Kernel Stein discrepancy minimization for MCMC thinning with application to cardiac electrophysiology

Marina Riabiz

Work with: Wilson Chen, Jon Cockayne, Pawel Swietach, Steve Niederer, Lester Mackey, Chris Oates

MASCOT-NUM, Paris
15^{th} November 2021

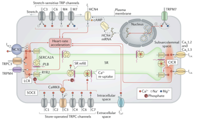
King's College London
The Alan Turing Institute
British Heart Foundation

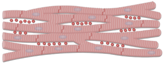
Outline

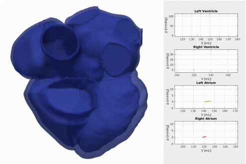
Introduction

Motivation

Computational cardiology: multi-scale and multi-physics integrated models of the hearth (*Digital Twin*)

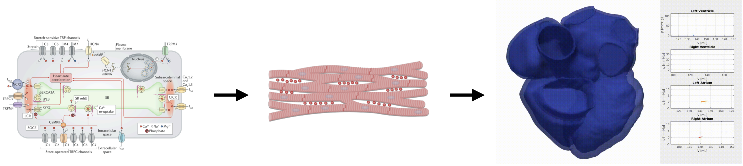






Motivation

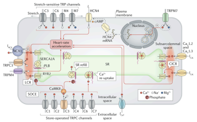
Computational cardiology: multi-scale and multi-physics integrated models of the hearth (*Digital Twin*)

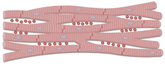


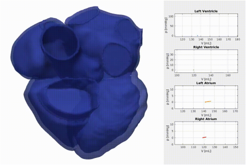
- Possible parameter inference via MCMC at cell scale, but hard to **assess the quality** of samples (*finite computing budget*)

Motivation

Computational cardiology: multi-scale and multi-physics integrated models of the hearth (*Digital Twin*)

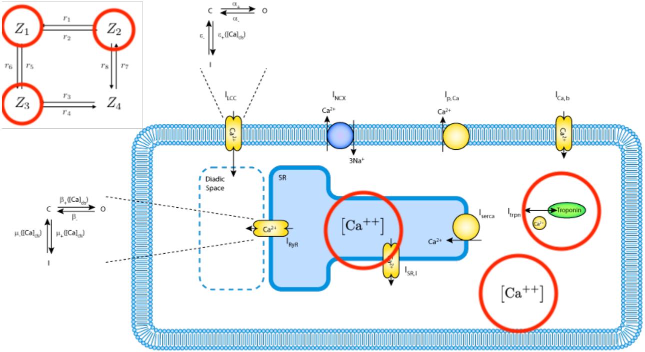






- Possible parameter inference via MCMC at cell scale, but hard to **assess the quality** of samples (*finite computing budget*)
- Computational **complexity increases** at higher scales (*compress samples to use as experimental design*)

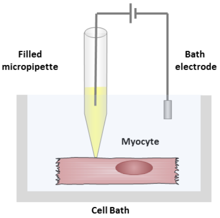
Biological Model of Calcium Transients in the Cell



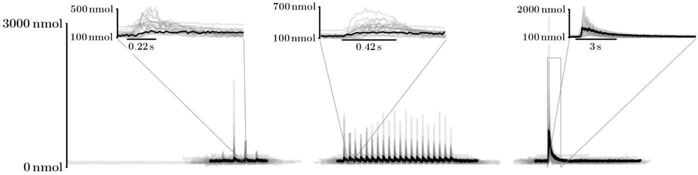
- **[Hinch 2004]** ordinary differential model, with 6 state variables

Experimental Investigation and Data

- 3-parts patch-clamp experiment on 20 ventricular myocytes



- Traces of calcium concentration in the cytoplasm



Statistical Model

- **Cell ODE model** with unknown parameters $x \in \mathbb{R}^d$, $d = 38$

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u; x)$$
$$u(0) = u_0$$

with solution $u(t; \theta) \in \mathbb{R}^6$, and u_0 assumed to be known

Statistical Model

- **Cell ODE model** with unknown parameters $x \in \mathbb{R}^d$, $d = 38$

$$\begin{aligned} \frac{\mathrm{d}u}{\mathrm{d}t} &= f(t, u; x) \\ u(0) &= u_0 \end{aligned}$$

  with solution $u(t; \theta) \in \mathbb{R}^6$, and u_0 assumed to be known

- **Gaussian measurement error** model relates the data y to the ODE (with known σ)

$$p(y|x) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - u_1(t_i; x))^2}{2\sigma^2}\right)$$

Statistical Model

- **Cell ODE model** with unknown parameters $x \in \mathbb{R}^d$, $d = 38$

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u; x)$$
$$u(0) = u_0$$

  with solution $u(t;\theta) \in \mathbb{R}^6$, and u_0 assumed to be known

- **Gaussian measurement error** model relates the data y to the ODE (with known σ)

$$p(y|x) := \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - u_1(t_i;x))^2}{2\sigma^2}\right)$$

