Stein's Method, Learning, and Inference

Lester Mackey

Microsoft Research New England

Collaborators: Jackson Gorham, Andrew Duncan, Sebastian Vollmer, Jonathan Huggins, Wilson Chen, Alessandro Barp, Francois-Xavier Briol, Mark Girolami, Chris Oates, Murat Erdogdu, Ohad Shamir, Marina Riabiz, Jon Cockayne, Pawel Swietach, Steven Niederer, Anant Raj, Heishiro Kanagawa, Wittawat Jitkrittum, Kenji Fukumizu, Arthur Gretton, Jiaxin Shi, Chang Liu, Yuhao Zhou, Jessica Hwang, Michalis Titsias, and Carl-Johann Simon-Gabriel

Motivation: Large-scale Posterior Inference

Example: Bayesian logistic regression

1. Feature vectors: $v_l \in \mathbb{R}^d$ for each datapoint $l = 1, \dots, L$
2. Binary class labels: $Y_l \in \{0, 1\}$, $\mathbb{P}(Y_l = 1 \mid v_l, \beta) = \frac{1}{1+e^{-\langle \beta, v_l \rangle}}$
3. Unknown parameter vector: $\beta \sim \mathcal{N}(0, I)$

v_1	v_2	v_3	v_4
3	1	0	9
0	6	8	4
5	0	3	1
2	9	7	0
Class 0		Class 1	

- Generative model simple to express
- Posterior distribution over unknown parameters is complex
  - Normalization constant unknown, exact integration intractable

Standard inferential approach: Use Markov chain Monte Carlo (MCMC) to (eventually) draw samples from the posterior distribution

- **Benefit:** Approximates intractable posterior expectations $\mathbb{E}_P[h(Z)] = \int p(x)h(x)dx$ with asymptotically exact sample estimates $\mathbb{E}_Q[h(X)] = \frac{1}{n}\sum_{i=1}^n h(x_i)$
- **Problem:** Each new MCMC sample point x_i requires iterating over entire observed dataset: prohibitive when dataset is large!

Motivation: Large-scale Posterior Inference

Question: How do we scale Markov chain Monte Carlo (MCMC) posterior inference to massive datasets?

- **Benefit:** Approximates intractable posterior expectations $\mathbb{E}_P[h(Z)] = \int p(x)h(x)dx$ with asymptotically exact sample estimates $\mathbb{E}_Q[h(X)] = \frac{1}{n}\sum_{i=1}^n h(x_i)$
- **Problem:** Each point x_i requires iterating over entire dataset!

Template solution: Approximate MCMC with subset posteriors

[Welling and Teh, 2011, Ahn, Korattikara, and Welling, 2012, Korattikara, Chen, and Welling, 2014]

- Approximate standard MCMC procedure in a manner that makes use of only a small subset of datapoints per sample
- Reduced computational overhead leads to faster sampling and reduced Monte Carlo variance
- Introduces asymptotic bias: target distribution is not stationary
- Hope that for fixed amount of sampling time, variance reduction will outweigh bias introduced

Motivation: Large-scale Posterior Inference

Template solution: Approximate MCMC with subset posteriors

[Welling and Teh, 2011, Ahn, Korattikara, and Welling, 2012, Korattikara, Chen, and Welling, 2014]

- Hope that for fixed amount of sampling time, variance reduction will outweigh bias introduced

Introduces new challenges

- How do we compare and evaluate samples from approximate MCMC procedures?
- How do we select samplers and their tuning parameters?
- How do we quantify the bias-variance trade-off explicitly?

