Kernel Stein discrepancy minimization for MCMC thinning in cardiac electrophysiology

Marina Riabiz

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

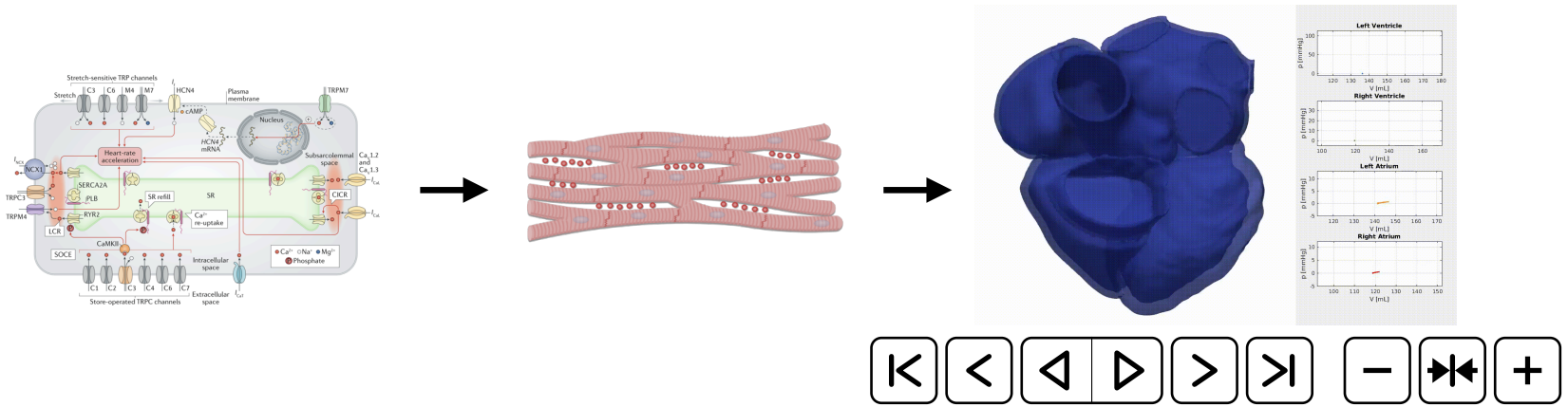
Inference for Expensive Systems in Mathematical Biology
23^{rd} May 2022

Outline

Introduction

Motivation

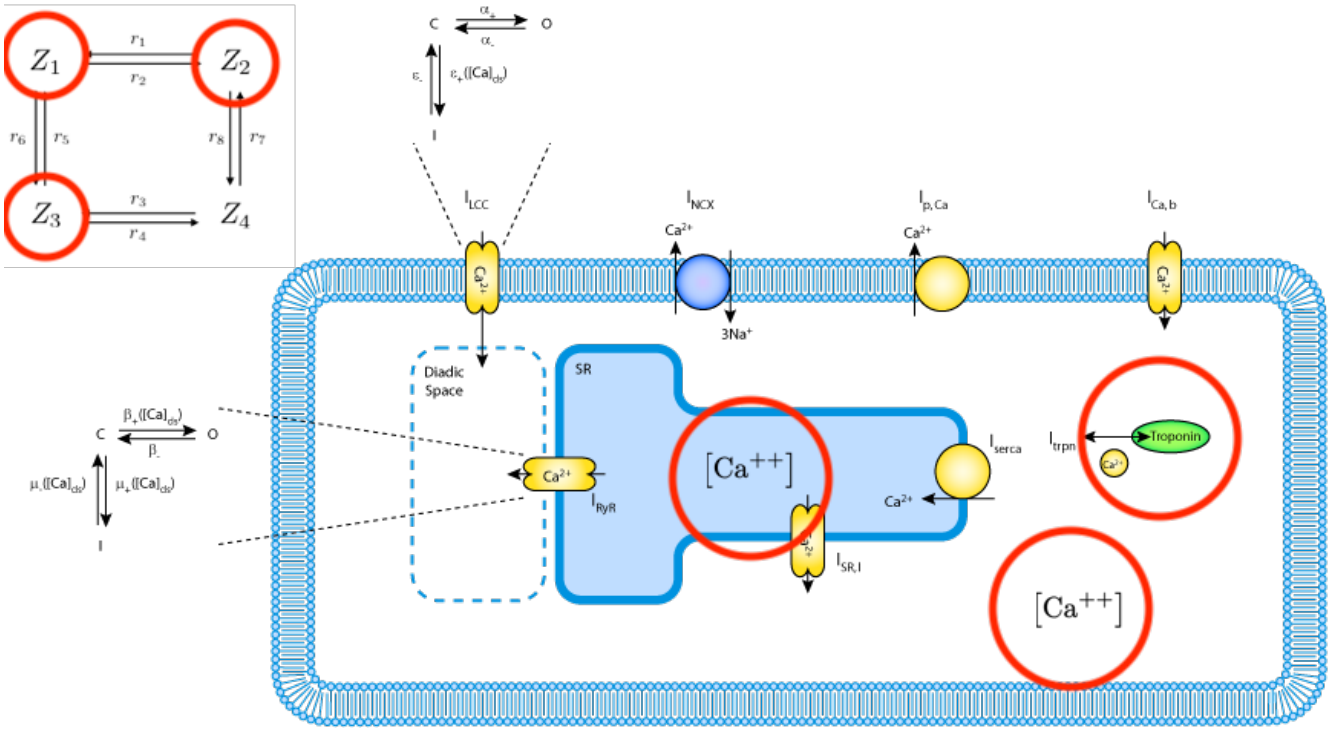
Computational cardiology: multi-scale and multi-physics integrated models of the hearth (*Digital Twin*[1])



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

[1] Strocchi et al. 2020; Niederer et al. 2021

Biological Model of Calcium Transients in the Cell

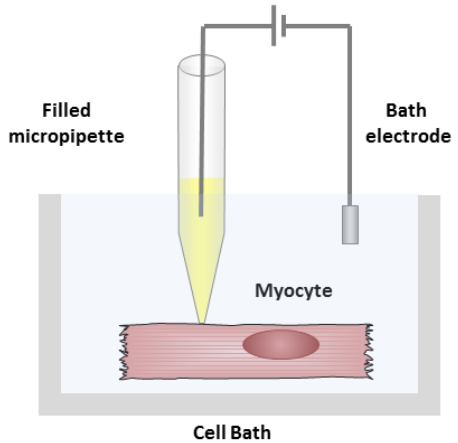


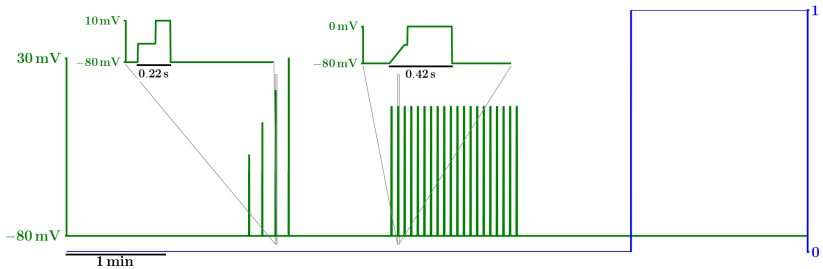
- Ordinary differential model, with 6 state variables[1]

[1] Hinch et al. 2004

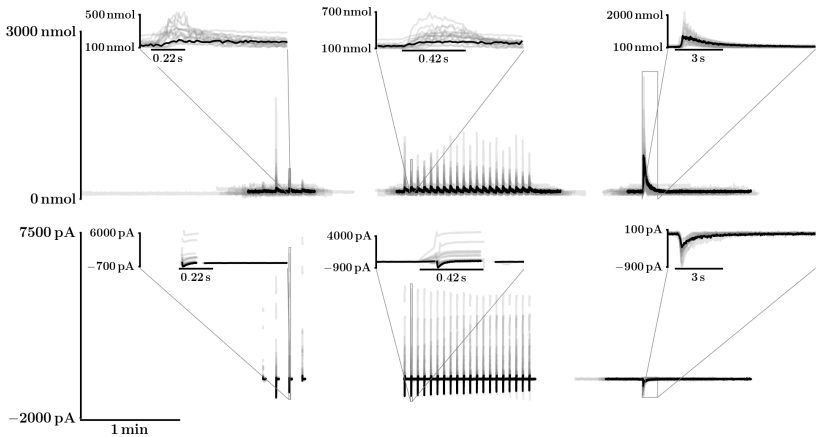
Experimental Investigation and Data

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





- Free $[Ca]^{2+}$ and transmembrane current time-courses



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**
- The goal is to obtain samples from the **posterior**

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

with $p(x)$ an appropriate prior density, and $p(y)$ an intractable d-dimensional integral

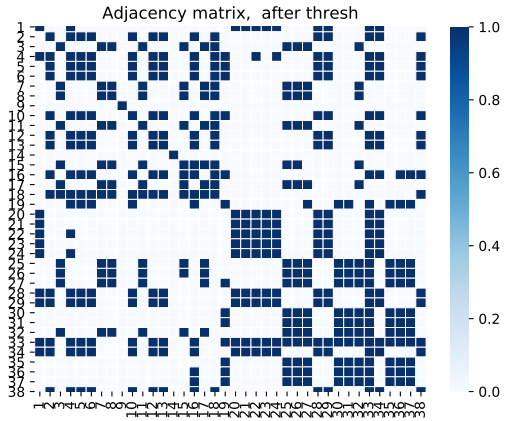
- Sampling from P via **Markov chain Monte Carlo** (MCMC) is a popular approach which requires only evaluation of the un-normalised form

