An Introduction to Stein Kernels

Robert Salomone

UNSW Sydney
r.salomone@unsw.edu.au

Introduction

- Exciting new area of research in Monte Carlo methods.
- Lots of papers in Machine Learning and Computational Statistics literature in the last few years.
- Main idea: Combine ideas from **Stein's Method** with **Reproducing Kernel Hilbert Space** (RKHS) theory.

Stein's Method: A History Lesson

- Stein's (1972) method — general **theoretical tool** for **bounding the distance between two distributions** p and q.

Integral Probability Metrics

- Let $\mathcal{F}$ is some class of **test functions** $\phi : \mathbb{R}^d \to \mathbb{R}$.
- For two random variables, $\boldsymbol{X} \sim q$, and $\boldsymbol{Y} \sim p$, an **Integral Probability Metric** (IPM) takes the form

$$d_{\mathcal{F}}(q, p) = \sup_{\phi \in \mathcal{F}} \left| \mathbb{E}_q \, \phi(\boldsymbol{X}) - \mathbb{E}_p \, \phi(\boldsymbol{Y}) \right|.$$

- Essentially, **worst case** error over all functions in $\mathcal{F}$ obtained by integrating using measure associated with q instead of p.
- Special choice of $\mathcal{F}$ can **recover some familiar friends**: e.g.,Total Variation, Maximum Mean Discrepancy, and Wasserstein.
- For a sequence of distributions $\{q^N\}_{N=1}^{\infty}$, a **sufficiently rich** choice of $\mathcal{F}$ ensures $d_{\mathcal{F}}(q^N, p) \to 0 \quad \Longrightarrow \quad q^N \to p$.

Stein's Insight

- How do you actually bound (let alone **compute**)

$$\sup_{\phi \in \mathcal{F}} \left| \mathbb{E}_q \, \phi(\boldsymbol{X}) - \mathbb{E}_p \, \phi(\boldsymbol{Y}) \right|,$$

  when it has **intractable expectations** and **supremum**?
- **Steins Insight**: Turn the difference of expectations into a **single computable expectation**.
- Main idea: Find an Operator $\mathcal{A}_p$ with property

$$\mathbb{E}\mathcal{A}_p\phi(\boldsymbol{X}) = 0 \text{ for all } \phi \in \widetilde{\mathcal{F}} \Longleftrightarrow \boldsymbol{X} \sim p,$$

  then, we have the IPM:

$$\mathbb{S}(q, p) = \sup_{\phi \in \widetilde{\mathcal{F}}} |\mathbb{E}_q[\mathcal{A}_p\phi(\boldsymbol{x})]|. \tag{1}$$

- Stein found such an operator for p being the univariate **Normal** distribution.

Barbour's Generator Approach

- Big step forward.
- Let $\{\boldsymbol{Z}_t\}_{t\geq 0}$ denote a Markov process with stationary distribution p.
- The **infinitesimal generator** $\mathcal{A}$ of $\{\boldsymbol{Z}_t\}_{t\geq 0}$ is defined pointwise by

$$\mathcal{A}\phi(\boldsymbol{x}) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\phi(\boldsymbol{Z}_t) \,|\, \boldsymbol{Z}_0 = \boldsymbol{x}_0]\bigg|_{t=0}.$$

- **Key Point**: This generator satisfies $\mathbb{E}_p \mathcal{A}\phi(\boldsymbol{X}) = 0$ under very mild conditions on ϕ and $\mathcal{A}$.
- So, let's create a **Stein Operator** for p with support on all of $\mathbb{R}^d$.

Langevin–Stein Operator

- **Langevin diffusion** $\{\boldsymbol{Z}_t\}_{t\geq 0}$ is the process:

$$\mathrm{d}\boldsymbol{Z}_t = \nabla \log p(\boldsymbol{Z}) + \sqrt{2}\mathrm{d}\boldsymbol{W}_t$$

  and has stationary distribution p.
- For $\boldsymbol{X}$ with density p, and $f \in \mathcal{C}^2$, the generator of the Langevin SDE with stationary distribution p is

$$\mathcal{A}f(\boldsymbol{x}) = \Delta f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot \nabla \log p(\boldsymbol{x}),$$

- As $f \in \mathcal{C}^2$, rewriting $\nabla f = \phi$ gives us

$$\mathcal{A}f(\boldsymbol{x}) = \nabla \phi(\boldsymbol{x}) + \phi(\boldsymbol{x}) \cdot \nabla \log p(\boldsymbol{x}),$$

  the RHS of which may seem more familiar.
- This is often called **the** Stein Operator in ML literature, but it really is just **a** Stein Operator.