Difficulty: Standard evaluation criteria like effective sample size, trace plots, and variance diagnostics assume convergence to the target distribution and do not account for asymptotic bias

This talk: Introduce new quality measures suitable for comparing the quality of approximate MCMC samples

Quality Measures for Samples

Challenge: Develop measure suitable for comparing the quality of *any* two samples approximating a common target distribution

Given

- **Continuous target distribution** P with support $\mathcal{X} = \mathbb{R}^d$ and density p
  - p known up to normalization, integration under P is intractable
- **Sample points** $x_1, \ldots, x_n \in \mathcal{X}$
  - Define **discrete distribution** Q_n with, for any function h, $\mathbb{E}_{Q_n}[h(X)] = \frac{1}{n}\sum_{i=1}^n h(x_i)$ used to approximate $\mathbb{E}_P[h(Z)]$
  - We make no assumption about the provenance of the x_i

Goal: Quantify how well $\mathbb{E}_{Q_n}$ approximates $\mathbb{E}_P$ in a manner that

I. Detects when a sample sequence is converging to the target

II. Detects when a sample sequence is not converging to the target

III. Is computationally feasible

Integral Probability Metrics

Goal: Quantify how well $\mathbb{E}_{Q_n}$ approximates $\mathbb{E}_P$

Idea: Consider an **integral probability metric (IPM)** [Müller, 1997]

$$d_{\mathcal{H}}(Q_n, P) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{Q_n}[h(X)] - \mathbb{E}_P[h(Z)]|$$

- Measures maximum discrepancy between sample and target expectations over a class of real-valued test functions $\mathcal{H}$
- When $\mathcal{H}$ sufficiently large, convergence of $d_{\mathcal{H}}(Q_n, P)$ to zero implies $(Q_n)_{n\geq 1}$ converges weakly to P (Requirement II)

Problem: Integration under P intractable!

$\Rightarrow$ Most IPMs cannot be computed in practice

Idea: Only consider functions with $\mathbb{E}_P[h(Z)]$ known *a priori* to be 0

- Then IPM computation only depends on Q_n!
- How do we select this class of test functions?
- Will the resulting discrepancy measure track sample sequence convergence?
- How do we solve the resulting optimization problem in practice?

Stein's Method

Stein's method [1972] provides a recipe for controlling convergence:

1. **Identify operator $\mathcal{T}$ and set $\mathcal{G}$** of functions $g : \mathcal{X} \to \mathbb{R}^d$ with
$$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \quad \text{for all} \quad g \in \mathcal{G}.$$
$\mathcal{T}$ and $\mathcal{G}$ together define the **Stein discrepancy** [Gorham and Mackey, 2015]
$$\boldsymbol{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})} \triangleq \sup_{g \in \mathcal{G}} |\mathbb{E}_{Q_n}[(\mathcal{T}g)(X)]| = d_{\mathcal{T}\mathcal{G}}(Q_n, P),$$
an IPM-type measure with no explicit integration under P

2. **Lower bound $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})$ by reference IPM $d_{\mathcal{H}}(Q_n, P)$** $\Rightarrow (Q_n)_{n\geq 1}$ converges to P whenever $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}) \to 0$ (Requirement II)
   - Performed once, in advance, for large classes of distributions

3. **Upper bound $\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})$ by any means necessary** to demonstrate convergence to 0 (Requirement I)

Standard use: As analytical tool to prove convergence

Our goal: Develop Stein discrepancy into practical quality measure

Identifying a Stein Operator $\mathcal{T}$

Goal: Identify operator $\mathcal{T}$ for which $\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0$ for all $g \in \mathcal{G}$

Approach: Generator method of Barbour [1988, 1990], Götze [1991]

- Identify a Markov process $(Z_t)_{t\geq 0}$ with stationary distribution P
- Under mild conditions, its **infinitesimal generator**

$$(\mathcal{A}u)(x) = \lim_{t\to 0} (\mathbb{E}[u(Z_t) \mid Z_0 = x] - u(x))/t$$

  satisfies $\mathbb{E}_P[(\mathcal{A}u)(Z)] = 0$

Overdamped Langevin diffusion: $dZ_t = \frac{1}{2}\nabla \log p(Z_t)dt + dW_t$

- Generator: $(\mathcal{A}_P u)(x) = \frac{1}{2}\langle \nabla u(x), \nabla \log p(x)\rangle + \frac{1}{2}\langle \nabla, \nabla u(x)\rangle$
- **Stein operator:** $(\mathcal{T}_P g)(x) \triangleq \langle g(x), \nabla \log p(x)\rangle + \langle \nabla, g(x)\rangle$

  [Gorham and Mackey, 2015, Oates, Girolami, and Chopin, 2016]