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

but it is not a silver bullet

Challenges in Bayesian Inference for ODEs

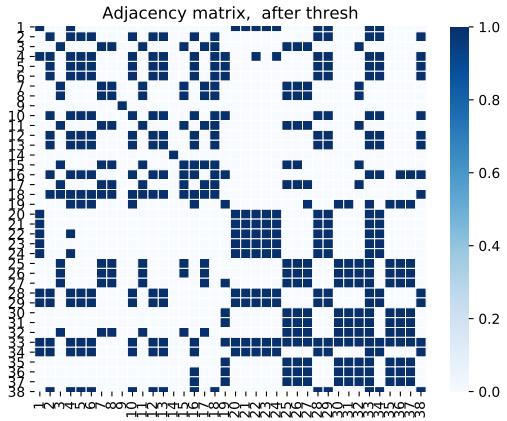
- **Parameters tightly coupled together** $\implies$ posterior effectively supported on a sub-manifold of $\mathcal{X}$. See Fisher information matrix:



- **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

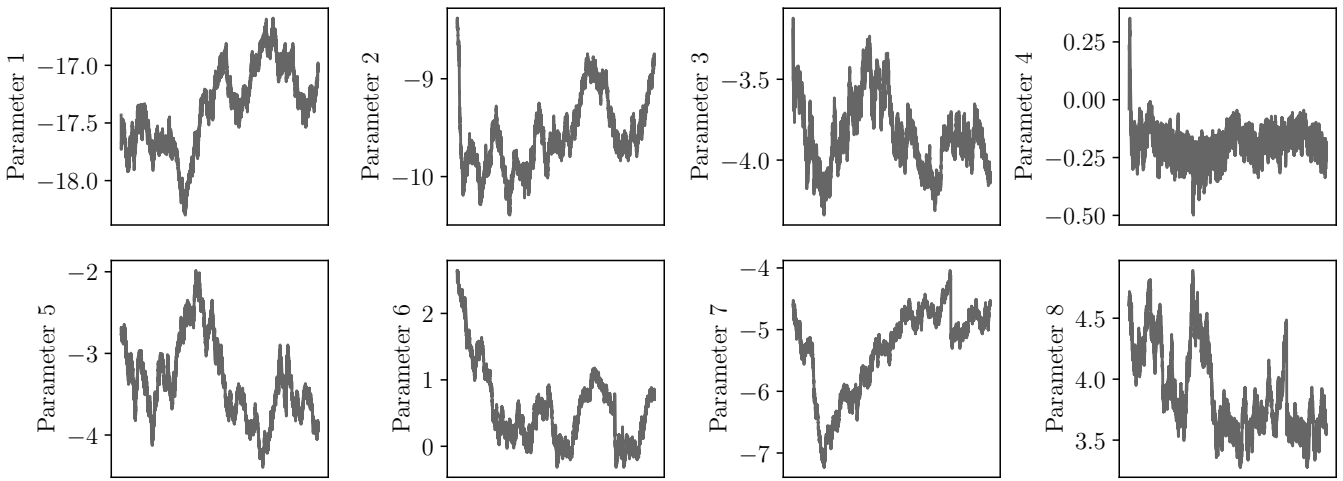
- **Parameters tightly coupled together** $\implies$ posterior effectively supported on a sub-manifold of $\mathcal{X}$. See Fisher information matrix:



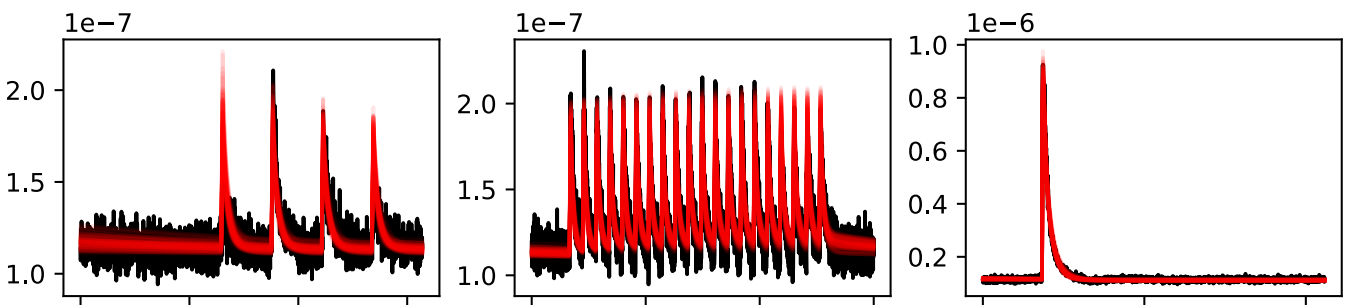
- **Gradient-based MCMC** can perform **poorly** (difficult to tune) and require computing sensitivities of the ODE at **high computing cost**
- **Failure of the ODE solver** for $u(\cdot; x)$ can occur for some values of $x \in \mathcal{X}$. Unclear how to address this without introducing **bias**

MCMC Cardiac Cell Model

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



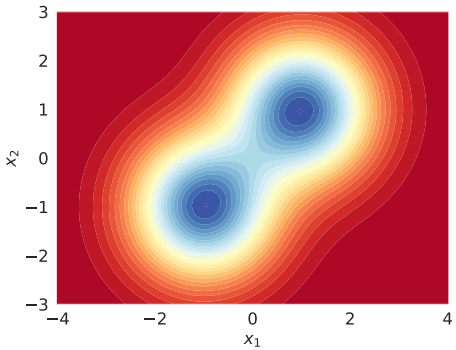
Fits



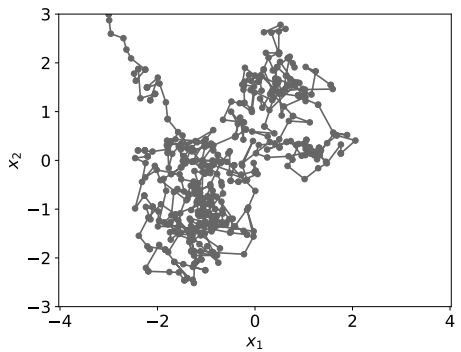
Optimal Thinning of MCMC Output

Notation and Problem

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



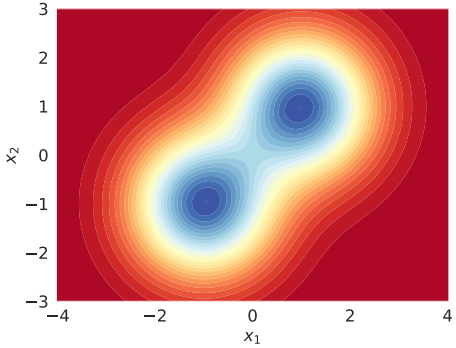
P distribution of interest, supported on $\mathbb{R}^d$



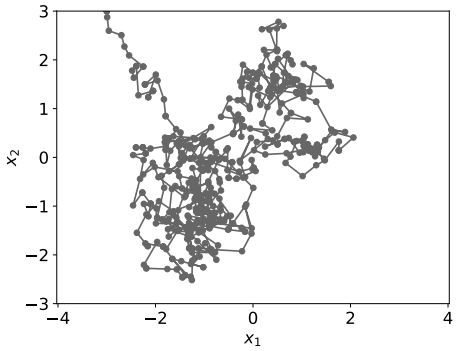
$(X_i)_{i=1}^n$ samples from a P-invariant Markov chain

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

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

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

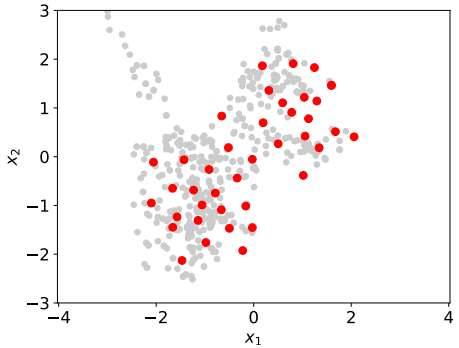
$\rightarrow$ burn-in tackles bias, but it increases variance if b is large

$\rightarrow$ 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

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

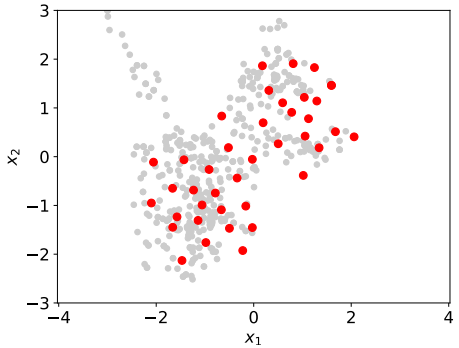


[1] Riabiz et al. 2020

Optimal MCMC Postprocessing

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

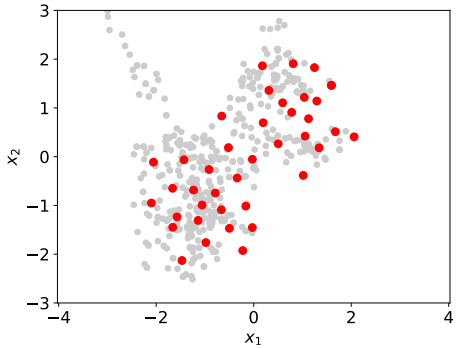
$$S = \underset{\substack{S \subset \{1,\dots,n\} \\ |S|=m}}{\arg\min} \underbrace{\text{diff}}_{(*)} \left(\frac{1}{m} \sum_{i \in S} \delta(X_i), P \right)$$

