Optimal Thinning of MCMC Output

Chris. J. Oates
Newcastle University
Alan Turing Institute

February 2022 @ DataSig Seminar Series

Lloyd's Register Foundation

Newcastle University

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.

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.

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.

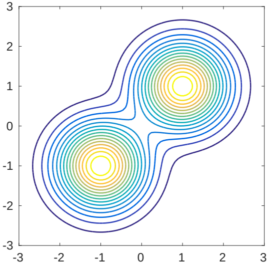
Sampling from P via Markov chain Monte Carlo (MCMC) is a popular approach which requires only evaluation of the un-normalised form

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

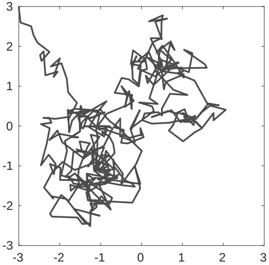
but it is not a silver bullet.

An Ideal Post-Processing Method

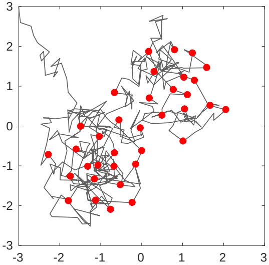
In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior P:



P



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



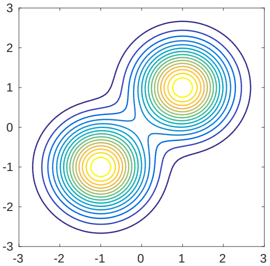
Representative Subset
$(\theta_i)_{i \in S}$

Desiderata:

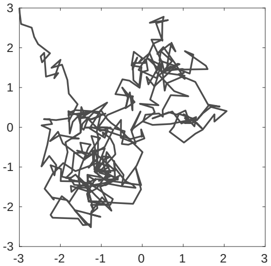
- Fix problems with MCMC (*automatic identification of burn-in; mitigation of poor mixing; number of points proportional to the probability mass in a region; etc.*)
- Compressed representation of the posterior, to reduce any downstream computational load.

An Ideal Post-Processing Method

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:

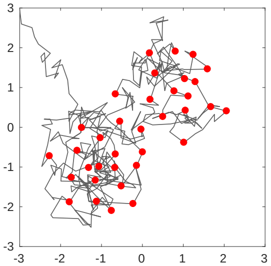


P



MCMC output

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



Representative Subset

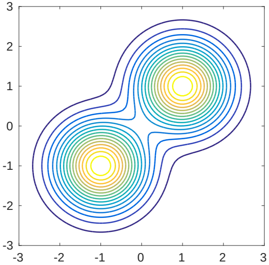
$(\theta_i)_{i \in S}$

Desiderata:

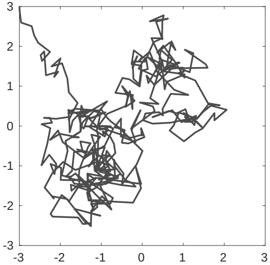
- Fix problems with MCMC (*automatic identification of burn-in; mitigation of poor mixing; number of points proportional to the probability mass in a region; etc.*)
- Compressed representation of the posterior, to reduce any downstream computational load.

An Ideal Post-Processing Method

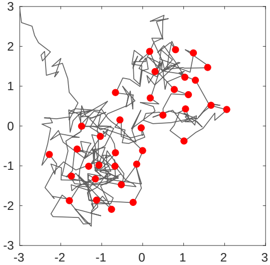
In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior P:



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

"Pick a representative subset from the MCMC output"

Idea:
$$\operatorname*{arg\,min}_{\substack{S\subset\{1,\dots,n\}\\|S|=m}} \underbrace{\mathrm{diff}}_{(*)}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right)$$

Remarks:

- "Nice idea, but we don't have access to P."
- "Combinatorial optimisation is a hard problem."

Our strategy is to use **Stein's Method** to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$.