  - Depends on P only through $\nabla \log p$; computable even if p cannot be normalized!
  - $\mathbb{E}_P[(\mathcal{T}_P g)(Z)] = 0$ for all $g : \mathcal{X} \to \mathbb{R}^d$ in **classical Stein set**

$$\mathcal{G}_{\|\cdot\|} = \left\{ g : \sup_{x\neq y} \max\left(\|g(x)\|^*, \|\nabla g(x)\|^*, \frac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x-y\|} \right) \leq 1 \right\}$$

Detecting Convergence and Non-convergence

Goal: Show **classical Stein discrepancy** $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ if and only if $(Q_n)_{n\geq 1}$ converges to P

- In the univariate case ($d = 1$), known that for many targets P, $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ only if Wasserstein $d_{\mathcal{W}_{\|\cdot\|}}(Q_n, P) \to 0$

  [Stein, Diaconis, Holmes, and Reinert, 2004, Chatterjee and Shao, 2011, Chen, Goldstein, and Shao, 2011]
- Few multivariate targets have been analyzed (see [Reinert and Röllin, 2009, Chatterjee and Meckes, 2008, Meckes, 2009] for multivariate Gaussian)

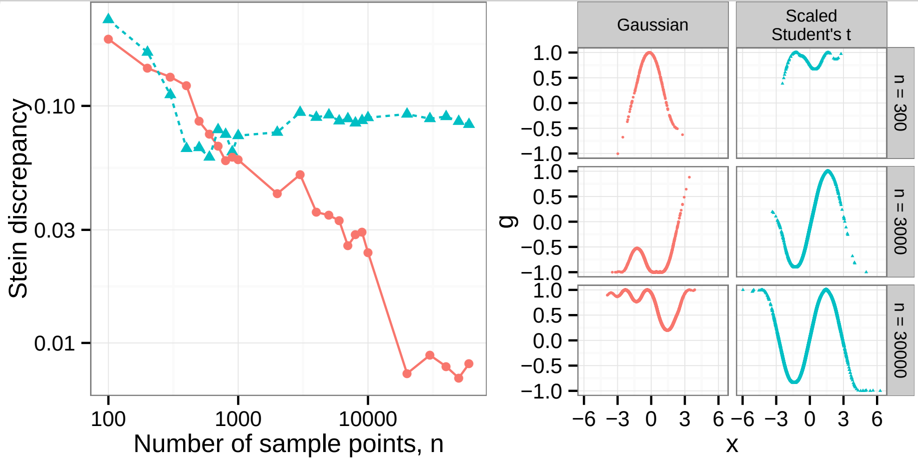
New contribution [Gorham, Duncan, Vollmer, and Mackey, 2019]

Theorem (Stein Discrepancy-Wasserstein Equivalence)

If the Langevin diffusion couples at an integrable rate and $\nabla \log p$ is Lipschitz, then $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0 \Leftrightarrow d_{\mathcal{W}_{\|\cdot\|}}(Q_n, P) \to 0$.

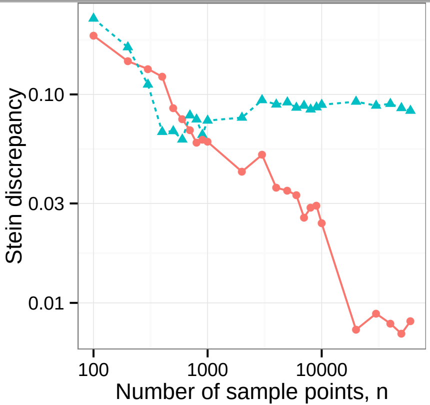
- Examples: strongly log concave P, Bayesian logistic regression or robust t regression with Gaussian priors, Gaussian mixtures
- Conditions not necessary: template for bounding $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$