- **Variability in cell response** is explained through different parameters x

Bayesian Inverse Model

- Model expert-derived **priors** and **system un-identifiability**

Bayesian Inverse Model

- Model expert-derived **priors** and **system un-identifiability**
- The goal is to obtain samples from the **posterior**

$$P \ : \ p(x|y) \quad = \quad \frac{p(y|x)p(x)}{p(y)}$$

where $p(x)$ is an appropriate prior density

Bayesian Inverse Model

- Model expert-derived **priors** and **system un-identifiability**
- The goal is to obtain samples from the **posterior**

$$P \ : \ p(x|y) \quad = \quad \frac{p(y|x)p(x)}{p(y)}$$

  where $p(x)$ is an appropriate prior density

- This raises technical challenges as the **normalisation constant**

$$p(y) \quad = \quad \int_{\mathcal{X}} p(y|x)p(x)\mathrm{d}x$$

  is an intractable d-dimensional integral

Bayesian Inverse Model

- Model expert-derived **priors** and **system un-identifiability**
- The goal is to obtain samples from the **posterior**

$$P \; : \; p(x|y) \quad = \quad \frac{p(y|x)p(x)}{p(y)}$$

  where $p(x)$ is an appropriate prior density

- This raises technical challenges as the **normalisation constant**

$$p(y) \quad = \quad \int_{\mathcal{X}} p(y|x)p(x)\mathrm{d}x$$

  is an intractable d-dimensional integral

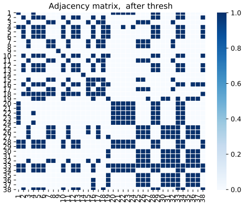
- Sampling from P via **Markov chain Monte Carlo** (MCMC) is a popular approach which requires only evaluation of the un-normalised form

$$p(x) := p(y|x)p(x)$$

  but it is not a silver bullet

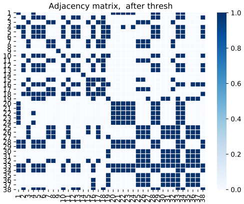
Challenges in Bayesian Inference for ODEs

- **Parameters tightly coupled together** $\implies$ posterior effectively supported on a sub-manifold of $\mathcal{X}$. See Fisher information matrix:



Challenges in Bayesian Inference for ODEs

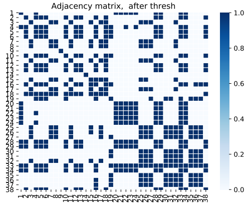
- **Parameters tightly coupled together** $\Longrightarrow$ posterior effectively supported on a sub-manifold of $\mathcal{X}$. See Fisher information matrix:



- **Gradient-based MCMC** can perform **poorly** (difficult to tune) and require computing sensitivities of the ODE at **high computing cost**

Challenges in Bayesian Inference for ODEs

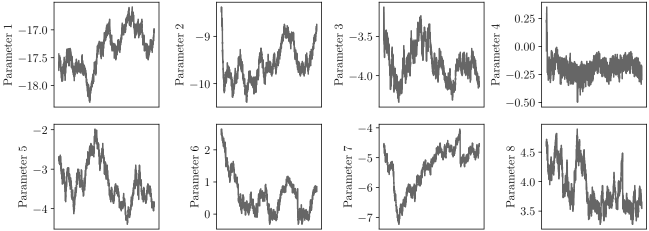
- **Parameters tightly coupled together** $\implies$ posterior effectively supported on a sub-manifold of $\mathcal{X}$. See Fisher information matrix:



- **Gradient-based MCMC** can perform **poorly** (difficult to tune) and require computing sensitivities of the ODE at **high computing cost**
- **Failure of the ODE solver** for $u(\cdot; x)$ can occur for some values of $x \in \mathcal{X}$. Unclear how to address this without introducing **bias**

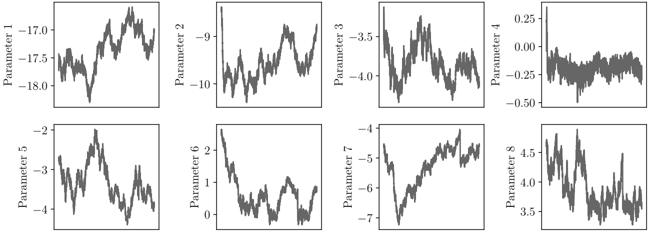
MCMC Cardiac Cell Model

Random walk MCMC run (weeks) for estimatating x ($d = 38$)

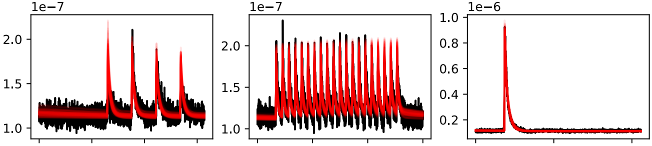


MCMC Cardiac Cell Model

Random walk MCMC run (weeks) for estimatating x ($d = 38$)



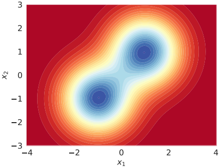
Fits



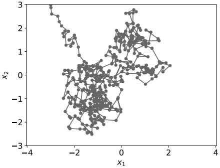
Optimal Thinning of MCMC Output

Notation and Problem

"How to remove bias from MCMC output and provide a compressed representation of the output?"