[1] Riabiz et al. 2020

Optimal MCMC Postprocessing

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

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

Stein Thinning[1]**:** Need to specify discrepancy $(*)$ and optimization procedure

[1] Riabiz et al. 2020

Step 1: Choice of Discrepancy

We start with an **integral probability metric** based on a class of test functions $\mathcal{F}$ that is *measure-determining*[1]

$$\text{diff}\left(\frac{1}{m}\sum_{i\in S}\delta(X_i), P\right) \quad := \quad \sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i\in S} f(X_i) - \int f(x)\mathrm{d}P(x)\right|$$

[1] $\text{diff} = 0 \Leftrightarrow \frac{1}{m}\sum_{i\in S}\delta(X_i) = P$

Step 1: Choice of Discrepancy

We start with an **integral probability metric** based on a class of test functions $\mathcal{F}$ that is *measure-determining*[1]

$$\text{diff}\left(\frac{1}{m}\sum_{i\in S}\delta(X_i), P\right) \quad := \quad \sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i\in S} f(X_i) - \int f(x)\mathrm{d}P(x)\right|$$

Two problems with computing IPMs

[1] $\text{diff} = 0 \Leftrightarrow \frac{1}{m}\sum_{i\in S}\delta(X_i) = P$

Step 1: Choice of Discrepancy

We start with an **integral probability metric** based on a class of test functions $\mathcal{F}$ that is *measure-determining*[1]

$$\text{diff}\left(\frac{1}{m}\sum_{i\in S}\delta(X_i), P\right) \quad := \quad \sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i\in S} f(X_i) - \int f(x)\mathrm{d}P(x)\right|$$

Two problems with computing IPMs

Solution comes from the 'freedom' in choosing $\mathcal{F}$. Jointly:

- Write supremum in closed form by choosing $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS) (*MMD*)
- Based on Stein's method, choose $\mathcal{F} = \mathcal{F}_P$ such that $\int f(x)\mathrm{d}P(x) = 0$ (*Stein discrepancy*)

[1] $\text{diff} = 0 \Leftrightarrow \frac{1}{m}\sum_{i\in S}\delta(X_i) = P$

Step 1: Kernel Stein Discrepancy

The result is the **kernel Stein discrepancy:**[1]

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

where k_P is the kernel of a RKHS, that depends on

- evaluations of $\nabla \log p$ (*no normalizing constant*)
- a base kernel k

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

Step 1: Kernel Stein Discrepancy

The result is the **kernel Stein discrepancy:**[1]

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

where k_P is the kernel of a RKHS, that depends on

- evaluations of $\nabla \log p$ (*no normalizing constant*)
- a base kernel k

Conditions on k and P ensure that the KSD is **convergence determining** (KSD $\to 0$ implies $\frac{1}{m}\sum_{i\in S}\delta(X_i) \Rightarrow P$) when k is the inverse-multiquadric kernel[2] with hyper-parameter Γ

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

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

[2] Gorham et al., 2017

KSD Convergence Control

When Q is an **empirical measure**

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

where k_P is the kernel of a RKHS, that depends on

- evaluations of $\nabla \log p$ (*no normalizing constant*)
- a base kernel k

KSD Convergence Control

When Q is an **empirical measure**

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

where k_P is the kernel of a RKHS, that depends on

- evaluations of $\nabla \log p$ (*no normalizing constant*)
- a base kernel k

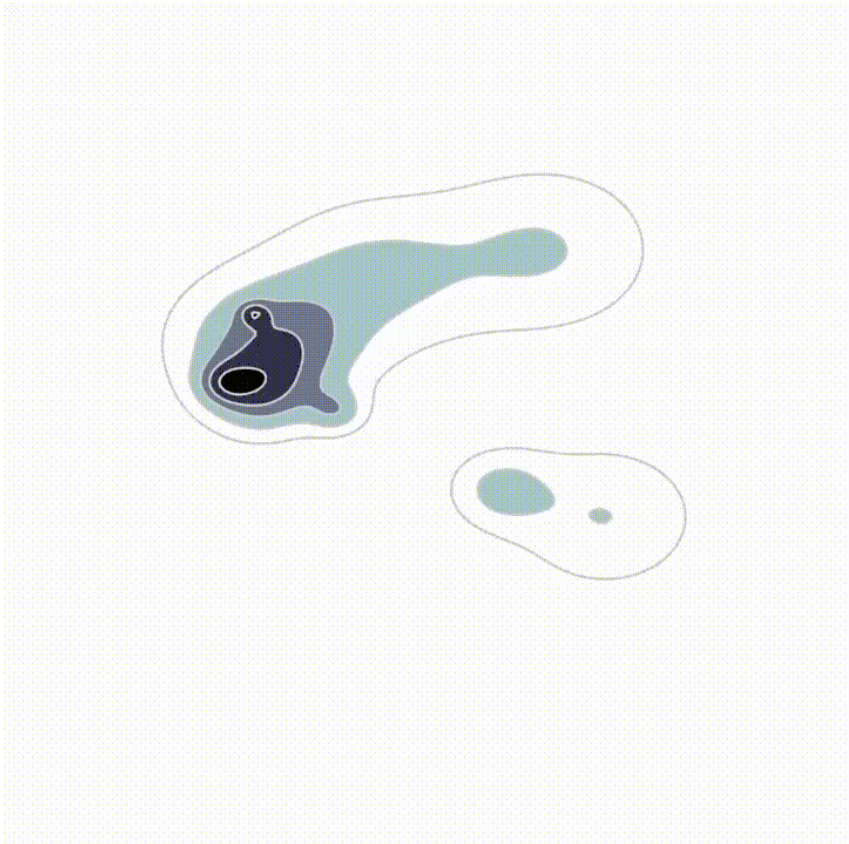
Conditions on k and P ensure that the KSD is **convergence determining** (KSD $\to 0$ implies $\frac{1}{m}\sum_{i\in S}\delta(X_i) \Rightarrow P$) when k is the inverse-multiquadric kernel[1] with hyper-parameter Γ

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

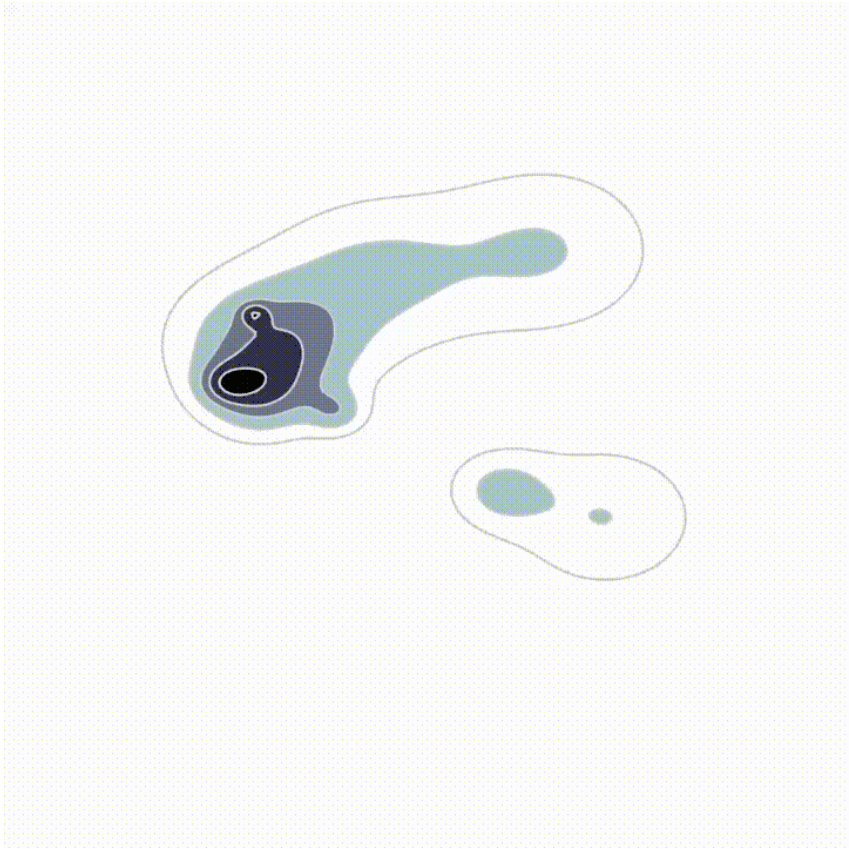
[1] Gorham et al., 2017

Step 2: KSD Optimization

The point set $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$ is obtained by greedy minimization of the KSD



Stein Thinning Example[1]