Optimal Thinning of MCMC Output

“Pick a representative subset from the MCMC output”

Idea:
$$\mathop{\arg\min}_{\substack{S \subset \{1,\dots,n\} \\ |S|=m}} \underbrace{\mathrm{diff}}_{(*)}\left(\frac{1}{m} \sum_{i \in S} \delta(\theta_i), P \right)$$

Remarks:

- “Nice idea, but we don’t have access to P.”
- “Combinatorial optimisation is a hard problem.”

Our strategy is to use **Stein’s Method** to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$.

Optimal Thinning of MCMC Output

"Pick a representative subset from the MCMC output"

Idea:
$$\operatorname*{arg\,min}_{\substack{S \subset \{1,\ldots,n\} \\ |S|=m}} \underbrace{\mathrm{diff}}_{(*)} \left(\frac{1}{m} \sum_{i \in S} \delta(\theta_i), P \right)$$

Remarks:

- "Nice idea, but we don't have access to P."
- "Combinatorial optimisation is a hard problem."

Our strategy is to use **Stein's Method** to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$.

Optimal Thinning of MCMC Output

"Pick a representative subset from the MCMC output"

Idea:

$$\operatorname*{arg\,min}_{\substack{S \subset \{1,\dots,n\} \\ |S|=m}} \underbrace{\mathrm{diff}}_{(*)} \left(\frac{1}{m} \sum_{i \in S} \delta(\theta_i), P \right)$$

Remarks:

- "Nice idea, but we don't have access to P."
- "Combinatorial optimisation is a hard problem."

Our strategy is to use **Stein's Method** to manufacture a function $(*)$ that can be computed without the normalisation constant $\pi(y)$.

Outline

Kernel Stein Discrepancy

Stein Thinning of MCMC Output

Stein's Method in Computational Statistics

Kernel Stein Discrepancy

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\|\cdot\|_{\mathcal{K}}$:

$$\begin{aligned} \operatorname{diff}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) &:= \sup_{\|f\|_{\mathcal{K}}\le 1}\left|\frac{1}{m}\sum_{i\in S} f(\theta_i) - \mathbb{E}_{\vartheta\sim P}[f(\vartheta)]\right| \\ &=: D_{\mathcal{K},P}\left(\{\theta_i\}_{i\in S}\right) \end{aligned}$$

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta,\cdot)\mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\|\cdot\|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta,\cdot)\mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

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

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\| \cdot \|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

Let's try to compute this:

$$D_{\mathcal{K},P}(\{\theta_i\}_{i \in S})^2 = \frac{1}{m^2} \sum_{i,j \in S} k(\theta_i, \theta_j) - \frac{2}{m} \sum_{i \in S} k_P(\theta_i) + k_{P,P}$$

where $k_P := \int k(\theta, \cdot) \mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

Approximation in Reproducing Kernel Hilbert Spaces (RKHS)

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

Consider an **integral probability metric** based on $\|\cdot\|_{\mathcal{K}}$:

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

which is known as the *worst-case integration error* for the RKHS $\mathcal{K}$.

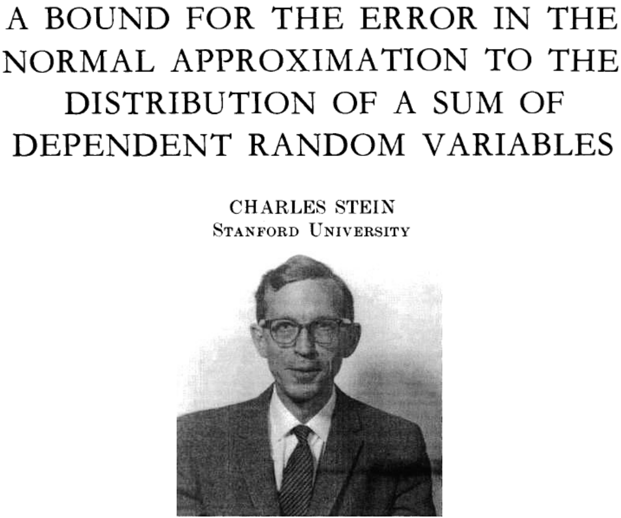
Let's try to compute this:

$$D_{\mathcal{K},P}(\{\theta_i\}_{i\in S})^2 \quad = \quad \frac{1}{m^2}\sum_{i,j\in S} k(\theta_i, \theta_j) - \frac{2}{m}\sum_{i\in S} k_P(\theta_i) + k_{P,P}$$

where $k_P := \int k(\theta, \cdot)\mathrm{d}P(\theta) \in \mathcal{K}$ and $k_{P,P} := \int k_P \mathrm{d}P$.