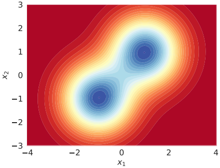
P distribution of interest, supported on $\mathbb{R}^d$



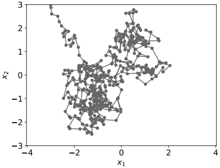
$(X_i)_{i=1}^n$ samples from a P-invariant Markov chain

Notation and Problem

"How to remove bias from MCMC output and provide a compressed representation of the output?"



P distribution of interest, supported on $\mathbb{R}^d$



$(X_i)_{i=1}^n$ samples from a P-invariant Markov chain

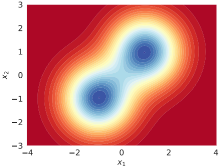
- Without MCMC postprocessing

$$P \approx \frac{1}{n}\sum_{i=1}^{n} \delta(X_i)$$

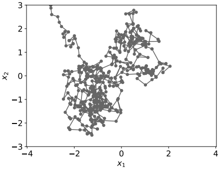
$\rightarrow$ bias if X_1 is sampled 'far from' P, and n is small
$\rightarrow$ correlated samples worsen quality of Monte Carlo estimators

Notation and Problem

"How to remove bias from MCMC output and provide a compressed representation of the output?"



P distribution of interest, supported on $\mathbb{R}^d$



$(X_i)_{i=1}^n$ samples from a P-invariant Markov chain

- Traditional postprocessing: estimate b (burn-in) and t (thinning)

$$P \approx \frac{1}{\lfloor (n-b)/t \rfloor} \sum_{i=1}^{\lfloor (n-b)/t \rfloor} \delta(X_{b+it})$$

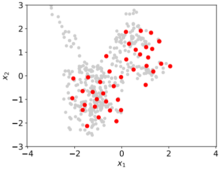
$\rightarrow$ burn-in tackles bias, but it increases variance if b is large

$\rightarrow$ thinning also tends to increase variance

Optimal MCMC Postprocessing

Desiderata: Find $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$, $m \ll n$, so that

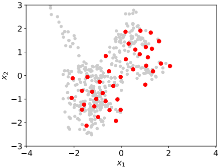
$$P \approx \frac{1}{m} \sum_{i=1}^{m} \delta(X_{\pi(i)})$$



Optimal MCMC Postprocessing

Desiderata: Find $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$, $m \ll n$, so that

$$P \approx \frac{1}{m} \sum_{i=1}^{m} \delta(X_{\pi(i)})$$



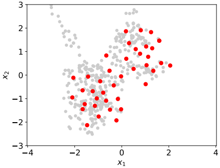
Idea: Find S by minimizing a discrepancy measure between the empirical distribution and P

$$S = \underset{\substack{S \subset \{1,\ldots,n\} \\ |S|=m}}{\arg\min} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i \in S} \delta(X_i), P \right)$$

Optimal MCMC Postprocessing

Desiderata: Find $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$, $m \ll n$, so that

$$P \approx \frac{1}{m} \sum_{i=1}^{m} \delta(X_{\pi(i)})$$



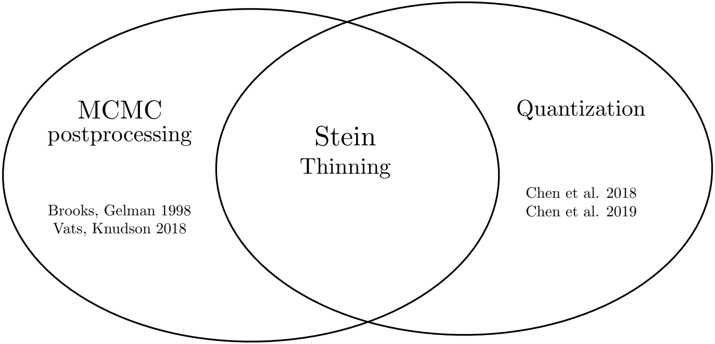
Idea: Find S by minimizing a discrepancy measure between the empirical distribution and P

$$S = \underset{\substack{S \subset \{1,\ldots,n\} \\ |S|=m}}{\arg\min} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i \in S} \delta(X_i), P \right)$$

Need to specify

1. **meaningful** and **computable** discrepancy $(*)$
2. **optimization** procedure

Stein Thinning



Step 1
Choice of Discrepancy

Worst Integration Error

- We start with an **integral probability metric**[1]

$$\mathsf{diff}\Bigg(\underbrace{\frac{1}{m}\sum_{i\in S}\delta(X_i)}_{Q}, P\Bigg) = \sup_{f\in\mathcal{F}}\left|\int f(x)\mathrm{d}Q(x) - \int f(x)\mathrm{d}P(x)\right|$$

$$=: \mathsf{IPM}_{\mathcal{F}}(Q,P)$$

based on a class of test functions $\mathcal{F}$ that is *measure-determining*:

$$\mathsf{IPM}_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow Q = P$$

[1] Müller 1997

Worst Integration Error

- We start with an **integral probability metric**[1]

$$\mathsf{diff}\Big(\underbrace{\frac{1}{m}\sum_{i\in S}\delta(X_i), P}_{Q}\Big) = \sup_{f\in\mathcal{F}}\left|\int f(x)\mathrm{d}Q(x) - \int f(x)\mathrm{d}P(x)\right|$$