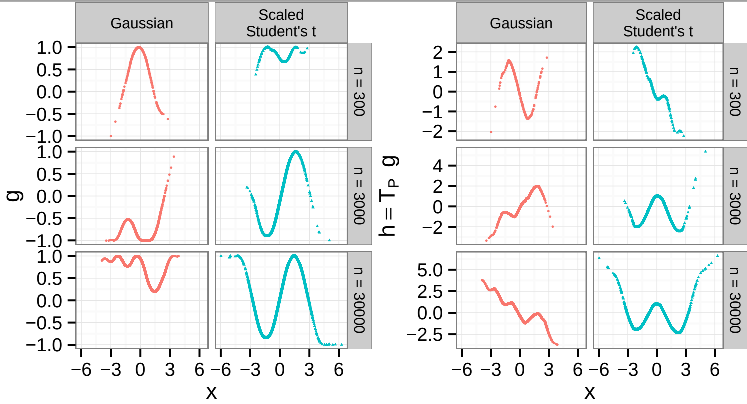
A Simple Example



- For target $P = \mathcal{N}(0,1)$, compare i.i.d. $\mathcal{N}(0,1)$ sample sequence $Q_{1:n}$ to scaled Student's t sequence $Q'_{1:n}$ with matching variance
- Expect $\mathcal{S}(Q_{1:n}, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q,G_1}) \to 0$ & $\mathcal{S}(Q'_{1:n}, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q,G_1}) \not\to 0$

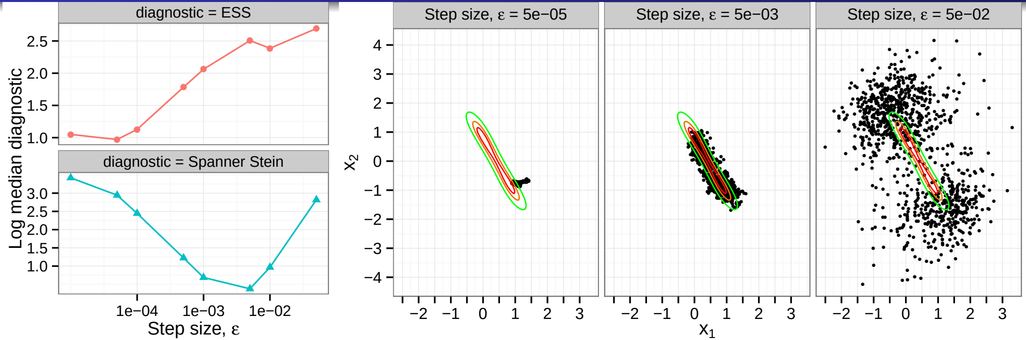
A Simple Example





- **Middle:** Recovered optimal functions g
- **Right:** Associated test functions $h(x) \triangleq (\mathcal{T}_P g)(x)$ which best discriminate sample Q from target P

Selecting Sampler Hyperparameters



Target posterior density: $p(x) \propto \pi(x) \prod_{l=1}^{L} \pi(y_l \mid x)$

Stochastic Gradient Langevin Dynamics [Welling and Teh, 2011]

$$x_{k+1} \sim \mathcal{N}(x_k + \frac{\epsilon}{2}(\nabla \log \pi(x_k) + \frac{L}{|\mathcal{B}_k|} \sum_{l \in \mathcal{B}_k} \nabla \log \pi(y_l|x_k)), \epsilon I)$$

- Random batch $\mathcal{B}_k$ of datapoints used to draw each sample point
  - Step size ϵ too small $\Rightarrow$ slow mixing
  - Step size ϵ too large $\Rightarrow$ sampling from very different distribution
  - Standard diagnostics like **effective sample size** (ESS) do not account for this bias

Alternative Stein Sets $\mathcal{G}$

Goal: Identify a more "user-friendly" Stein set $\mathcal{G}$ than the classical

Approach: Reproducing kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [Oates, Girolami, and Chopin, 2016, Chwialkowski, Strathmann, and Gretton, 2016, Liu, Lee, and Jordan, 2016]