Problem: We need to choose k carefully, so that k_P and $k_{P,P}$ can be evaluated. How?

A Brief History of Stein



A BOUND FOR THE ERROR IN THE NORMAL APPROXIMATION TO THE DISTRIBUTION OF A SUM OF DEPENDENT RANDOM VARIABLES

CHARLES STEIN
STANFORD UNIVERSITY

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 Class $\mathcal{F}$, if it holds that

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

Example (Stein, 1972)

- $P = N(\mu, \sigma^2)$ with density function $p(x)$
- $\mathcal{A} : f \mapsto \frac{\nabla(fp)}{p}$
- $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} \text{ s.t. } \nabla(fp) \in L^1(\mathbb{R}) \text{ and } \lim_{x \searrow -\infty} f(\theta)p(\theta) = \lim_{\theta \nearrow +\infty} f(\theta)p(\theta)\}.$

Stein Characterisation

Definition (Stein Characterisation)

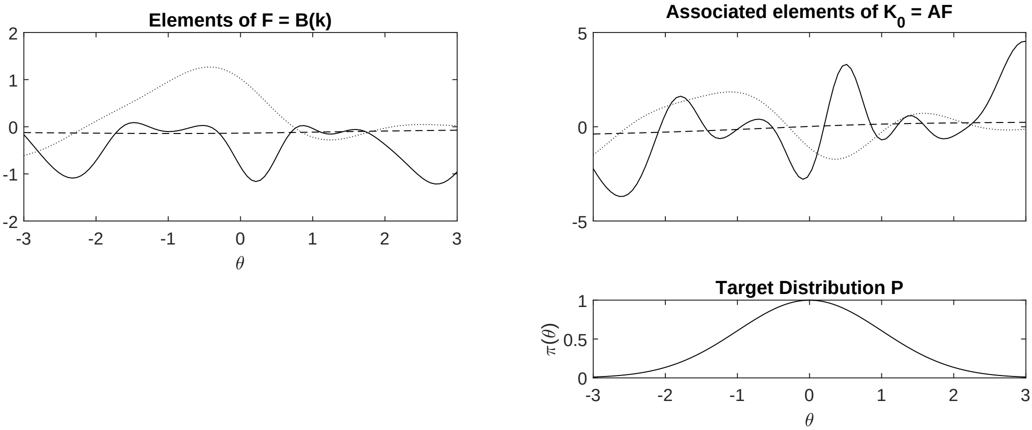
A distribution P is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Class $\mathcal{F}$, if it holds that

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

Example (Stein, 1972)

- $P = N(\mu, \sigma^2)$ with density function $p(x)$
- $\mathcal{A} : f \mapsto \frac{\nabla(fp)}{p}$
- $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} \text{ s.t. } \nabla(fp) \in L^1(\mathbb{R}) \text{ and } \lim_{x \searrow -\infty} f(\theta)p(\theta) = \lim_{\theta \nearrow +\infty} f(\theta)p(\theta)\}.$

Stein Characterisation


Elements of F = B(k)