$$=: \mathsf{IPM}_{\mathcal{F}}(Q, P)$$

based on a class of test functions $\mathcal{F}$ that is *measure-determining*:

$$\mathsf{IPM}_{\mathcal{F}}(P, Q) = 0 \Leftrightarrow Q = P$$

- Two problems with computing $\mathsf{IPM}_{\mathcal{F}}(P, Q)$

[1] Müller 1997

Worst Integration Error

- We start with an **integral probability metric**[1]

$$\mathsf{diff}\Bigg(\underbrace{\frac{1}{m}\sum_{i\in S}\delta(X_i)}_{Q}, P\Bigg) = \sup_{f\in\mathcal{F}}\left|\int f(x)\mathrm{d}Q(x) - \int f(x)\mathrm{d}P(x)\right|$$

$$=: \mathsf{IPM}_{\mathcal{F}}(Q,P)$$

based on a class of test functions $\mathcal{F}$ that is *measure-determining*:

$$\mathsf{IPM}_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow Q = P$$

- Two problems with computing $\mathsf{IPM}_{\mathcal{F}}(P,Q)$
- **Solution** comes from the 'freedom' in choosing $\mathcal{F}$
  - e.g. write supremum in closed form by choosing $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS) and get MMD

[1] Müller 1997

Stein's Method

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

$$X \sim P \quad \text{iff} \quad \int \mathcal{A}_P g(x) \mathrm{d}P(x) = 0 \quad \forall g \in \mathcal{G}$$

Stein's Method

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

$$X \sim P \quad \text{iff} \quad \int \mathcal{A}_P g(x) \mathrm{d}P(x) = 0 \quad \forall g \in \mathcal{G}$$

Stein Discrepancy: Given a Stein characterisation $(\mathcal{A}_P, \mathcal{G})$, the Stein discrepancy between a distribution P and an approximation Q is defined as the maximum deviation from the Stein identity

$$\mathsf{SD}(Q, P) := \sup_{g \in \mathcal{G}} \left| \int \mathcal{A}_P g(x) \mathrm{d}Q(x) \right|$$

Stein's Method

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

$$X \sim P \quad \text{iff} \quad \int \mathcal{A}_P g(x) \mathrm{d}P(x) = 0 \quad \forall g \in \mathcal{G}$$

Stein Discrepancy: Given a Stein characterisation $(\mathcal{A}_P, \mathcal{G})$, the Stein discrepancy between a distribution P and an approximation Q is defined as the maximum deviation from the Stein identity

$$\mathsf{SD}(Q, P) := \sup_{f \in \mathcal{F} = \mathcal{A}_P \mathcal{G}} \left| \int f(x) \mathrm{d}Q(x) \right|$$

Stein Operators in Hilbert Spaces

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{K}$ of functions from $\mathcal{X}$ to $\mathbb{R}$[1]

Theorem[Chwialkowski 2016] $(d = 1)$: Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_P[(\Delta k(X, X))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}_P, \mathcal{G})$, consisting of

$$\mathcal{A}_P g = \frac{\nabla(gp)}{p}, \qquad \mathcal{G} = \{g \in \mathcal{K} : \|g\|_{\mathcal{K}} \leq 1\}.$$

[1] i.e $\forall x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{K}$ and $f(x) = \left\langle f, k(x, \cdot) \right\rangle_{\mathcal{K}}$ whenever $f \in \mathcal{K}$

Stein Operators in Hilbert Spaces

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the reproducing kernel of a RKHS $\mathcal{K}$ of functions from $\mathcal{X}$ to $\mathbb{R}$[1]

Theorem[Chwialkowski 2016] $(d = 1)$: Suppose that k is bounded, symmetric, cc-universal and satisfies $\mathbb{E}_P[(\Delta k(X, X))^2] < \infty$. Then P has Stein characterisation $(\mathcal{A}_P, \mathcal{G})$, consisting of

$$\mathcal{A}_P g = \frac{\nabla(gp)}{p}, \qquad \mathcal{G} = \{g \in \mathcal{K} : \|g\|_\mathcal{K} \le 1\}.$$

Theorem[Oates 2017] $(d = 1)$: The functions $\mathcal{A}_P g$ just defined are precisely the elements of the unit ball in the RKHS $\mathcal{K}_P := \mathcal{A}_P \mathcal{K}$ with kernel

$$\begin{aligned} k_P(x, y) &= \nabla_x \nabla_y k(x, y) + \frac{\nabla_x p(x)}{p(x)} \nabla_y k(x, y) \\ &\quad + \frac{\nabla_y p(y)}{p(y)} \nabla_x k(x, y) + \frac{\nabla_x p(x)}{p(x)} \frac{\nabla_y p(y)}{p(y)} k(x, y) \end{aligned}$$

In particular, under regularity conditions, $\int h \mathrm{d}P = 0, \forall h \in \mathcal{K}_P$

[1] i.e $\forall x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{K}$ and $f(x) = \langle f, k(x, \cdot) \rangle_\mathcal{K}$ whenever $f \in \mathcal{K}$