So...

- The general form of **Stein Discrepancy** is

$$\mathbb{S}(q, p) = \sup_{\phi \in \widetilde{\mathcal{F}}} |\mathbb{E}_q[\mathcal{A}_p \phi(\boldsymbol{x})]|. \tag{2}$$

- Simpler, but to do something **practical** with this would ideally be able to solve the above **exactly**.
- Seems impossible, but using **kernel methods** get us around it.

Kernel Methods: A Crash Course

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive semi-definite function.
- Positive semi-definite function: The **Gram Matrix** $\mathrm{K} = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij}$ is positive definite for any number of points.
- Concrete example for now: for $\sigma^2 > 0$, **rbf kernel** is $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(||\boldsymbol{x} - \boldsymbol{y}||^2/2\sigma^2)$.
- **For our purposes**, it suffices to know that such kernel functions allow us to work implicitly (e.g., solve optimization problems) in complicated, possibly infinite dimensional, spaces of **functions**.

The Magic

- Let $\mathcal{H}$ be a **Reproducing Kernel Hilbert Space**[1].
- **The Magic** — insist that $\widetilde{\mathcal{F}}$ be the **unit ball** of $\mathcal{H}$:

$$\widetilde{\mathcal{F}} = \{\phi \in \mathcal{H} : ||\phi||_{\mathcal{H}} \leq 1\}.$$

- Then,

$$\mathbb{S}(q, p) = \sup_{\phi \in \widetilde{\mathcal{F}}} |\mathbb{E}_q[\mathcal{A}_p \phi(\boldsymbol{x})]|, \tag{3}$$

has the **closed form** solution

$$\sqrt{\mathbb{E}_{q \times q}\, k_p(\boldsymbol{X}, \boldsymbol{Y})},$$

where k_p is a special type of kernel called a **(Reproducing) Stein Kernel**. More on what precisely k_p is in a little while.

[1] intuitively, a space of functions induced by the choice of $k(\cdot, \cdot)$

Properties of Stein Kernels

- 1. We had **Stein's Identity**: $\mathbb{E}_p \mathcal{A}_p \phi(\boldsymbol{X}) = 0$.
  Now, for $\boldsymbol{X}, \boldsymbol{Y} \sim p$, we have

$$\underbrace{\mathbb{E}_{p \times p}\, k_p(\boldsymbol{X}, \boldsymbol{Y}) = 0}_{\textbf{Kernelized Stein Identity}}.$$

2. For $q \neq p$: $\mathbb{E}_{q \times q}\, k_p(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$, and thus

$$\underbrace{\sqrt{\mathbb{E}_{q \times q}\, k_p(\boldsymbol{X}, \boldsymbol{Y})}}_{\textbf{Kernelized Stein Discrepancy}} \geq 0.$$

Empirical Measures

- Recall, $\mathbb{S}(q,p) = \sup_{\phi \in \widetilde{\mathcal{F}}} |\mathbb{E}_q[\mathcal{A}_p \phi(\boldsymbol{x})]| = \sqrt{\mathbb{E}_{q\times q} k_p(\boldsymbol{X},\boldsymbol{Y})}$ for iid $\boldsymbol{X}$ and $\boldsymbol{Y}$. It is tempting to think we need to **estimate** this.
- However, often q is the empirical distribution of a **set of N points** $\{\boldsymbol{x}_k\}_{k=1}^N$, i.e.,

$$\mathbb{E}_{q^N} \phi(\boldsymbol{X}) = \frac{1}{N} \sum_{k=1}^{N} \phi(\boldsymbol{x}_k).$$

- We do this all the time in Monte Carlo / Quasi Monte Carlo (use an empirical measure).
- The KSD in such a case is

$$\mathbb{S}(q^N,p) = \sqrt{\mathbb{E}_{q^N\times q^N} k_p(\boldsymbol{X},\boldsymbol{Y})} = \sqrt{\frac{1}{N^2} \sum_{j=1}^{N}\sum_{k=1}^{N} k_p(\boldsymbol{x}_j,\boldsymbol{x}_k)}.$$

- This is not a **Monte Carlo estimate** of the KSD, this **is** the KSD of your point set! An **explicitly computable** IPM.