[1] https://www.youtube.com/watch?v=WwmTeLrNmOQ&t=6s

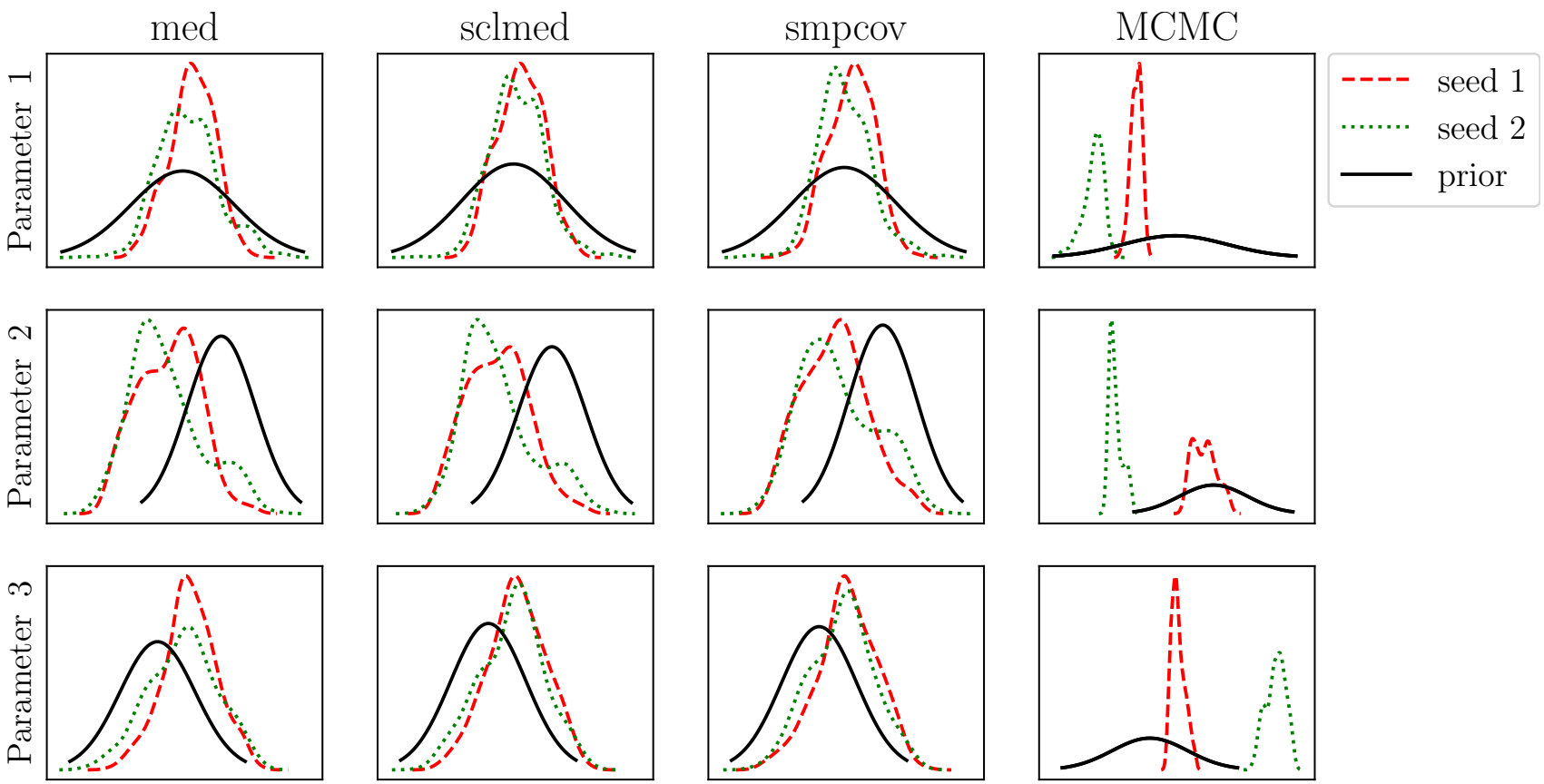
Results

Theoretical Results

Guarantee consistency of the empirical distribution obtained, considering

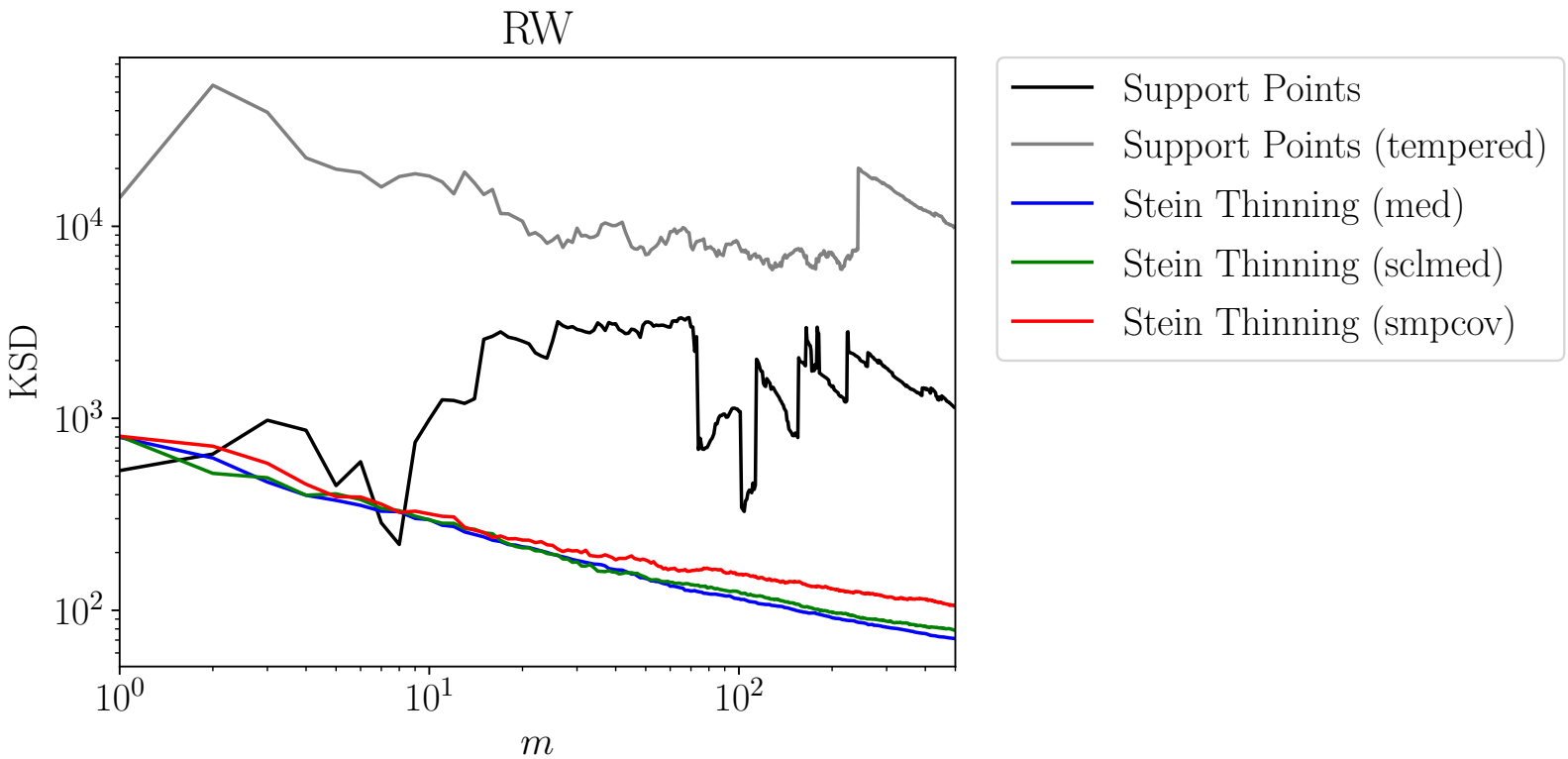
- type of KSD optimization procedure (**greedy**)
- **randomness** of the MCMC output
- possible **bias** in the Markov chain (e.g. **tempered** target)