Kernel Stein Discrepancy[1]

$$\mathrm{IPM}^2_{\mathcal{F}}(Q,P) = \sup_{f\in\mathcal{F}} \left|\int f(x)\mathrm{d}Q(x) - \int f(x)\mathrm{d}P(x)\right|^2$$

$\mathcal{F} = \mathcal{F}_k$

$$\mathrm{MMD}^2(Q,P) = \iint k(x,y)\mathrm{d}Q(x)\mathrm{d}Q(y) - 2\iint k(x,y)\mathrm{d}Q(x)\mathrm{d}P(y) + \iint k(x,y)\mathrm{d}P(x)\mathrm{d}P(y)$$

$\mathcal{F} = \mathcal{A}_P\mathcal{G}$

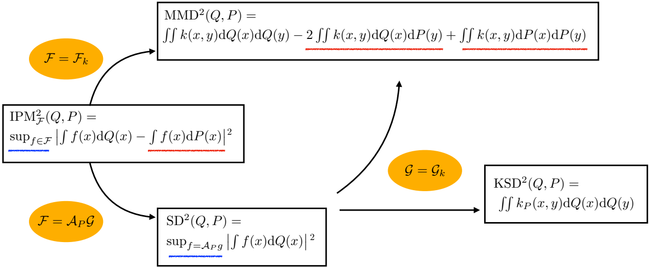
$$\mathrm{SD}^2(Q,P) = \sup_{f=\mathcal{A}_P g} \left|\int f(x)\mathrm{d}Q(x)\right|^2$$

$\mathcal{G} = \mathcal{G}_k$

$$\mathrm{KSD}^2(Q,P) = \iint k_P(x,y)\mathrm{d}Q(x)\mathrm{d}Q(y)$$

[1] Chwialkowski et al., 2016; Liu et al., 2016, Gorham et al., 2017

Kernel Stein Discrepancy[1]



When Q is an empirical measure

$$\mathsf{KSD}\left(\frac{1}{m}\sum_{i\in S}\delta(X_i), P\right) = \sqrt{\frac{1}{m^2}\sum_{i,j\in S} k_P(X_i, X_j)}$$

where k_P depends on evaluations of a base kernel k and $\nabla \log p$

[1] Chwialkowski et al., 2016; Liu et al., 2016, Gorham et al., 2017

KSD Convergence Control

Conditions on P (*distantly dissipative*) ensure that the KSD is **convergence determining**:

$$\text{KSD} \ \to 0 \text{ implies } \frac{1}{m}\sum_{i\in S}\delta(X_i) \Rightarrow P$$

when the base kernel k is the inverse-multiquadric kernel[1] with hyper-parameter Γ

$$k(x,y) := (1 + \|\Gamma^{-1/2}(x-y)\|^2)^{-1/2}$$

[1] Gorham et al., 2017

KSD Convergence Control

Conditions on P (*distantly dissipative*) ensure that the KSD is **convergence determining**:

$$\text{KSD } \to 0 \text{ implies } \frac{1}{m}\sum_{i\in S}\delta(X_i) \Rightarrow P$$

when the base kernel k is the inverse-multiquadric kernel[1] with hyper-parameter Γ

$$k(x,y) := (1 + \|\Gamma^{-1/2}(x-y)\|^2)^{-1/2}$$

- It makes sense to minimize the KSD
- The choice of Γ can affect the outcome in practice

[1] Gorham et al., 2017

Step 2
Optimization Procedure

Step 2: Greedy Minimization of KSD

Given the MCMC output $(X_i)_{i=1}^n$ and a kernel k_P, the point set $S = \{\pi_{(1)} \dots, \pi_{(m)}\} \subset \{1, \dots, n\}^m$ is obtained as:

$$\pi_{(j)} \quad \in \quad \underset{i=1,\dots,n}{\operatorname{argmin}} \, \mathsf{KSD}\left(\frac{1}{j} \left[\delta(X_i) + \sum_{j'=1}^{j-1} \delta(X_{\pi(j')}) \right], P \right) \qquad j = 1, \dots, m$$

Step 2: Greedy Minimization of KSD

Given the MCMC output $(X_i)_{i=1}^n$ and a kernel k_P, the point set $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$ is obtained as:

$$\pi_{(j)} \quad \in \quad \operatorname*{argmin}_{i=1,\ldots,n} \mathsf{KSD}\left(\frac{1}{j} \left[\delta(X_i) + \sum_{j'=1}^{j-1} \delta(X_{\pi(j')}) \right], P \right) \qquad j = 1, \ldots, m$$

- The cost of the algorithm is $O(nm^2)$, hence comparable to that of producing the MCMC output $O(n)$ when m is fixed
- The same x_i can be selected more than once (necessary if $m > n$)