- A reproducing kernel k is symmetric ($k(x, y) = k(y, x)$) and positive semidefinite ($\sum_{i,l} c_i c_l k(z_i, z_l) \geq 0, \forall z_i \in \mathcal{X}, c_i \in \mathbb{R}$)
  - Gaussian: $k(x, y) = e^{-\frac{1}{2}\|x-y\|_2^2}$, IMQ: $k(x, y) = \frac{1}{(1+\|x-y\|_2^2)^{1/2}}$
- Generates a reproducing kernel Hilbert space (RKHS) $\mathcal{K}_k$
- Define the **kernel Stein set** [Gorham and Mackey, 2017]
$$\mathcal{G}_k \triangleq \{g = (g_1, \ldots, g_d) \mid \|v\|^* \leq 1 \text{ for } v_j \triangleq \|g_j\|_{\mathcal{K}_k}\}$$
- Yields closed-form **kernel Stein discrepancy (KSD)**
$$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) = \|w\| \text{ for } w_j \triangleq \sqrt{\sum_{i,i'=1}^n k_0^j(x_i, x_{i'})}.$$
  - Reduces to parallelizable pairwise evaluations of **Stein kernels**
$$k_0^j(x, y) \triangleq \frac{1}{p(x)p(y)} \nabla_{x_j} \nabla_{y_j} (p(x)k(x, y)p(y))$$

Detecting Non-convergence

Goal: Show $(Q_n)_{n\geq 1}$ converges to P whenever $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$

Theorem (Univariate KSD detects non-convergence [Gorham and Mackey, 2017])

Suppose $P \in \mathcal{P}$ and $k(x,y) = \Phi(x-y)$ for $\Phi \in C^2$ with a non-vanishing generalized Fourier transform. If $d = 1$, then $(Q_n)_{n\geq 1}$ converges weakly to P whenever $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$.

- $\mathcal{P}$ is the set of targets P with Lipschitz $\nabla \log p$ and strongly log concave tails ($\frac{\langle \nabla \log(p(x)/p(y)), y-x \rangle}{\|x-y\|_2^2} \geq k$ for $\|x-y\|_2 \geq r$)
  - Includes Bayesian logistic and Student's t regression with Gaussian priors, Gaussian mixtures with common covariance, ...
- Justifies use of KSD with popular Gaussian, Matérn, or inverse multiquadric kernels k **in the univariate case**

Detecting Non-convergence

Goal: Show $(Q_n)_{n\geq 1}$ converges to P whenever $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$

- In higher dimensions, KSDs based on common kernels fail to detect non-convergence, even for Gaussian targets P

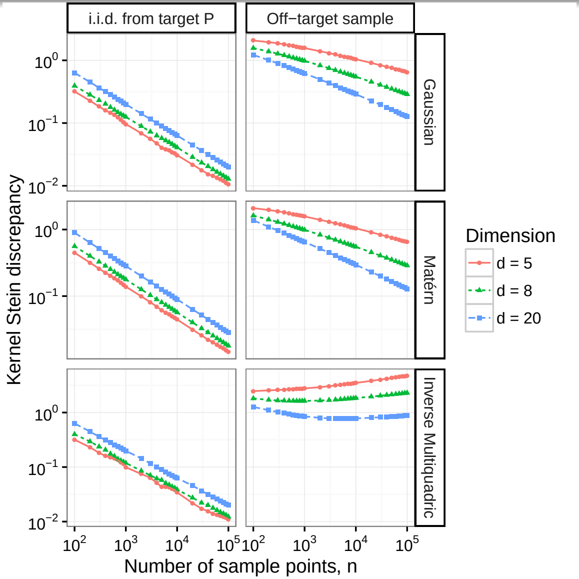
Theorem (KSD fails with light kernel tails [Gorham and Mackey, 2017])

Suppose $d \geq 3$, $P = \mathcal{N}(0, I_d)$, and $\alpha \triangleq (\frac{1}{2} - \frac{1}{d})^{-1}$. If $k(x,y)$ and its derivatives decay at a $o(\|x-y\|_2^{-\alpha})$ rate as $\|x-y\|_2 \to \infty$, then $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ for some $(Q_n)_{n\geq 1}$ not converging to P.