Hinch Cell Model - Marginals

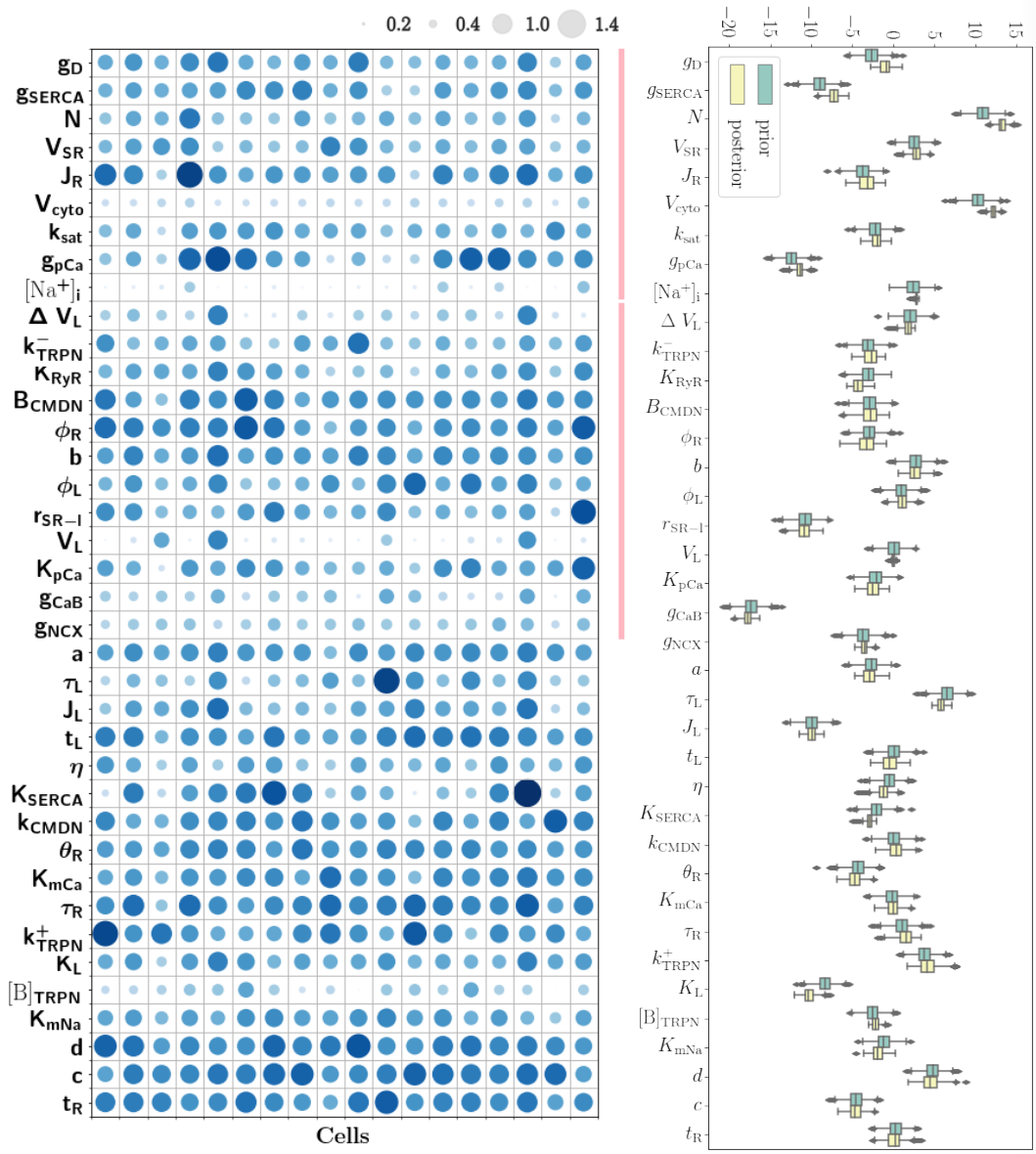


- MCMC targeting the original posteriror is stuck in local modes
- Stein-thinned MCMC targeting tempered posteriors is consistent across seeds, and choice of preconditioner Γ

Hinch Model - KSD



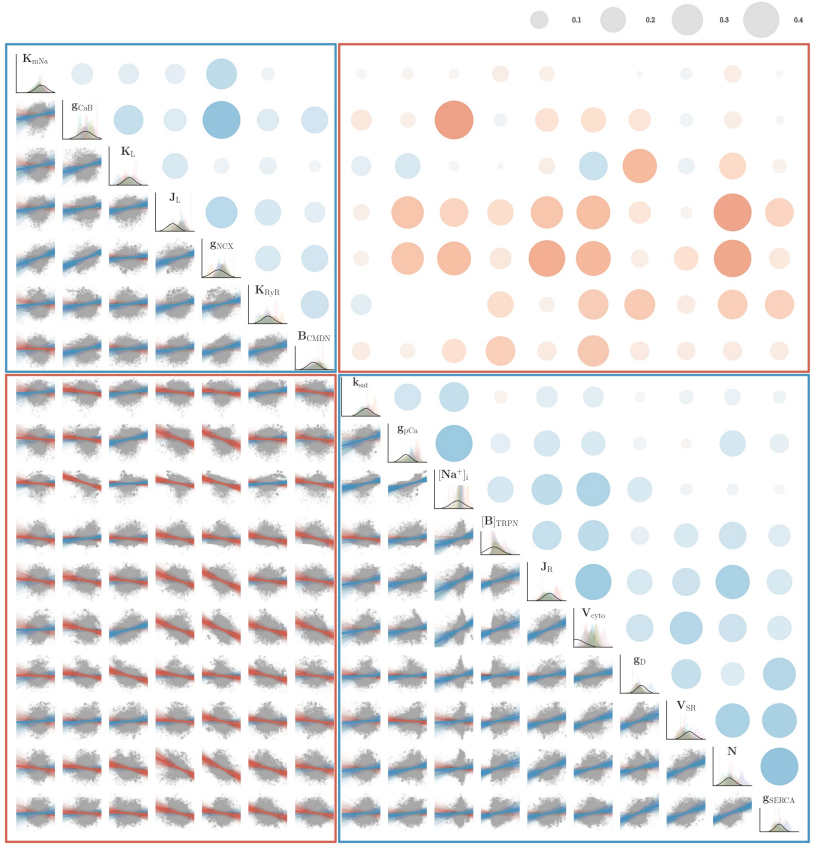
- Tempered MCMC output: ST achieves lower KSD values than SP, because it corrects for bias caused by tempering
- Standard MCMC output: ST achieves lower KSD values than SP, that is negatively affected by the non-convergence of MCMC


0.2
0.4
1.0
1.4
g_D
g_{SERCA}
N
V_{SR}
J_R
V_{cyto}
k_{sat}
g_{pCa}
$[Na^+]_i$
ΔV_L
k^-_{TRPN}
K_{RyR}
B_{CMDN}
ϕ_R
b
ϕ_L
r_{SR-I}
V_L
K_{pCa}
g_{CaB}
g_{NCX}
a
τ_L
J_L
t_L
η
K_{SERCA}
k_{CMDN}
θ_R
K_{mCa}
τ_R
k^+_{TRPN}
K_L
$[B]_{TRPN}$
K_{mNa}
d
c
t_R
Cells
prior
posterior
−20
−15
−10
−5
0
5
10
15

Parameters Correlation

- **Co-regulation** of calcium-handling proteins is at the basis of understanding how myocytes **preserve their functional activity** in different conditions
- To infer co-regulation, it is necessary to combine information derived from a **population** of cells
- In a Bayesian paradigm, this would be achieved through **hierarchical modelling**, but the available cohort of **18 cells was not sufficient** to estimate the high-dimensional correlation matrix of the ODEs parameters
- This was instead **approximated by averaging sample correlation matrices** computed at each Stein thinned MCMC iteration.

Parameters Correlation



- Meaningful correlations, between parameters related to **protein activity** (not kinetics)
- Indicating the existence of a **compensatory mechanism** between cell-membrane and intracellular proteins

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.
- S. Mak and V. R. Joseph. Support points. The Annals of Statistics, 46(6A):2562-2592, 2018.
- S. A. Niederer, M. S. Sacks, M. Girolami, K. Willcox. Scaling digital twins from the artisanal to the industrial. Nature Computational Science, 2021
- M. Strocchi et al. A publicly available virtual cohort of fourchamber heart meshes for cardiac electromechanics simulations, PloS one 15.6 (2020): e0235145
- G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. InterStat, 5(16.10):1249-1272, 2004.
- D. Vats and C. Knudson. Revisiting the Gelman-Rubin diagnostic. arXiv:1812.09384, 2018.

MCMC

Markov Chain Monte Carlo (MCMC): Construct an ergodic Markov chain (X_n) that has P as invariant distribution

MCMC

Markov Chain Monte Carlo (MCMC): Construct an ergodic Markov chain (X_n) that has P as invariant distribution

MCMC

Markov Chain Monte Carlo (MCMC): Construct an ergodic Markov chain (X_n) that has P as invariant distribution

- Markov chain: (X_n) is a correlated and ordered sequence of random variables, such that the current variable X_{n+1} is conditionally independent of the past $(X_m)_{m<n}$ given the last variable X_n

MCMC

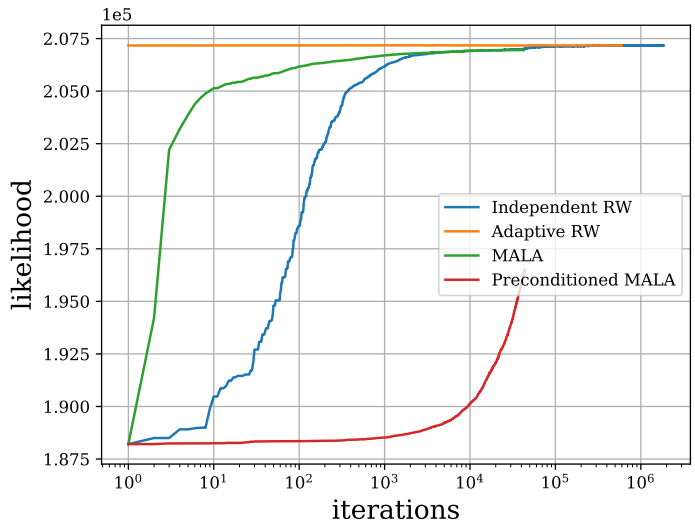
Markov Chain Monte Carlo (MCMC): Construct an ergodic Markov chain (X_n) that has P as invariant distribution

- Markov chain: (X_n) is a correlated and ordered sequence of random variables, such that the current variable X_{n+1} is conditionally independent of the past $(X_m)_{m<n}$ given the last variable X_n
- P-invariant and ergodic Markov chain: the probability distribution of the chain approximates P in the limit $n \to \infty$

Poor Performance of Several MCMC Methods

Let $\mathcal{L}(x) := \log p(x)$ and $z_n \sim \mathcal{N}(0, I)$. We implemented Metropolis-Hastings with **proposals**:

$$x_n = x_{n-1} + H\nabla\mathcal{L}(x_{n-1}) + Gz_n$$



Proposal	H	G
`RW`	0	ϵI
`ADA-RW`	0	$\sqrt{\hat{\Sigma}}$
`MALA`	$\frac{\epsilon^2}{2} I$	ϵI
`P-MALA`	$\frac{\epsilon^2}{2} M^{-1}(x_{n-1})$	$\epsilon\sqrt{M^{-1}(x_{n-1})}$