Associated elements of K_0 = AF

Target Distribution P

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta \sim P}[(\Delta k(\vartheta, \vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta, \theta') &= \nabla_\theta \nabla_{\theta'} k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \nabla_{\theta'} k(\theta, \theta') \\ &\quad + \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} \nabla_\theta k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} k(\theta, \theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta \sim P}[(\Delta k(\vartheta, \vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A} f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta, \theta') &= \nabla_\theta \nabla_{\theta'} k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \nabla_{\theta'} k(\theta, \theta') \\ &\quad + \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} \nabla_\theta k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} k(\theta, \theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta \sim P}[(\Delta k(\vartheta, \vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta, \theta') \quad &= \quad \nabla_\theta \nabla_{\theta'} k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \nabla_{\theta'} k(\theta, \theta') \\ &\quad + \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} \nabla_\theta k(\theta, \theta') + \frac{\nabla_\theta p(\theta)}{p(\theta)} \frac{\nabla_{\theta'} p(\theta')}{p(\theta')} k(\theta, \theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta\sim P}[(\Delta k(\vartheta,\vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A},\mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta,\theta') \quad &= \quad \nabla_\theta \nabla_{\theta'} k(\theta,\theta') + [\nabla_\theta \log p(\theta)]\ \nabla_{\theta'} k(\theta,\theta') \\ &\quad + [\nabla_{\theta'} \log p(\theta') \nabla_\theta]\ k(\theta,\theta') + [\nabla_\theta \log p(\theta)]\ [\nabla_{\theta'} \log p(\theta')]\ k(\theta,\theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta \sim P}[(\Delta k(\vartheta, \vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta, \theta') &= \nabla_\theta \nabla_{\theta'} k(\theta, \theta') + [\nabla_\theta \log p(\theta)] \, \nabla_{\theta'} k(\theta, \theta') \\ &\quad + [\nabla_{\theta'} \log p(\theta') \nabla_\theta] \, k(\theta, \theta') + [\nabla_\theta \log p(\theta)] \, [\nabla_{\theta'} \log p(\theta')] \, k(\theta, \theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Stein Operators in Hilbert Spaces

(Going to stick to $d = 1$.)

Theorem (Chwialkowski et al. [2016])

Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_{\vartheta \sim P}[(\Delta k(\vartheta, \vartheta))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}, \mathcal{F})$, consisting of

$$\mathcal{A}f = \frac{\nabla(fp)}{p}, \qquad \mathcal{F} = \mathcal{B}(k) := \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leq 1\}.$$

Theorem (O, Girolami and Chopin [2017])

The functions $\mathcal{A}f$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_0 := \mathcal{A}\mathcal{K}$ with kernel

$$\begin{aligned} k_0(\theta, \theta') \quad = \quad & \nabla_\theta \nabla_{\theta'} k(\theta, \theta') + [\nabla_\theta \log p(\theta)] \; \nabla_{\theta'} k(\theta, \theta') \\ & + [\nabla_{\theta'} \log p(\theta') \nabla_\theta] \; k(\theta, \theta') + [\nabla_\theta \log p(\theta)] \; [\nabla_{\theta'} \log p(\theta')] \; k(\theta, \theta'). \end{aligned}$$

In particular, under regularity conditions, $(k_0)_P = 0$ and $(k_0)_{P,P} = 0$ are trivially computed.

Solution: Use k_0 in an integral probability metric!

Kernel Stein Discrepancy

The kernel Stein discrepancy [KSD; Chwialkowski et al., 2016, Liu et al., 2016] is just the worst-case integration error for the Stein RKHS $\mathcal{K}_0$:

$$\begin{aligned} \mathrm{KSD}\left(\frac{1}{m}\sum_{i \in S} \delta(\theta_i), P\right) &:= D_{\mathcal{K}_0,P}\left(\{\theta_i\}_{i \in S}\right) \\ &= \sqrt{\frac{1}{m^2}\sum_{i,j \in S} k_0(\theta_i, \theta_j) - \frac{2}{m}\sum_{i \in S} \cancel{(k_0)_P(\theta_i)} + \cancel{(k_0)_{P,P}}} \end{aligned}$$

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

$$\begin{array}{ccccc} d_{\text{Dud}}\left(\frac{1}{m}\sum_{i \in S}\delta(\theta_i), P\right) & & \mathrm{KSD}\left(\frac{1}{m}\sum_{i \in S}\delta(\theta_i), P\right) & & d_{\text{Wass}}\left(\frac{1}{m}\sum_{i \in S}\delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD is based on $k(\theta, \theta')$ 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 just the worst-case integration error for the Stein RKHS $\mathcal{K}_0$:

$$\begin{aligned} \mathrm{KSD}\left(\frac{1}{m}\sum_{i \in S}\delta(\theta_i), P\right) &:= D_{\mathcal{K}_0,P}\left(\{\theta_i\}_{i\in S}\right) \\ &= \sqrt{\frac{1}{m^2}\sum_{i,j\in S} k_0(\theta_i,\theta_j)} \end{aligned}$$