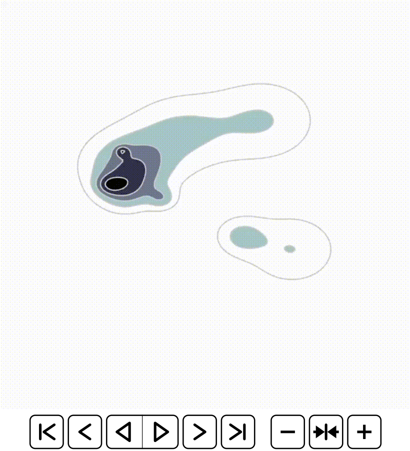
Step 2: Greedy Minimization of KSD

Given the MCMC output $(X_i)_{i=1}^n$ and a kernel k_P, the point set $S = \{\pi_{(1)} \dots, \pi_{(m)}\} \subset \{1, \dots, n\}^m$ is obtained as:

$$\pi_{(j)} \quad \in \quad \operatorname*{argmin}_{i=1,\dots,n} \mathsf{KSD}\left(\frac{1}{j}\left[\delta(X_i) + \sum_{j'=1}^{j-1} \delta(X_{\pi(j')})\right], P\right) \qquad j = 1, \dots, m$$

- The cost of the algorithm is $O(nm^2)$, hence comparable to that of producing the MCMC output $O(n)$ when m is fixed
- The same x_i can be selected more than once (necessary if $m > n$)
- Myopic + no mini-batching

Stein Thinning Example[1]



[1] https://en.wikipedia.org/wiki/Stein_discrepancy
https://www.youtube.com/watch?v=WwmTeLrNmOQ&t=6s

Results

Theoretical Results

Theoretical Results

Guarantee consistency of the empirical distribution obtained, considering

- type of KSD optimization procedure (**greedy**)

Theoretical Results

Guarantee consistency of the empirical distribution obtained, considering

- type of KSD optimization procedure (**greedy**)
- **randomness** of the MCMC output

Theoretical Results

Guarantee consistency of the empirical distribution obtained, considering

- type of KSD optimization procedure (**greedy**)
- **randomness** of the MCMC output
- possible **bias** in the Markov chain

Result 1: Convergence for fixed $(x_i)_{i=1}^n$, as $m \to \infty$

$$\mathrm{KSD}\left(\frac{1}{m}\sum_{j=1}^{m}\delta(x_{\pi(j)}), P\right)^2 \le \mathrm{KSD}\left(\sum_{i=1}^{n} w_i^* \delta(x_i), P\right)^2 + \\ + \left(\frac{1+\log(m)}{m}\right) \max_{i=1,\ldots,n} k_P(x_i, x_i)$$

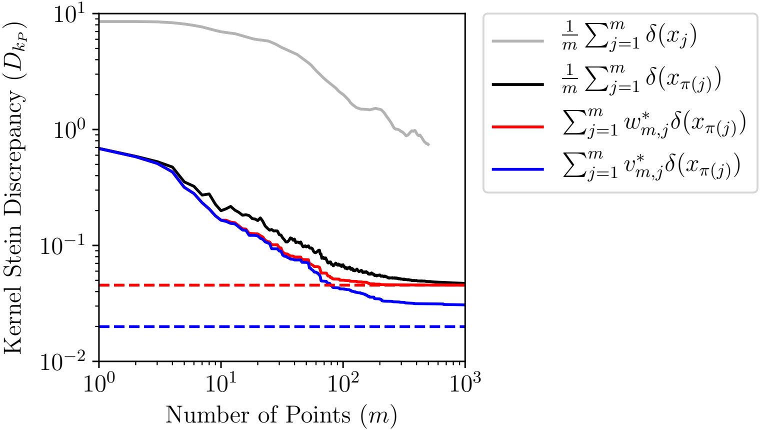
where the weights $w^* = (w_1^*, \ldots, w_n^*)$ satisfy

$$w^* \in \mathop{\arg\min}_{\substack{1_n^\top w = 1 \\ w \ge 0}} \mathrm{KSD}\left(\sum_{i=1}^{n} w_i \delta(x_i), P\right)$$

and $\sum_{i=1}^n w_i^* \delta(x_i)$ is the optimal weighted empirical distribution based on $(x_i)_{i=1}^n$, with cost $O(n^3)$[1]

[1]Liu, Lee 2017, Hodgkinson et al. 2020

Illustration of Result 1



(where v^* are optimal weights without positivity constraint)

V-Uniform Ergodicity

For a function $V : \mathcal{X} \to [1, \infty)$, a function $f : \mathcal{X} \to \mathbb{R}$ and a measure Q on $\mathcal{X}$ we denote

$$\|f\|_V := \sup_{x \in \mathcal{X}} |f(x)|/V(x),$$

$$\|Q\|_V := \sup_{\|f\|_V \leq 1} \left| \int f \mathrm{d}Q \right|$$

A Markov chain with transition kernel P^n is said to be V-*uniformly ergodic*[1] if there exist constants $R \in [0, \infty)$, $\rho \in [0, 1)$, such that

$$\|P^n(x, \cdot) - P\|_V \leq RV(x)\rho^n$$

for all $n \in \mathbb{N}$ and all initial states $x \in \mathcal{X}$.

[1] Meyn and Tweedie, 2012)

Result 2: L^2 Convergence

Let $(X_i)_{i \in \mathbb{N}}$ be a P-invariant, time-homogeneous, reversible Markov chain, generated using a V-uniformly ergodic transition kernel, such that $V(x) \geq \sqrt{k_P(x,x)}$[1] for all $x \in \mathcal{X}$.