Linear case: FIM and Posterior Covariance

Given Σ_0 prior covariance and F the FIM, the posterior covariance in the **linear case** is

$$\begin{aligned}\Sigma_{post} &= (F + \Sigma_0^{-1})^{-1} \\ &= \Sigma_0^{1/2}(\underbrace{\Sigma_0^{T/2} F \Sigma_0^{1/2}}_{G} + I_P)^{-1}\Sigma_0^{T/2}\end{aligned}$$

Linear case: FIM and Posterior Covariance

Given Σ_0 prior covariance and F the FIM, the posterior covariance in the **linear case** is

$$\begin{aligned}\Sigma_{post} &= (F + \Sigma_0^{-1})^{-1} \\ &= \Sigma_0^{1/2}(\underbrace{\Sigma_0^{T/2} F \Sigma_0^{1/2}}_{G} + I_P)^{-1}\Sigma_0^{T/2}\end{aligned}$$

- The **sparsity** of F (or G, the **prior-preconditioned FIM**) indicates **conditional independence** between parameters $\Rightarrow$ blocking strategies in the samplers

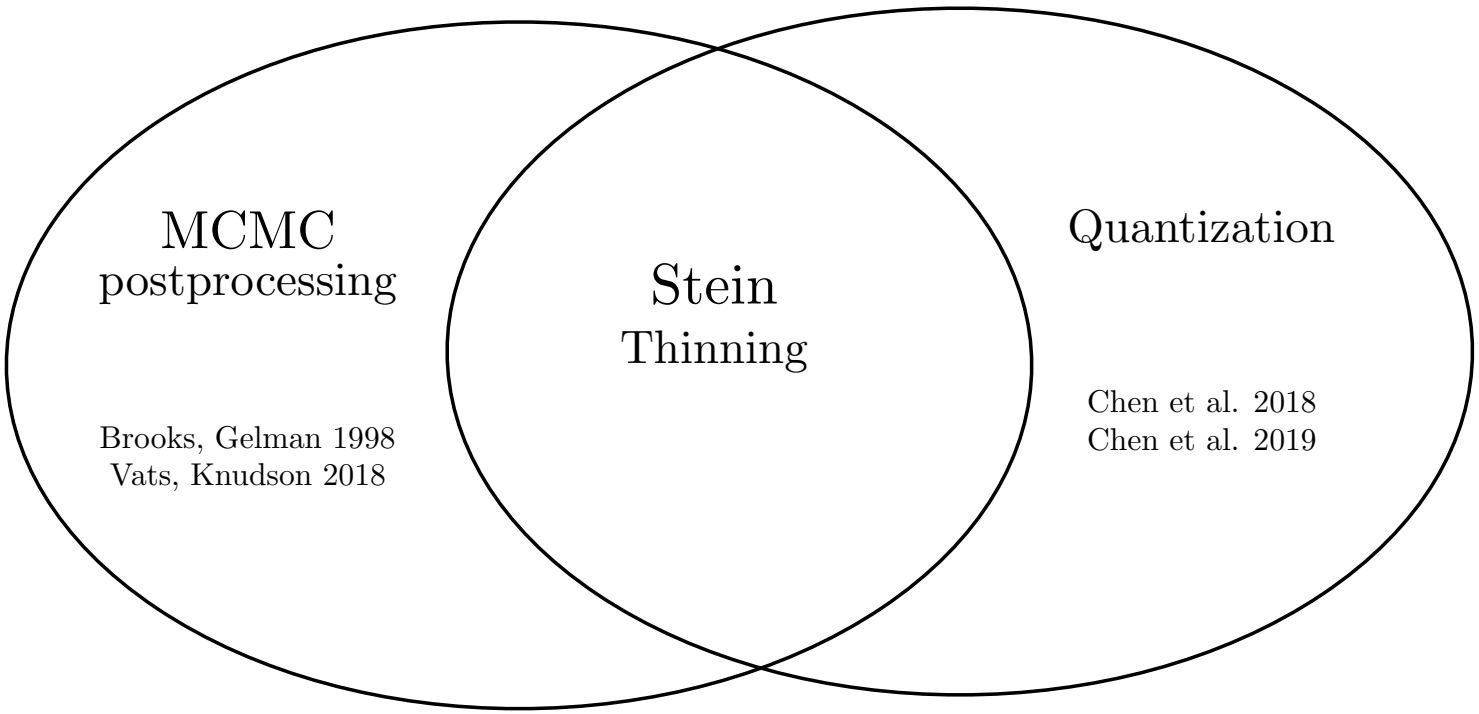
Linear case: FIM and Posterior Covariance

Given Σ_0 prior covariance and F the FIM, the posterior covariance in the **linear case** is

$$\begin{aligned}\Sigma_{post} &= (F + \Sigma_0^{-1})^{-1} \\ &= \Sigma_0^{1/2}(\underbrace{\Sigma_0^{T/2} F \Sigma_0^{1/2}}_{G} + I_P)^{-1}\Sigma_0^{T/2}\end{aligned}$$

- The **sparsity** of F (or G, the **prior-preconditioned FIM**) indicates **conditional independence** between parameters $\Rightarrow$ blocking strategies in the samplers
- F (or G) is s.t. F^{-1} is indicative of **posterior covariance**, after bringing the parameters to comparable scales. The decay of its spectrum indicates **identifiability**

Stein Thinning[1]



[1] Riabiz et al. 2020

$\mathcal{F}$ that leads from IPM to MMD

- We start with the IMP(μ, ν)
- *kernel*: symmetric and positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- k *reproduces* a Hilbert space of functions $\mathcal{H}$ from $\mathcal{X} \to \mathbb{R}$ if $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}$,
  1. $k(\cdot, x) \in \mathcal{H}$
  2. $\langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x)$

  There is a one-to-one mapping between the kernel k and the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$, so we write $\mathcal{H}(k)$
- Choosing the set $\mathcal{F}$ to be the unit ball $\mathcal{B}(k) := \{f \in \mathcal{H}(k) : \langle f, f \rangle_{\mathcal{H}(k)} \leq 1\}$ of the RKHS $\mathcal{H}(k)$ enables the supremum to be written in closed form and defines the MMD:

$$\mathrm{MMD}_{\mu,k}(\mu, \nu)^2 = \iint k(x, y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) - 2\iint k(x, y)\mathrm{d}\nu(x)\mathrm{d}\mu(y) + \iint k(x, y)\mathrm{d}\mu(x)\mathrm{d}\mu(y)$$

- Under suitable conditions on k and $\mathcal{X}$ it can be shown that
  1. MMD is a metric on $\mathcal{P}(\mathcal{X})$
  2. if $\mathrm{MMD}_{\mu,k}(\nu) \to 0$ then $\nu \Rightarrow \mu$

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

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

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 $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{K}}$ 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

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

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, $\int h \mathrm{d}P = 0$, $\forall h \in \mathcal{K}_P$

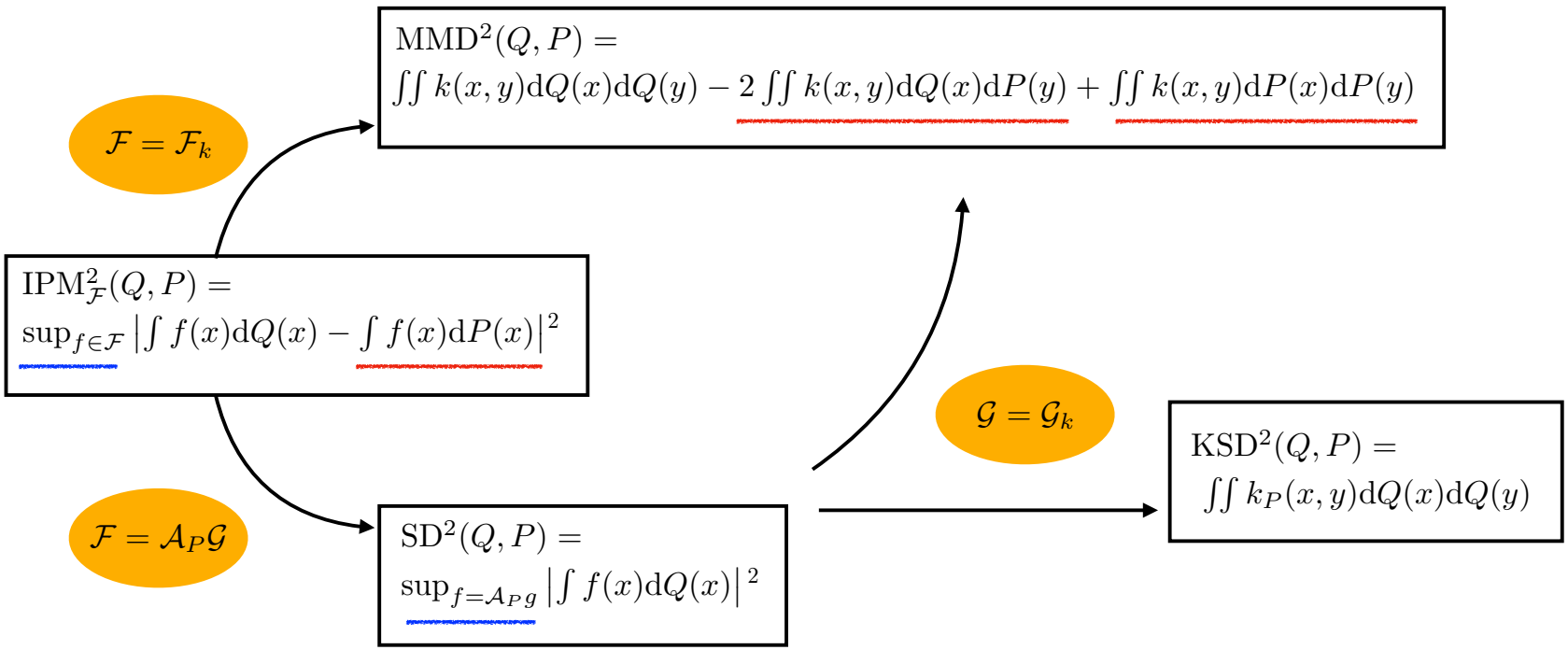
[1] i.e $\forall x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{K}$ and $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{K}}$ whenever $f \in \mathcal{K}$