An Explicitly Computable Integral Probability Metric!

How do you construct Stein kernels though?

Step 1: Choose a Base Kernel

- We begin with a **base** kernel, a symmetric positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For example:

kernel	$k(\boldsymbol{x}, \boldsymbol{y})$	parameters
Polynomial	$(\boldsymbol{x}^\top \boldsymbol{y} + c)^d$	$c \in \mathbb{R}_+, d \in \mathbb{N}$
Gaussian (rbf)	$\exp(\|\|\boldsymbol{x} - \boldsymbol{y}\|\|^2 / 2\sigma^2)$	$\sigma^2 > 0$
Inverse Multiquadric (IMQ)	$(1 + \|\|\boldsymbol{x} - \boldsymbol{y}\|\|_2^2)^{-\beta}$	$\beta \in (\mathbf{0}, \mathbf{1})$

- Many others: Matérn, Exponential, you can easily create your own.

Step 2: "Steinalize" the Base Kernel

- **Main Idea**: Hit the base kernel with a Stein Operator on both arguments, i.e., $\mathcal{A}_p^{\boldsymbol{x}}$ and $\mathcal{A}_p^{\boldsymbol{y}}$.
- In the case where $\mathcal{A}$ is the **Langevin–Stein** operator, this yields

$$\begin{aligned} k_p(\boldsymbol{x}, \boldsymbol{y}) :=& \nabla \log p(\boldsymbol{x})^\top k(\boldsymbol{x}, \boldsymbol{y}) \nabla \log p(\boldsymbol{y}) \\ &+ \nabla \log p(\boldsymbol{x})^\top \nabla_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}) \\ &+ \nabla \log p(\boldsymbol{y})^\top \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}) \\ &+ \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

- Still no need of the normalizing constant!

Properties of Stein Kernels (Continued)

- Stein Kernels are **actually Kernels**. We have

$$(\text{symmetric } k \succ 0) \Longrightarrow (\text{symmetric } k_p \succ 0).^{2}$$

- KSD with **IMQ kernel** and **Langevin–Stein** operator controls weak convergence[3], i.e.,

$$\mathbb{S}(q^N, p) \to 0 \quad \Longrightarrow \quad q^N \to p.$$

- This property is called **convergence-determining**.
- Surprisingly, one can construct cases where common kernels **fail** to have this property.

---

[2] Oates et al., 2017 for Langevin–Stein kernels. More general results by Hodgkinson et al., 2020.

[3] Gorham and Mackey, 2017, Theorem 8

Succinctly put...

- The **Stein Operator** removes the need to compute the expectations $\mathbb{E}_p \phi$.
- The **Stein Kernel** saves you from solving for the supremum.
- For IMQ kernel, KSD is still equal to some **convergence determining** IPM.
- So, what can we actually do with this?