- Gaussian ($k(x,y) = e^{-\frac{1}{2}\|x-y\|_2^2}$) and Matérn kernels fail for $d \geq 3$
- Inverse multiquadric kernels ($k(x,y) = (1 + \|x-y\|_2^2)^{\beta}$) with $\beta < -1$ fail for $d > \frac{2\beta}{1+\beta}$
- The violating sample sequences $(Q_n)_{n\geq 1}$ are simple to construct

Problem: Kernels with light tails ignore excess mass in the tails

The Importance of Kernel Choice



- KSDs with light-tailed kernels fail to detect non-convergence when $d \geq 3$!
- Target $P = \mathcal{N}(0, I_d)$
- Off-target Q_n has all $\|x_i\|_2 \leq 2n^{1/d}\log n$, $\|x_i - x_j\|_2 \geq 2\log n$
- Gaussian and Matérn KSDs driven to 0 by an off-target sequence that does not converge to P
- IMQ KSD does not have this deficiency: $k(x, y) = (1 + \|x - y\|_2^2)^{-\frac{1}{2}}$

Detecting Non-convergence

Goal: Show $(Q_n)_{n\geq 1}$ converges to P whenever $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$

- Consider the inverse multiquadric (IMQ) kernel

$$k(x, y) = (c^2 + \|x - y\|_2^2)^\beta \text{ for some } \beta < 0, c \in \mathbb{R}.$$

- IMQ KSD fails to detect non-convergence when $\beta < -1$
- However, IMQ KSD detects non-convergence when $\beta \in (-1, 0)$

Theorem (IMQ KSD detects non-convergence [Gorham and Mackey, 2017])

Suppose $P \in \mathcal{P}$ and $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for $\beta \in (-1, 0)$. If $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$, then $(Q_n)_{n\geq 1}$ converges weakly to P.

Detecting Convergence

Goal: Show $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ whenever $(Q_n)_{n\geq 1}$ converges to P

Proposition (KSD detects convergence [Gorham and Mackey, 2017])

If $k \in C_b^{(2,2)}$ and $\nabla \log p$ Lipschitz and square integrable under P, then $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ whenever the Wasserstein distance $d_{\mathcal{W}_{\|\cdot\|_2}}(Q_n, P) \to 0$.

- Covers Gaussian, Matérn, IMQ, and other common bounded kernels k

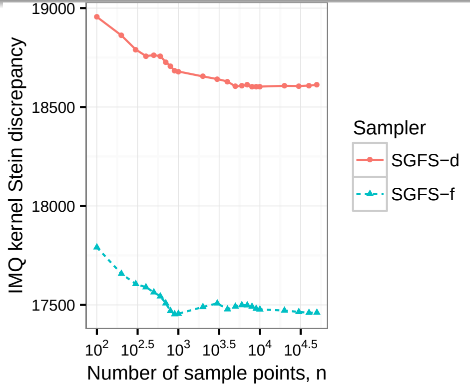
Selecting Samplers

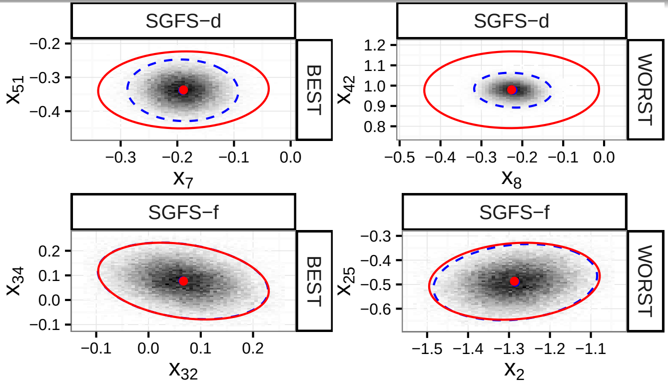
Stochastic Gradient Fisher Scoring (SGFS)

[Ahn, Korattikara, and Welling, 2012]

- Approximate MCMC procedure designed for scalability
  - Approximates Metropolis-adjusted Langevin algorithm but does not use Metropolis-Hastings correction
  - Target P is not stationary distribution
- **Goal:** Choose between two variants
  - SGFS-f inverts a $d \times d$ matrix for each new sample point
  - SGFS-d inverts a diagonal matrix to reduce sampling time
- **MNIST handwritten digits** [Ahn, Korattikara, and Welling, 2012]
  - 10000 images, 51 features, binary label indicating whether image of a 7 or a 9
- Bayesian logistic regression posterior P