Stein's Method Example

Example (Stein, 1972)

- $P = N(\mu, \sigma^2)$ with density function $p(x)$
- $\mathcal{A}_P : g \mapsto \frac{\nabla(gp)}{p}$
- $\mathcal{G} = \left\{ g : \mathbb{R} \to \mathbb{R} \text{ s.t. } \sup_{x \neq y} \max \left(|g(x)|, |\nabla g(x)|, \frac{|\nabla g(x) - \nabla g(y)|}{|x-y|} \right) \leq 1 \right\}$

Kernel Stein Discrepancy[1]



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

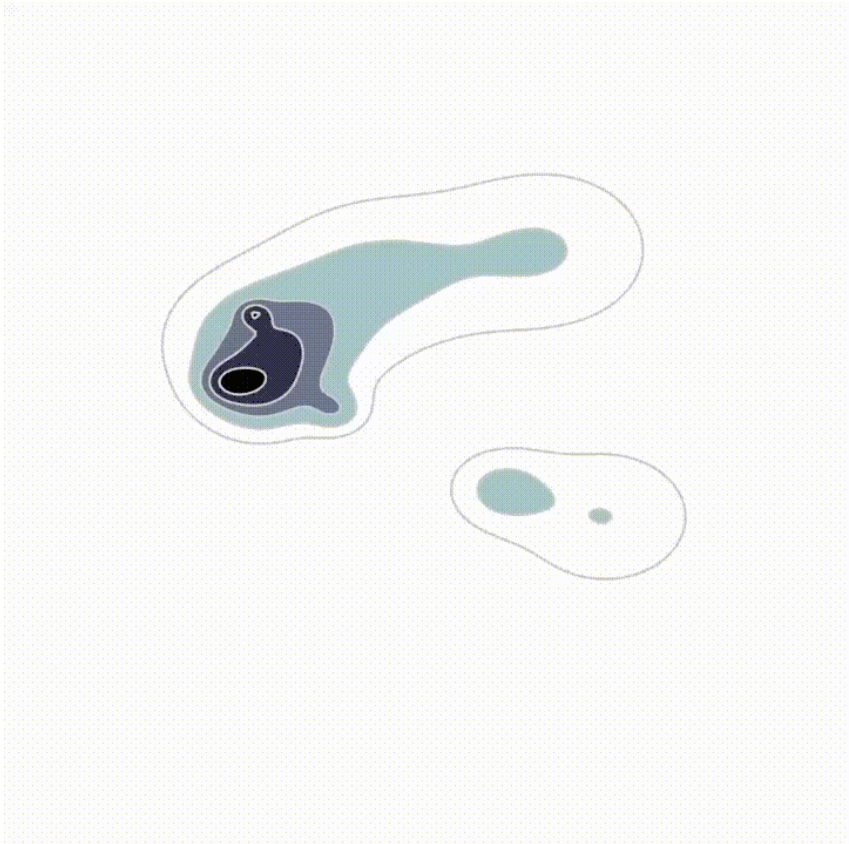
Choice of hyper-parameter Γ

Three settings for Γ in the base kernel k:

- **Median (`med`):** $\Gamma = \ell^2 I$, where
  $\ell = \text{med} := \text{median}\{\|X_i - X_j\| : 1 \leq i < j \leq n_0\}$
- **Scaled median (`sclmed`):** $\Gamma = \ell^2 I$, where $\ell = \text{med}/\sqrt{\log(m)}$
- **Sample covariance (`smpcov`):** Γ is the sample covariance matrix

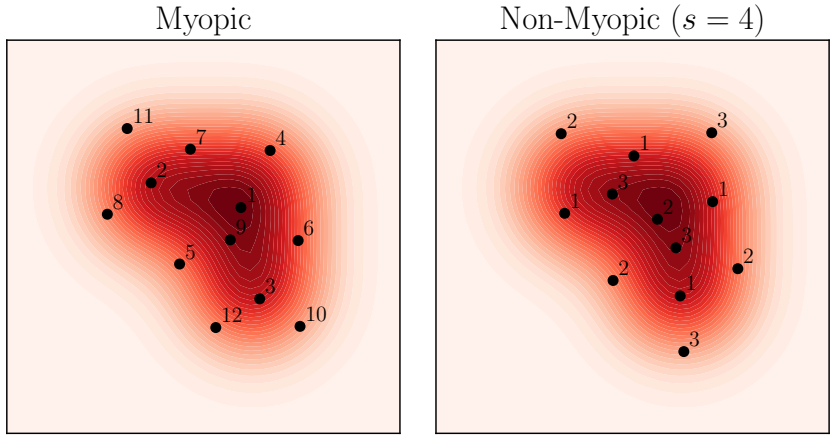
Step 2: KSD Optimization

The point set $S = \{\pi_{(1)} \ldots, \pi_{(m)}\} \subset \{1, \ldots, n\}^m$ is obtained by greedy minimization of the KSD



Step 2: Improved Minimization of KSD[1]

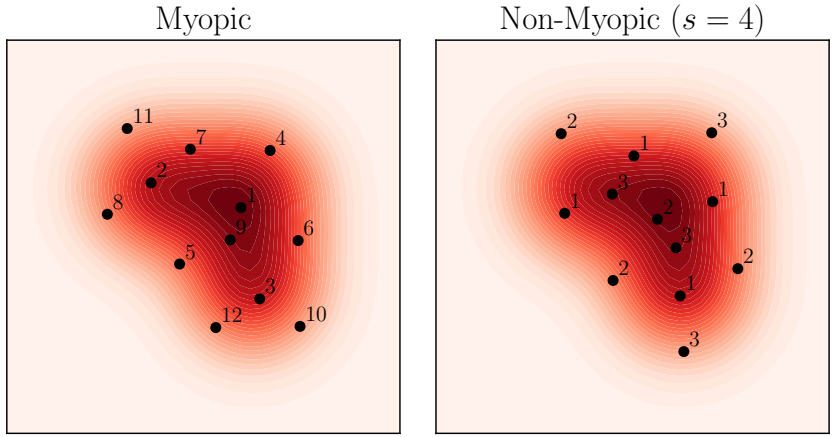
Non-myopic approach: improve the quality of the approximation by selecting $s > 1$ points at each iteration j



[1] Teymur et al., *Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy*, AISTATS 2021

Step 2: Improved Minimization of KSD[1]

Non-myopic approach: improve the quality of the approximation by selecting $s > 1$ points at each iteration j

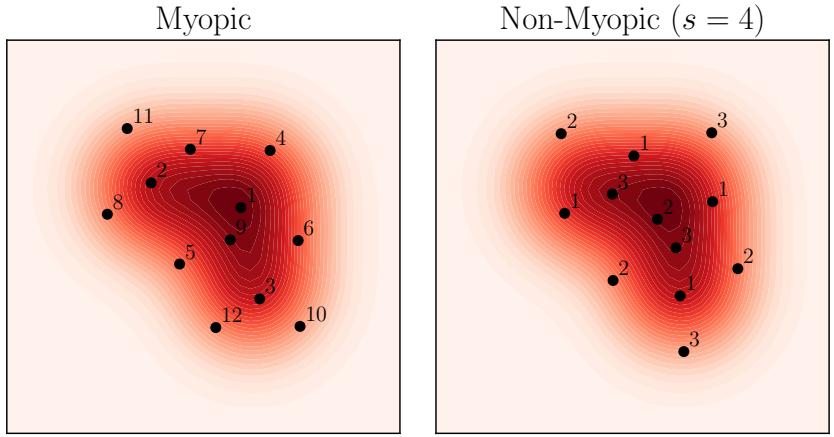


Mini-batching the set of candidate points $(X_i)_{i=1}^n$ to $(X_i^j)_{i=1}^b$ at each iteration j to reduce the length of the vector being scanned

[1] Teymur et al., *Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy*, AISTATS 2021

Step 2: Improved Minimization of KSD[1]

Non-myopic approach: improve the quality of the approximation by selecting $s > 1$ points at each iteration j



Mini-batching the set of candidate points $(X_i)_{i=1}^n$ to $(X_i^j)_{i=1}^b$ at each iteration j to reduce the length of the vector being scanned

Optimization is solved by re-casting to an **integer quadratic programme** with complexity $O(m^2 s^2 b^s)$, advantageous if $b \ll n$

[1] Teymur et al., *Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy*, AISTATS 2021