Methods

1. Measuring Sample Quality
2. Construct a Sequence of Quality Samples
3. Improve the Quality of Existing Samples

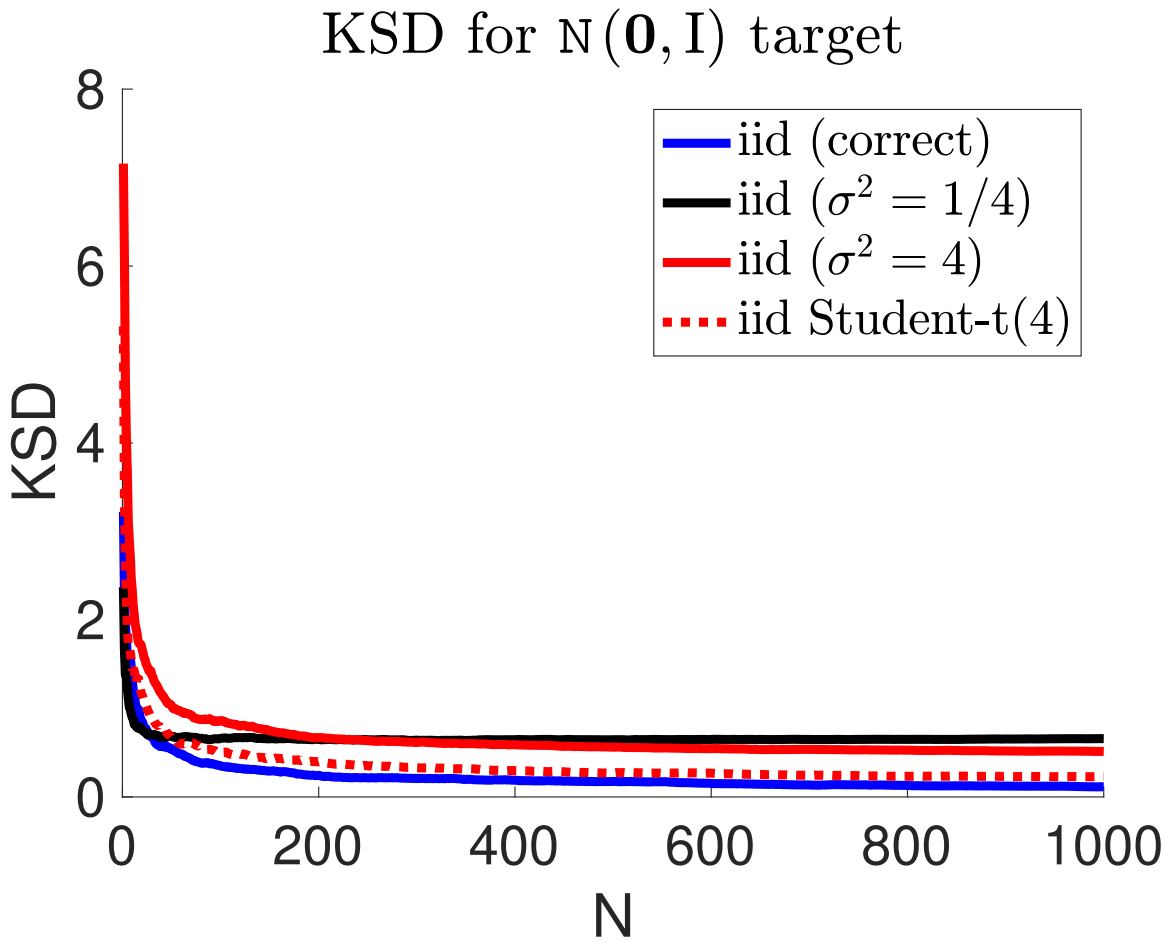
Measuring Sample Quality

- As mentioned, we have a **general tool** for assessing how well a collection of points approximate expectations with respect to a target distribution.
- How you got the points **doesn't matter**!
  - Asymptotically Unbiased MCMC (e.g., Metropolis–Hastings)
  - Asymptotically Biased MCMC (e.g., SGLD, ULA)
  - Sequential Monte Carlo / Sequentually Interacting MCMC
  - Constructed the Point Set Deterministically (e.g., QMC)