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

$$\begin{tikzcd} d_{\mathrm{Dud}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \arrow[d] & & \mathrm{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \arrow[d] & & d_{\mathrm{Wass}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \arrow[d] \\ 0 & \Leftarrow & 0 & \Leftarrow & 0 \end{tikzcd}$$

when the KSD is based on $k(\theta, \theta')$ 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 just the worst-case integration error for the Stein RKHS $\mathcal{K}_0$:

$$\begin{aligned} \mathrm{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \quad &:= \quad D_{\mathcal{K}_0,P}\left(\{\theta_i\}_{i\in S}\right) \\ &= \quad \sqrt{\frac{1}{m^2}\sum_{i,j\in S} k_0(\theta_i,\theta_j)} \end{aligned}$$

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

$$\begin{array}{ccccc} d_{\text{Dud}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) & & \mathrm{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) & & d_{\text{Wass}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD is based on $k(\theta,\theta')$ 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 just the worst-case integration error for the Stein RKHS $\mathcal{K}_0$:

$$\begin{aligned} \mathrm{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) &:= D_{\mathcal{K}_0,P}\left(\{\theta_i\}_{i\in S}\right) \\ &= \sqrt{\frac{1}{m^2}\sum_{i,j\in S} k_0(\theta_i,\theta_j)} \end{aligned}$$

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

$$\begin{array}{ccccc} d_{\mathrm{Dud}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) & & \mathrm{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) & & d_{\mathrm{Wass}}\left(\frac{1}{m}\sum_{i\in S}\delta(\theta_i), P\right) \\ \downarrow & \Leftarrow & \downarrow & \Leftarrow & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

when the KSD is based on $k(\theta,\theta')$ 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.)

Stein Thinning of MCMC Output

Stein Thinning of MCMC Output

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

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

$$\begin{aligned} i_1 &\in \operatorname*{arg\,max}_{i \in \{1,\dots,n\}} p(\theta_i | y) \\ i_m &\in \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} \mathrm{KSD}\left(\frac{1}{m} \sum_{j=1}^{m-1} \delta(\theta_{i_j}) + \frac{1}{m} \delta(\theta_i), P \right), \qquad m \geq 2 \\ &= \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} \sum_{j=1}^{m-1} k_0(\theta_i, \theta_{i_j}) + \frac{k_0(\theta_i, \theta_i)}{2} \end{aligned}$$

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"

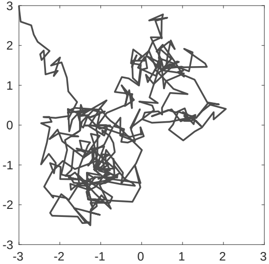
The "Stein Thinning" algorithm that we propose produces a subset $S = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ consisting of:

$$\begin{aligned}
i_1 &\in \operatorname*{arg\,max}_{i \in \{1,\dots,n\}} p(\theta_i | y) \\
i_m &\in \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} \text{KSD}\left(\frac{1}{m} \sum_{j=1}^{m-1} \delta(\theta_{i_j}) + \frac{1}{m} \delta(\theta_i), P \right), \qquad m \geq 2 \\
&= \operatorname*{arg\,min}_{i \in \{1,\dots,n\}} \sum_{j=1}^{m-1} k_0(\theta_i, \theta_{i_j}) + \frac{k_0(\theta_i, \theta_i)}{2}
\end{aligned}$$

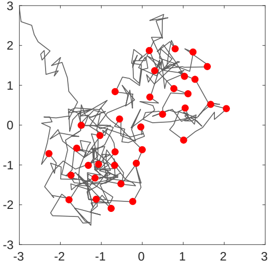
This requires searching over a finite set only and can therefore be exactly implemented. The cost of selecting the mth point is $O(mn)$.