Selecting Samplers





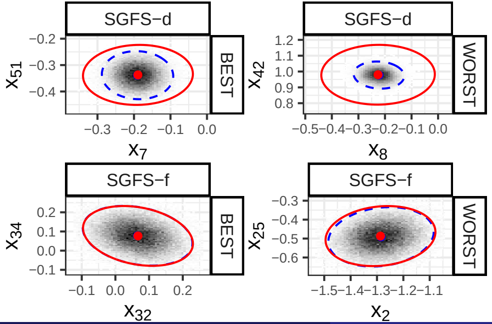
- **Left:** IMQ KSD quality comparison for SGFS Bayesian logistic regression (no surrogate ground truth used)
- **Right:** SGFS sample points (5×10^4) with marginal means and 95% confidence ellipses (blue) that align best / worst with surrogate ground truth sample (red)
- Small speed-up of SGFS-d ($0.0017s$ vs. $0.0019s$) outweighed by loss in accuracy

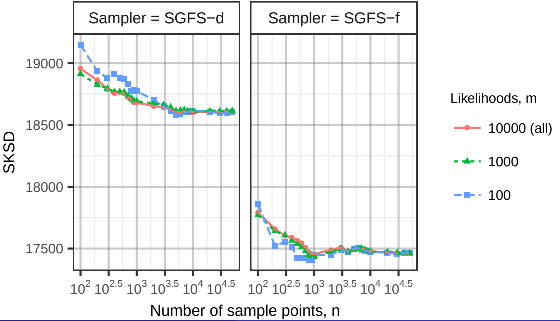
Issue: What if $\nabla \log p$ is too expensive to evaluate?

- Posterior $\nabla \log p(x) = \nabla \log \pi(x) + \sum_{l=1}^{L} \nabla \log \pi(y_l \mid x)$

Solution: Stochastic Stein Discrepancies [Gorham, Raj, and Mackey, 2020]

- Replace each $\nabla \log p(x_i)$ with stochastic gradient based on random datapoint batch: $\nabla \log \pi(x_i) + \frac{L}{|\mathcal{B}_i|} \sum_{l \in \mathcal{B}_i} \nabla \log \pi(y_l | x_i)$
- Resulting stochastic Stein discrepancies inherit convergence control of standard SDs with probability 1 [Gorham, Raj, and Mackey, 2020]





Goodness-of-fit testing

- Chwialkowski, Strathmann, and Gretton [2016] used the KSD $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ to test whether a sample was drawn from a target distribution P (see also Liu, Lee, and Jordan [2016])
- Test with default Gaussian kernel k experienced considerable loss of power as the dimension d increased
- We recreate their experiment with IMQ kernel ($\beta = -\frac{1}{2}, c = 1$)
  - For $n = 500$, generate sample $(x_i)_{i=1}^n$ with $x_i = z_i + u_i\, e_1$ $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and $u_i \overset{\text{iid}}{\sim}$ Unif$[0, 1]$. Target $P = \mathcal{N}(0, I_d)$.
  - Compare with standard normality test of Baringhaus and Henze [1988]

