], Fisher et al. [2021], ...

  $$\min_{Q \in \mathcal{Q}} D_{\mathcal{H}(k_P),P}(Q), \qquad \text{(e.g.)} \ \mathcal{Q} = \{T_\# Q_0 : T \text{ a neural network}\}$$

  Avoids the requirement in VI that T be a diffeomorphism (i.e. no need for normalising flows!).

- **Gradient Flow:** Korba et al. [2021]

  $$\frac{\partial Q_t}{\partial t} + \text{div}(Q_t v_{Q_t}) = 0, \qquad v_{Q_t} = -\nabla_{W_2} \mathcal{F}(Q_t), \qquad \mathcal{F}(Q) = \frac{1}{2} D_{\mathcal{H}(k_P),P}(Q)^2$$

  *not the same as SVGD [see Liu, 2017].

Broader Context: Optimisation over $\mathcal{P}(\Theta)$

Going beyond optimisation in Θ, we can consider optimisation in $\mathcal{P}(\Theta)$:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{\|f\|_{\mathcal{H}(k_P)} \leq 1} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020], ...

  Given $\{\theta_i\}_{i=1}^n$, construct $P_{\mathrm{SIS}} := \sum_{i=1}^n w_i \delta(\theta_i)$ where $w \in \operatorname*{arg\,min}_{\substack{w_1,\dots,w_n \geq 0 \\ w_1+\dots+w_n=1}} D_{\mathcal{H}(k_P),P}\left(\sum_{i=1}^n w_i \delta(\theta_i)\right)$

  Complexity $= O(n^3)$ but $P_{\mathrm{ST}} \to P_{\mathrm{SIS}}$ as $m \to \infty$ for n fixed.
- **Variational Inference:** Ranganath et al. [2016], Hu et al. [2018], Fisher et al. [2021], ...

  $$\min_{Q \in \mathcal{Q}} D_{\mathcal{H}(k_P),P}(Q), \qquad \text{(e.g.) } \mathcal{Q} = \{T_\# Q_0 : T \text{ a neural network}\}$$

  Avoids the requirement in VI that T be a diffeomorphism (i.e. no need for normalising flows!).
- **Gradient Flow:** Korba et al. [2021]

  $$\frac{\partial Q_t}{\partial t} + \mathrm{div}(Q_t v_{Q_t}) = 0, \qquad v_{Q_t} = -\nabla_{W_2} \mathcal{F}(Q_t), \qquad \mathcal{F}(Q) = \frac{1}{2} D_{\mathcal{H}(k_P),P}(Q)^2$$

  *not the same as SVGD [see Liu, 2017].

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Sampling with Stein Discrepancies

For any Stein characterisation $(\mathcal{A}, \mathcal{F})$ we can consider an associated Stein discrepancy [Gorham and Mackey, 2015]:

$$D_{\mathcal{H}(k_P),P}(Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_{\vartheta \sim Q}[f(\vartheta)]|$$

- **Beyond Euclidean State Spaces:** Riemannian manifolds [Barp et al., 2022, Le et al., 2020], discrete spaces [Xu and Reinert, 2021], ...
- **Beyond Kernel Stein Sets:** bounded Lipschitz [Gorham and Mackey, 2015], neural network [Grathwohl et al., 2020], ...
- **Beyond the Canonical Stein Operator:** diffusion Stein operators [Gorham et al., 2019], ...
- **Scalable Stein Discrepancies:** random features [Huggins and Mackey, 2018], data sub-sampling [Gorham et al., 2020], ...

The interaction between the sampling algorithms we have seen and the choice of Stein Discrepancy is not well-understood.

Broader Context: Alternatives to Direct Minimisation of Stein Discrepancy

- **Stein Variational Gradient Descent:** Liu and Wang [2016], Liu [2017], Liu and Zhu [2018], Detommaso et al. [2018], ...
- **MCMC with Stein Control Variates:** Assaraf and Caffarel [1999], Mira et al. [2013], CJO et al. [2017], Belomestny et al. [2017], South et al. [2022], ...

  $$\text{Given a QoI } f, \text{ seek } (u, c) \text{ such that } c + \frac{\nabla \cdot (p \nabla u)}{p} = f. \qquad \text{Then } c = \mathbb{E}_{\vartheta \sim P}[f(\vartheta)].$$

  In practice, an approximate solution u gives rise to a control variate $v = \nabla \cdot (p \nabla u)/p$ for use in MCMC.

A slightly more detailed introduction can be found in the survey:

- A. Anastasiou et al. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021

Thank you for your attention!

Broader Context: Alternatives to Direct Minimisation of Stein Discrepancy

- **Stein Variational Gradient Descent:** Liu and Wang [2016], Liu [2017], Liu and Zhu [2018], Detommaso et al. [2018], ...
- **MCMC with Stein Control Variates:** Assaraf and Caffarel [1999], Mira et al. [2013], CJO et al. [2017], Belomestny et al. [2017], South et al. [2022], ...

$$\text{Given a QoI } f, \text{ seek } (u, c) \text{ such that } c + \frac{\nabla \cdot (p \nabla u)}{p} = f. \qquad \text{Then } c = \mathbb{E}_{\vartheta \sim P}[f(\vartheta)].$$

In practice, an approximate solution u gives rise to a control variate $v = \nabla \cdot (p \nabla u)/p$ for use in MCMC.