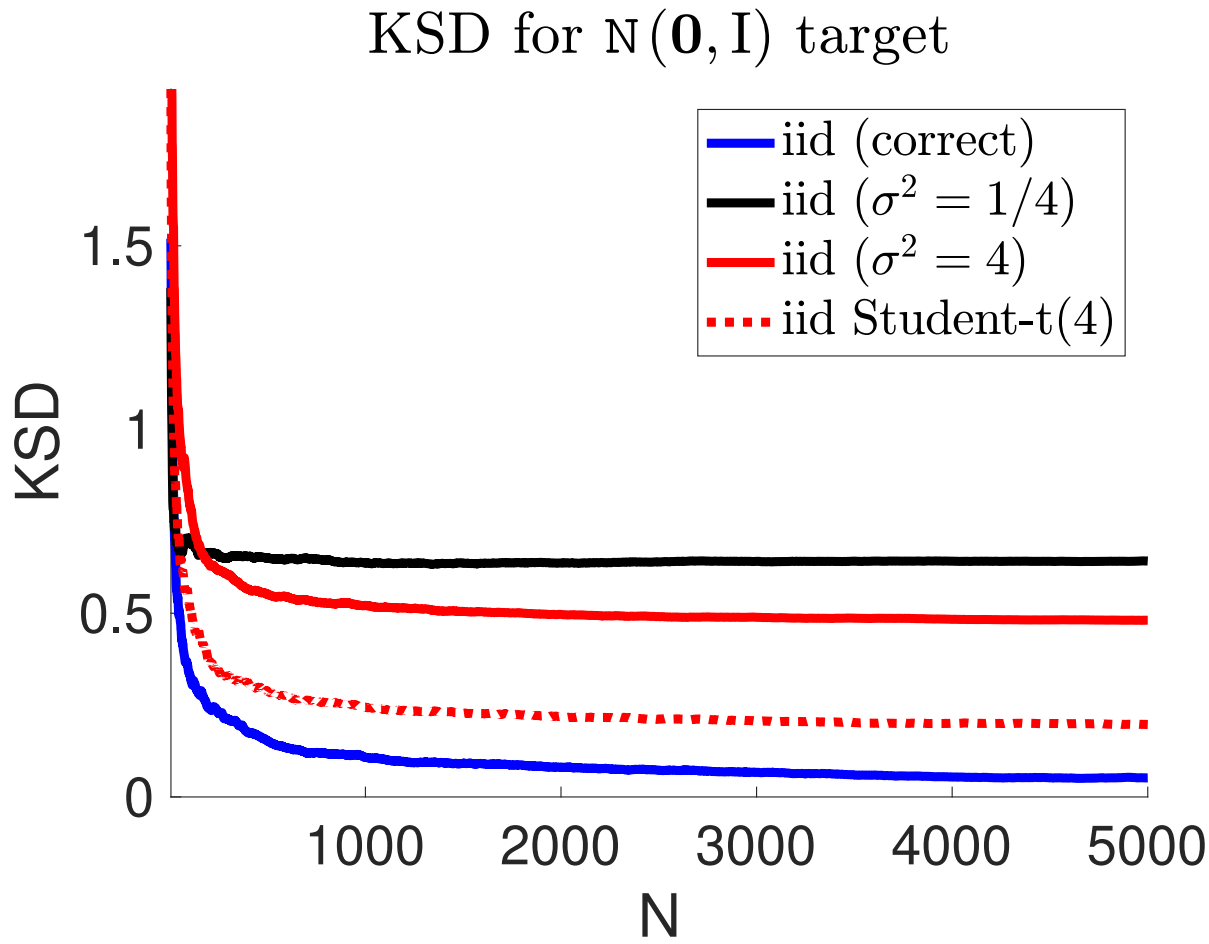
KSD in Action

- Suppose that our target p is a 10–dimensional standard normal distribution, $\mathcal{N}(\mathbf{0}, \mathrm{I})$.
- We will evaluate KSD with point sets generated via:
  1. (perfect) iid sampling from p.
  2. iid samples from $\mathcal{N}(\mathbf{0}, 0.25 \times \mathrm{I})$ (underdispersed)
  3. iid samples from $\mathcal{N}(\mathbf{0}, 4 \times \mathrm{I})$ (overdispersed)
  4. iid samples from student–t$(\mathbf{0}, \mathrm{I}, 4)$

KSD in Action



KSD in Action (More Samples)



So...

- **Practitioners**: You can use KSD to evaluate the quality of points generated by **any** sampling algorithm, and use this to tune hyperparameters (example forthcoming).
- **Methodologists**: You can use KSD to show your sampling methods are better than other sampling methods.

Hyperparameter Tuning

- In *Measuring Sample Quality with Kernels* (Gorham and Mackey, 2017), the authors consider tuning an **asymptotically biased** Slice Sampler.
- Effective Sample Size (ESS) would say set steps as large as possible (minimize autocorrelation), but this is clearly a bad idea (increases bias).
- However, KSD makes a different recommendation...