[1]The function $x \mapsto \sqrt{k_P(x,x)}$ can be understood in terms of $\|\nabla \log p(x)\|$

Result 2: L^2 Convergence

Let $(X_i)_{i \in \mathbb{N}}$ be a P-invariant, time-homogeneous, reversible Markov chain, generated using a V-uniformly ergodic transition kernel, such that $V(x) \geq \sqrt{k_P(x,x)}$[1] for all $x \in \mathcal{X}$. Under regularity conditions[2]

$$\mathbb{E}\left[\mathsf{KSD}\left(\frac{1}{m}\sum_{j=1}^{m}\delta(X_{\pi(j)}), P\right)^2\right] \leq \frac{\log(b)}{\gamma n} + \frac{CM}{n} + \left(\frac{1+\log(m)}{m}\right)\frac{\log(nb)}{\gamma}$$

- Mean-square convergence to 0 of the KSD as $m, n \to \infty$, $m \propto n$
- Non-asymptotic (non-tight) bound on the expected KSD squared

[1] The function $x \mapsto \sqrt{k_P(x,x)}$ can be understood in terms of $\|\nabla \log p(x)\|$

[2] The chain has explored regions of high probability under P

Result 3: Consistency (with Biased MCMC)

Let Q a be probability distribution on $\mathcal{X}$ with $P \ll Q$.
Let $(X_i)_{i \in \mathbb{N}}$ be a Q-invariant, time-homogeneous, reversible Markov chain generated using a V-uniformly ergodic transition kernel, such that $V(x) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\sqrt{k_P(x,x)}$.

Result 3: Consistency (with Biased MCMC)

Let Q a be probability distribution on $\mathcal{X}$ with $P \ll Q$.
Let $(X_i)_{i \in \mathbb{N}}$ be a Q-invariant, time-homogeneous, reversible Markov chain generated using a V-uniformly ergodic transition kernel, such that $V(x) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\sqrt{k_P(x,x)}$. Under regularity conditions[1]

$$\mathsf{KSD}\left(\frac{1}{m}\sum_{j=1}^{m}\delta(X_{\pi(j)}), P\right) \to 0, \quad a.s. \text{ as } m, n \to \infty$$

and

$$\frac{1}{m}\sum_{j=1}^{m}\delta(X_{\pi(j)}) \Rightarrow P, \quad a.s. \text{ as } m, n \to \infty$$

[1] Q is not too dissimilar from P

Empirical Results

Inverse posterior inference for systems of ODEs, MCMC output obtained through random walk Metropolis-Hastings

Empirical Results

Inverse posterior inference for systems of ODEs, MCMC output obtained through random walk Metropolis-Hastings

Comparison of empirical measures obtained via

- Traditional thinning [1]
- Support points [2]
- Stein thinning based on three settings for Γ in the base kernel k:
  - Median (`med`)
  - Scaled median (`sclmed`)
  - Sample covariance (`smpcov`)

[1] Brooks, Gelman 1998; Vats, Knudson 2018

[2] Mak, Joseph 2018

Empirical Results

Inverse posterior inference for systems of ODEs, MCMC output obtained through random walk Metropolis-Hastings

Comparison of empirical measures obtained via

- Traditional thinning [1]
- Support points [2]
- Stein thinning based on three settings for Γ in the base kernel k:
  - Median (`med`)
  - Scaled median (`sclmed`)
  - Sample covariance (`smpcov`)

Two performance measures:

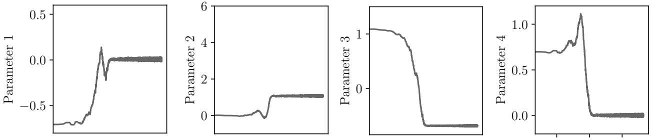
1. Energy distance[3]
2. KSD based on one setting for Γ

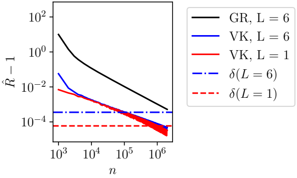
[1] Brooks, Gelman 1998; Vats, Knudson 2018

[2] Mak, Joseph 2018

[3] Criterion minimized by Support points

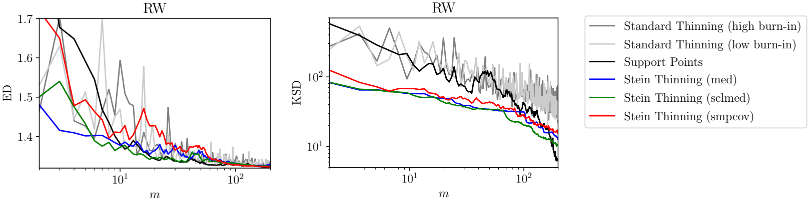
Goodwin Oscillator ($d = 4$) - Convergence Diagnostics