Stein Thinning of MCMC Output

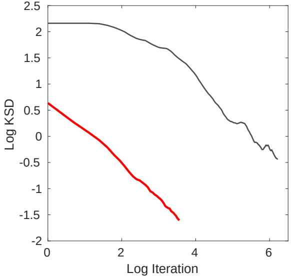
The figures we saw before were actually produced by Stein Thinning!



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



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

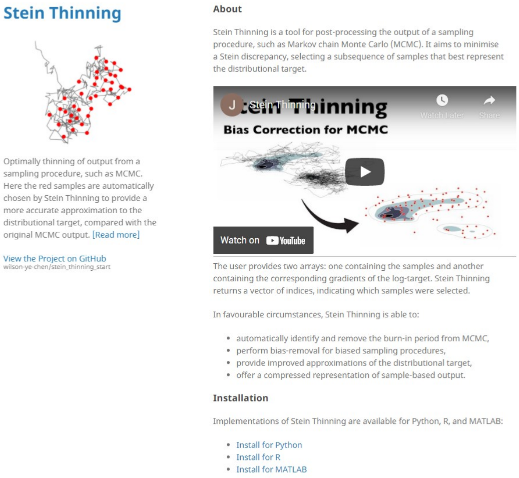


Performance
$m \mapsto \mathrm{KSD}\left(\frac{1}{m}\sum_{i \in S} \delta(\theta_i), P\right)$
(log-scales used)

The MCMC need not even be P-invariant; full details in:

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

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.

Installation

Implementations of Stein Thinning are available for Python, R, and MATLAB:

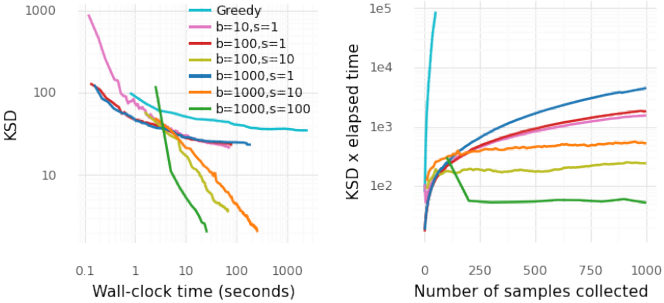
- Install for Python
- Install for R
- Install for MATLAB

▸ Link

Non-Myopic and Batch Extensions

However, 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, CJO. Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy. *AISTATS*, 2021.

Stein's Method in Computational Statistics

Stein's Method in Computational Statistics

Some other uses of Stein's method in facilitating Bayesian computation:

- **Stein Points:** Chen et al. [2018, 2019]
- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020]
- **Stein Variational Gradient Descent:** Liu and Wang [2016], ...
- **Control Variates:** CJO et al. [2017], South et al. [2022], ...
- **Variational Inference:** Fisher et al. [2021], Matsubara et al. [2022], ...

Recent advances in Stein discrepancies:

- **Diffusion-based Stein Operators:** Gorham and Mackey [2015], Gorham et al. [2019]
- **Stochastic Stein Discrepancy:** Huggins and Mackey [2018], Gorham et al. [2020]

Stein’s Method in Computational Statistics

Some other uses of Stein’s method in facilitating Bayesian computation:

- **Stein Points:** Chen et al. [2018, 2019]
- **Stein Importance Sampling:** Liu and Lee [2017], Hodgkinson et al. [2020]
- **Stein Variational Gradient Descent:** Liu and Wang [2016], ...
- **Control Variates:** CJO et al. [2017], South et al. [2022], ...
- **Variational Inference:** Fisher et al. [2021], Matsubara et al. [2022], ...

Recent advances in Stein discrepancies:

- **Diffusion-based Stein Operators:** Gorham and Mackey [2015], Gorham et al. [2019]
- **Stochastic Stein Discrepancy:** Huggins and Mackey [2018], Gorham et al. [2020]

References

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.

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.

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

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

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.

T. Matsubara, J. Knoblauch, F.-X. Briol, and CJO. Robust generalised Bayesian inference for intractable likelihoods. *JRSSB*, 2022.

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