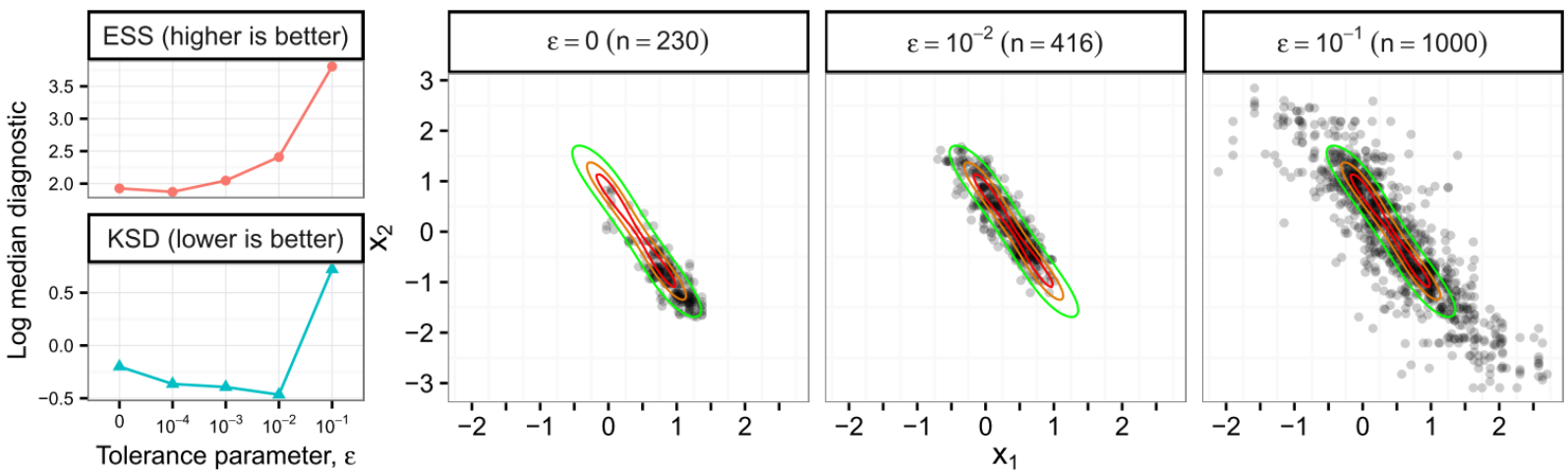


Figure: Figure from *Measuring Sample Quality with Kernels* (Gorham and Mackey, 2017)

Construct a Sequence of Quality Samples

Stein Points (Chen et. al., 2018)

- Main Idea: Construct point sequences that minimize the KSD!
- **Extensible Point Sets**.
- Two methods:
  1. Greedy Algorithm
  2. Herding Algorithm
- For empirical distribution of N Stein Points q^N obtained by both of the above methods, the authors establish that $\mathbb{S}(q^N, p) \to 0$.