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

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

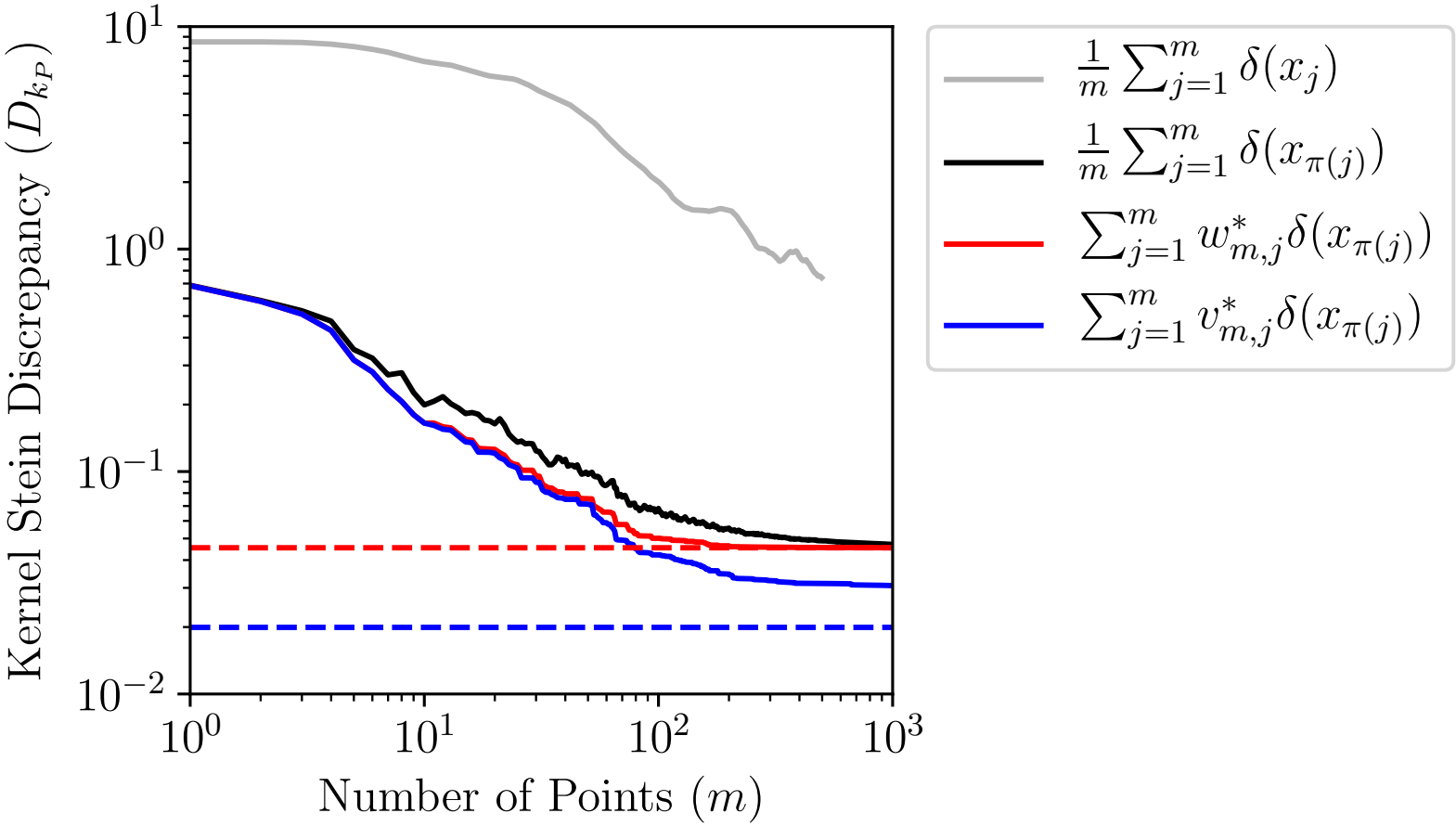
where the weights $w^* = (w_1^*, \ldots, w_n^*)$ satisfy

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

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

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

and

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

Result 3: Consistency (with Biased MCMC)

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

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

and

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

[1] $\tilde{P}$ 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:

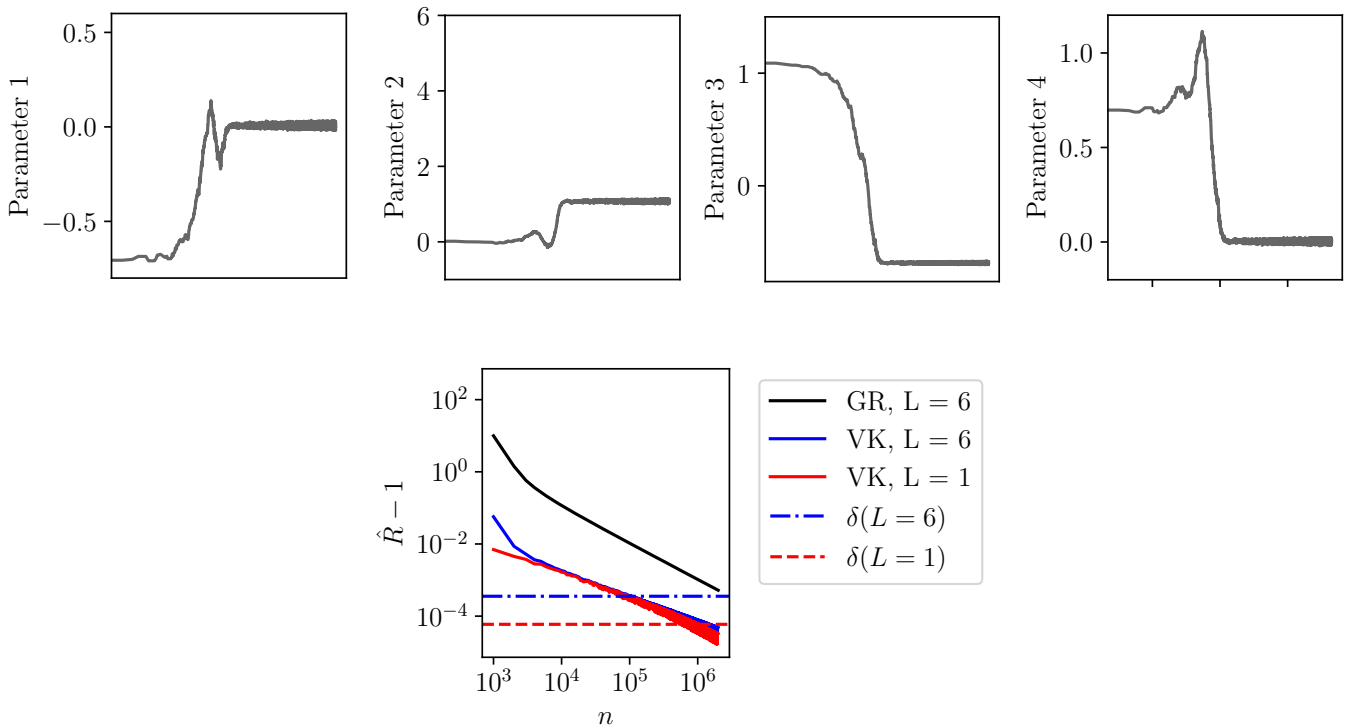
1. Energy distance[3]
2. KSD based on one setting for Γ

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

[2] Mak, Joseph 2018

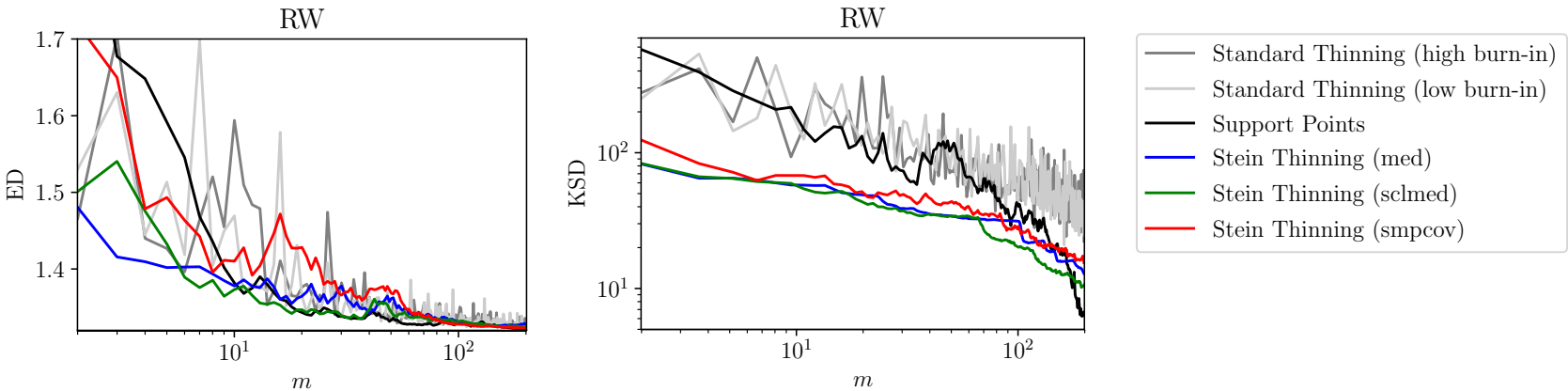
[3] Criterion minimized by Support points

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

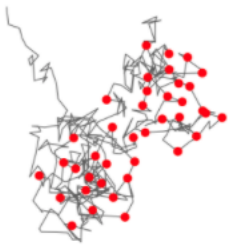
Goodwin Oscillator - Performance metrics



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

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

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