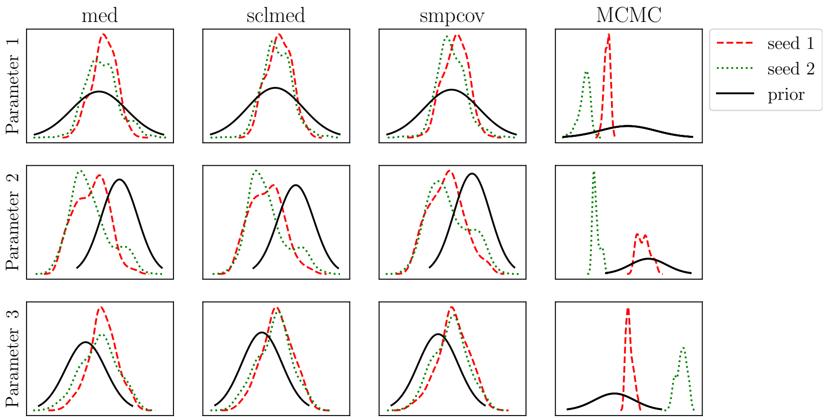
- Univariate and multivariate convergence diagnostics ($L \geq 1$ chains)
- Thresholding $\hat{R}$ leads to identify $\hat{b}$
- Bias-variance trade-off in fixed n scenario

Goodwin Oscillator - Performance metrics



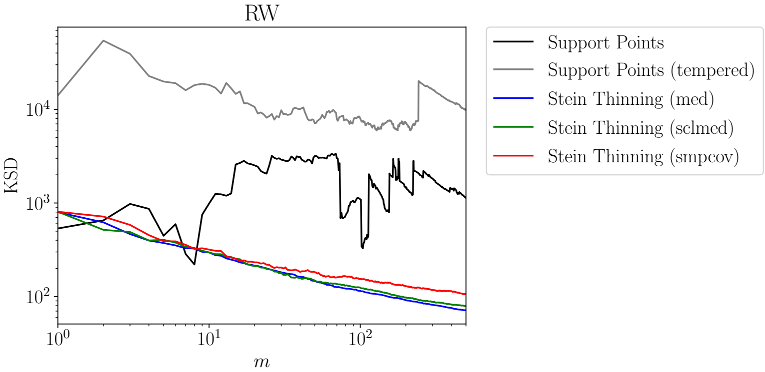
- Energy distance:
  - Not sensitive to details, needs high quality MCMC output
  - Does not provide convergence control

Hinch Cell Model - Marginals



- MCMC targeting the original posteriror is stuck in local modes
- Stein-thinned MCMC targeting tempered posteriors is consistent across seeds, and choice of preconditioner Γ

Hinch Model - KSD



- Tempered MCMC output: ST achieves lower KSD values than SP, because it corrects for bias caused by tempering
- Standard MCMC output: ST achieves lower KSD values than SP, that is negatively affected by the non-convergence of MCMC

Conclusions

Conclusions

Advantages

- automatically identify and remove the burn-in from MCMC output
- offer a compressed representation of sample-based output
- perform bias-removal for biased sampling procedures

Conclusions

Advantages

- automatically identify and remove the burn-in from MCMC output
- offer a compressed representation of sample-based output
- perform bias-removal for biased sampling procedures

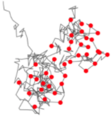
Caveats

- requires MCMC to have explored regions of high probability under P
- requires $\nabla \log p$, which might be expensive to compute (but could be computed in parallel as post-processing step)
- subject to pathologies if P has distant probabilities regions or P is high-dimensional and multi-modal
- not invariant to re-parametrizations

Stein Thinning[1]

Project webpage under development (`http://stein-thinning.org/`)

Stein Thinning



Optimally improves MCMC output via intelligent thinning and burn-in removal. The red dots are automatically chosen by Stein Thinning from the output of a slow-mixing MCMC sampler targeting a Gaussian mixture distribution [Read more].

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

This project is maintained by wilson-ye-chen

Hosted on GitHub Pages — Theme by orderedlist

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, to select a subset of the 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 representative samples were selected. In favourable circumstances, Stein Thinning is able to:

- automatically identify and remove the burn-in period from MCMC output,
- 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 currently available for Python and MATLAB:

- Install for Python
- Install for MATLAB
- Install for R (Coming soon!)

Get Started

In Python, MATLAB, or R, it takes a single function call to start Stein Thinning:

[1] Riabiz et al., Optimal Thinning of MCMC Output, arXiv:2005.03952, 2020

Thank you for your attention!

References

- R. Hinch, JL Greenstein, AJ Tanskanen, L Xu, and RL Winslow. A simplified local con-trol model of calcium-induced calcium release in cardiac ventricular myocytes. Biophysical journal, 87(6):3723-3736, 2004.
- K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of t. In ICML, 2016.
- Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-t tests. In International Conference on Machine Learning, pages 276-284, 2016.
- J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In Proceedings of the International Conference on Machine Learning, pages 1292-1301, 2017.
- Q. Liu and J. D. Lee. Black-box importance sampling. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.
- L. Hodgkinson, R. Salomone, and F. Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. arXiv:2001.09266, 2020.

References

- S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4):434-455, 1998.
- A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. Statistical science, 7(4):457-472, 1992.
- D. Vats and C. Knudson. Revisiting the Gelman-Rubin diagnostic. arXiv:1812.09384, 2018.
- S. Mak and V. R. Joseph. Support points. The Annals of Statistics, 46(6A):2562?2592, 2018.
- G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. InterStat, 5(16.10):1249?1272, 2004.