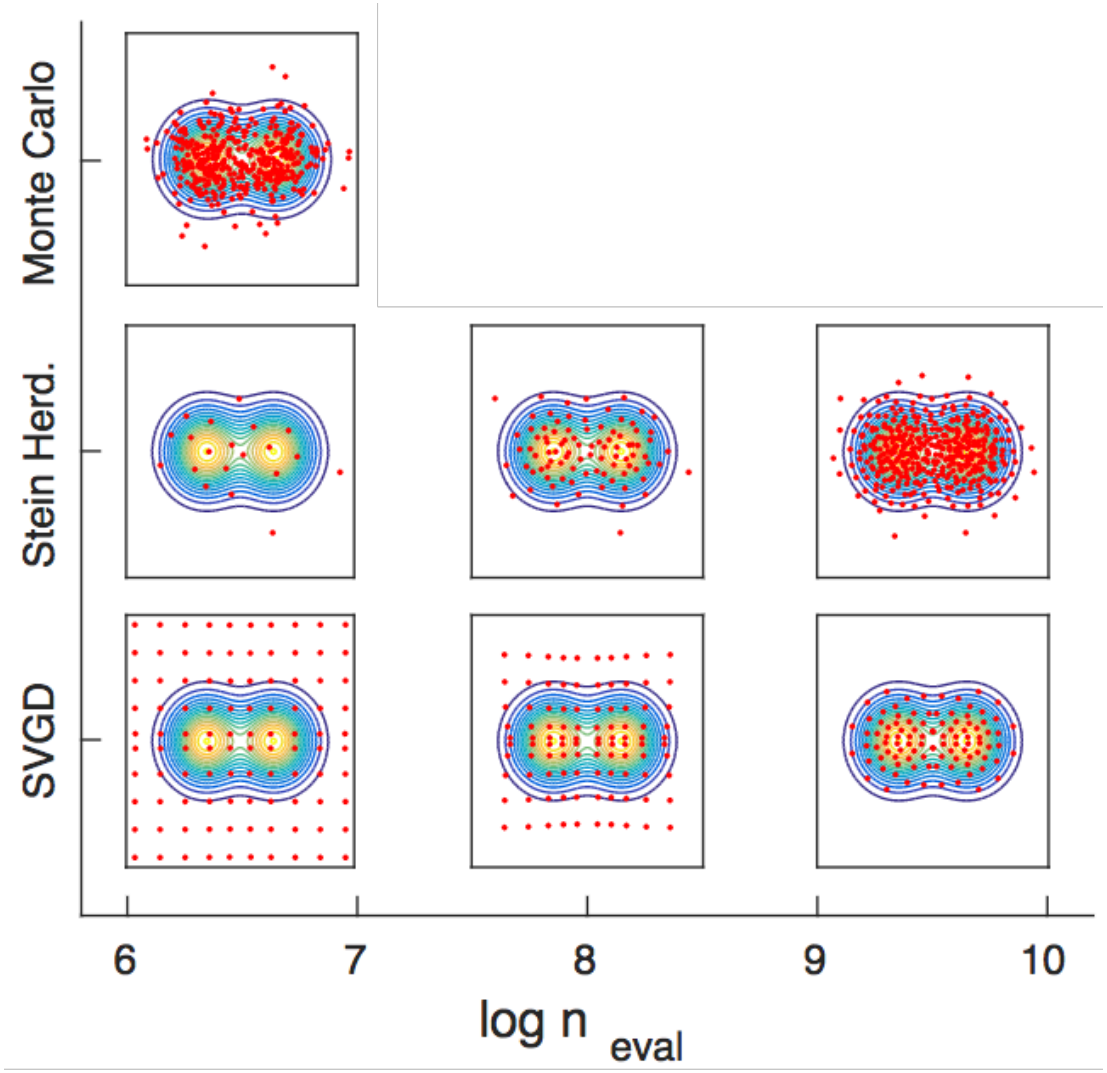


Figure: Modified Figure 1 from *Stein Points* paper.

Stein Importance Sampling (Liu and Lee, 2016)

- **Classic** Importance sampling. Instead of sampling from p, we sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ according to q and have

$$\sum_{k=1}^{N} \phi(\boldsymbol{X}_k) \frac{w_k}{\sum_{k=1}^{N} w_k}$$

  as an estimator of $\mathbb{E}_p \phi(\boldsymbol{X})$.
- Calculating the **importance weights** $w_k = p(\boldsymbol{X}_k)/q(\boldsymbol{X}_k)$ require evaluation of p and q pointwise, though they need not be normalized.
- **Consistent**, but has disadvantages. There may be better choices of weight functions.

Improve the Quality of Existing Samples

Stein Importance Sampling

- Idea is simple, for **fixed points**, consider **weighted** empirical distributions $q_{\boldsymbol{w}}^N$, i.e.,

$$\mathbb{E}_{q_{\boldsymbol{w}}^N}\phi(\boldsymbol{X}) = \sum_{k=1}^N w_k \phi(\boldsymbol{x}_k), \quad \boldsymbol{w} \geq 0, \sum_{k=1}^N w_k = 1.$$

- Define the **Stein Gram Matrix** $\mathrm{K}_p := (k_p(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij}$.
- We solve:

$$\begin{aligned}\boldsymbol{w}^\star &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left\{ \mathbb{S}(q_{\boldsymbol{w}}^N, p) : \boldsymbol{w} \geq 0, \sum_{k=1}^N w_k = 1 \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left\{ \boldsymbol{w}^\top \mathrm{K}_p \boldsymbol{w} : \boldsymbol{w} \geq 0, \sum_{k=1}^N w_k = 1 \right\}\end{aligned}$$

- Interestingly, density q need not be known, and the procedure reduces MSE quite well in practice.

New Result

- **Research Question:** Is it possible to use the output of a Markov Chain with ergodic distribution $q \neq p$ to construct **consistent** estimators for expectations under p by using **Stein Importance Sampling**?
- **Answer:** Yes! (under certain conditions) This is our new result. See Theorem 1 of

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

- In the above paper, we also provide **general theory** for construction of Stein Kernels on **arbitrary polish spaces** and conditions for such kernels to be convergence-determining.

Conspiring Scholars



Liam Hodgkinson Robert Salomone Fred Roosta

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

Some Other Methods

- **Black–Box Variational Inference**: Operator Variational Inference (Ranganath et al., 2016), Stein Variational Gradient Descent (Liu and Chen, 2016), Wild Variational Inference (Liu, 2016)
- **Reinforcement Learning:** Stein Variational Policy Gradient (Liu, 2017)
- **Generative Models**: SteinGAN (Wang et al., 2017)
- **Non-Parametric Control Variates**: Control Functionals (Oates et al., 2017)
- **Stein Variational Adaptive Importance Sampling** (Han and Liu, 2017).

Thankyou for your attention.