A slightly more detailed introduction can be found in the survey:

- A. Anastasiou et al. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021

Thank you for your attention!

Broader Context: Alternatives to Direct Minimisation of Stein Discrepancy

- **Stein Variational Gradient Descent:** Liu and Wang [2016], Liu [2017], Liu and Zhu [2018], Detommaso et al. [2018], ...
- **MCMC with Stein Control Variates:** Assaraf and Caffarel [1999], Mira et al. [2013], CJO et al. [2017], Belomestny et al. [2017], South et al. [2022], ...

  Given a QoI f, seek (u, c) such that $c + \frac{\nabla \cdot (p \nabla u)}{p} = f$. Then $c = \mathbb{E}_{\vartheta \sim P}[f(\vartheta)]$.

  In practice, an approximate solution u gives rise to a control variate $v = \nabla \cdot (p \nabla u)/p$ for use in MCMC.

A slightly more detailed introduction can be found in the survey:

- A. Anastasiou et al. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021

Thank you for your attention!

Broader Context: Alternatives to Direct Minimisation of Stein Discrepancy

- **Stein Variational Gradient Descent:** Liu and Wang [2016], Liu [2017], Liu and Zhu [2018], Detommaso et al. [2018], ...
- **MCMC with Stein Control Variates:** Assaraf and Caffarel [1999], Mira et al. [2013], CJO et al. [2017], Belomestny et al. [2017], South et al. [2022], ...

  $$\text{Given a QoI } f, \text{ seek } (u, c) \text{ such that } c + \frac{\nabla \cdot (p \nabla u)}{p} = f. \qquad \text{Then } c = \mathbb{E}_{\vartheta \sim P}[f(\vartheta)].$$

  In practice, an approximate solution u gives rise to a control variate $v = \nabla \cdot (p \nabla u)/p$ for use in MCMC.

A slightly more detailed introduction can be found in the survey:

- A. Anastasiou et al. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021

Thank you for your attention!

References I

A. Anastasiou et al. Stein's method meets statistics: A review of some recent developments. *arXiv:2105.03481*, 2021.

R. Assaraf and M. Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical Review Letters*, 83(23): 4682, 1999.

A. Barp, CJO, E. Porcu, and M. Girolami. A riemann–stein kernel method. *Bernoulli*, 2022.

D. Belomestny, L. Iosipoi, and N. Zhivotovskiy. Variance reduction via empirical variance minimization: convergence and complexity. *arXiv:1712.04667*, 2017.

W. Chen, L. Mackey, J. Gorham, F. Briol, and CJO. Stein points. In *ICML*, 2018.

W. Y. Chen, A. Barp, F. X. Briol, J. Gorham, L. Mackey, and CJO. Stein point Markov chain Monte Carlo. In *ICML*, 2019.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

CJO, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *JRSSB*, 79(3):695–718, 2017.

G. Detommaso, T. Cui, Y. Marzouk, R. Scheichl, and A. Spantini. A Stein variational Newton method. In *NeurIPS*, 2018.

M. A. Fisher, T. Nolan, M. M. Graham, D. Prangle, and CJO. Measure transport with kernel Stein discrepancy. *AISTATS*, 2021.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *NeurIPS*, 2015.

J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In *ICML*, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *AoAP*, 29 (5):2884–2928, 2019.

J. Gorham, A. Raj, and L. Mackey. Stochastic Stein discrepancies. In *NeurIPS*, 2020.

References II

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

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

T. Hu, Z. Chen, H. Sun, J. Bai, M. Ye, and G. Cheng. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

J. Huggins and L. Mackey. Random feature Stein discrepancies. In *NeurIPS*, 2018.

A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein discrepancy descent. In *ICML*, pages 5719–5730, 2021.

H. Le, A. Lewis, K. Bharath, and C. Fallaize. A diffusion approach to Stein's method on Riemannian manifolds. *arXiv:2003.11497*, 2020.

C. Liu and J. Zhu. Riemannian Stein variational gradient descent for bayesian inference. In *AAAI Conference on AI*, volume 32, 2018.

Q. Liu. Stein Variational Gradient Descent as Gradient Flow. In *NeurIPS*, pages 3118–3126, 2017.

Q. Liu and J. D. Lee. Black-box importance sampling. In *AISTATS*, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *NeurIPS*, 2016.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

A. Mira, R. Solgi, and D. Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator variational inference. In *NeurIPS*, volume 29, 2016.

M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and CJO. Optimal thinning of MCMC output. *JRSSB*, 2022.

References III

L. F. South, T. Karvonen, C. Nemeth, M. Girolami, and CJO. Semi-exact control functionals from Sard's method. *Biometrika*, 2022.

O. Teymur, J. Gorham, M. Riabiz, and CJO. Optimal quantisation of probability measures using maximum mean discrepancy. In *AISTATS*, 2021.

W. Xu and G. Reinert. A stein goodness-of-test for exponential random graph models. In *AISTATS*, pages 415–423, 2021.