Table: Mean power of multivariate normality tests across 400 simulations

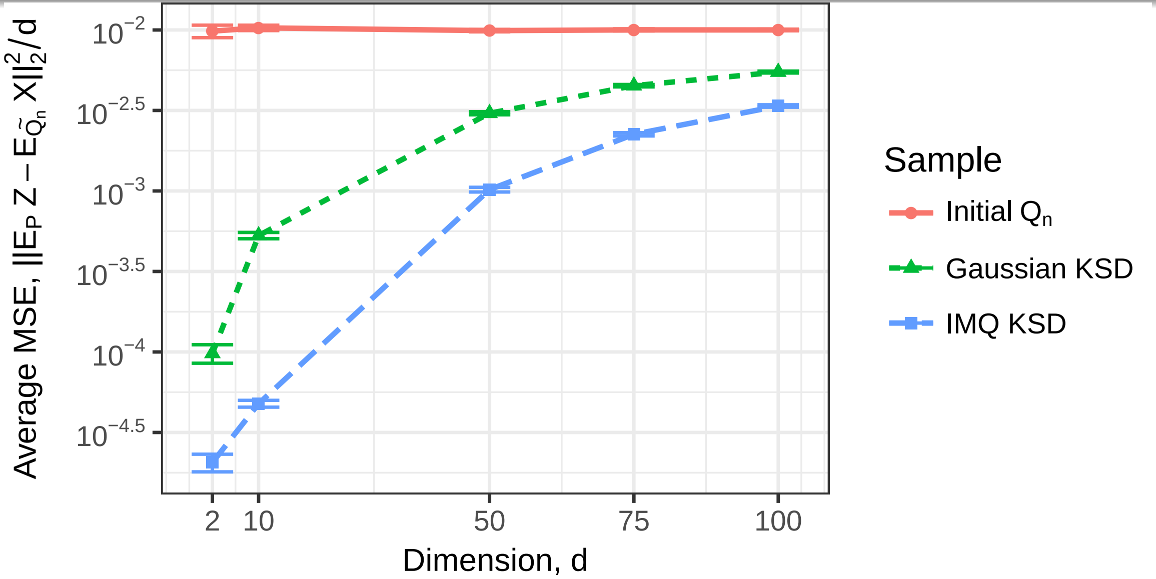
	d=2	d=5	d=10	d=15	d=20	d=25
B&H	1.0	1.0	1.0	0.91	0.57	0.26
Gaussian	1.0	1.0	0.88	0.29	0.12	0.02
IMQ	1.0	1.0	1.0	1.0	1.0	1.0

Beyond Sample Quality Comparison

Improving sample quality

- Given sample points $(x_i)_{i=1}^n$, can minimize KSD $\mathcal{S}(\tilde{Q}_n, \mathcal{T}_P, \mathcal{G}_k)$ over all weighted samples $\tilde{Q}_n = \sum_{i=1}^n q_n(x_i)\delta_{x_i}$ for q_n a probability mass function
- Liu and Lee [2017] do this with Gaussian kernel $k(x,y) = e^{-\frac{1}{h}\|x-y\|_2^2}$
  - Bandwidth h set to median of the squared Euclidean distance between pairs of sample points
- We recreate their experiment with the IMQ kernel $k(x,y) = (1 + \frac{1}{h}\|x-y\|_2^2)^{-1/2}$

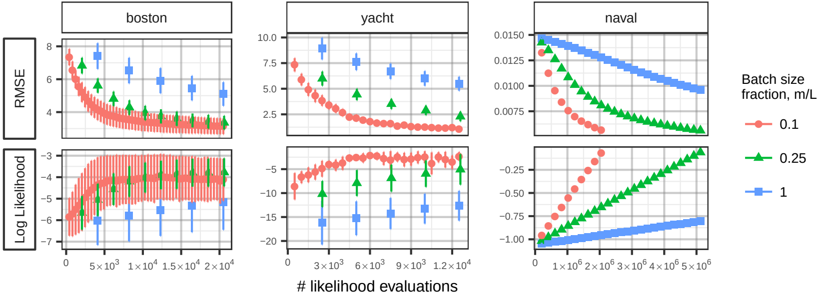
Improving Sample Quality



Generating High-quality Samples

Stein Variational Gradient Descent (SVGD) [Liu and Wang, 2016]

- Uses KSD to repeatedly update locations of n sample points:
  $x_i \leftarrow x_i + \frac{\epsilon}{n} \sum_{l=1}^{n} (k(x_l, x_i) \nabla \log p(x_l) + \nabla_{x_l} k(x_l, x_i))$
  - Approximates gradient step in KL divergence
  - Drives KSD to 0 at $O(1/\sqrt{n})$ rate [Balasubramanian, Banerjee, and Ghosal, 2024]
  - Simple to implement (but each update costs n^2 time)
- **Stochastic SVGD:** uses stochastic KSD $\Rightarrow$ same guarantees with many fewer likelihood evaluations [Gorham, Raj, and Mackey, 2020]



Generating High-quality Samples

Stein Points [Chen, Mackey, Gorham, Briol, and Oates, 2018]

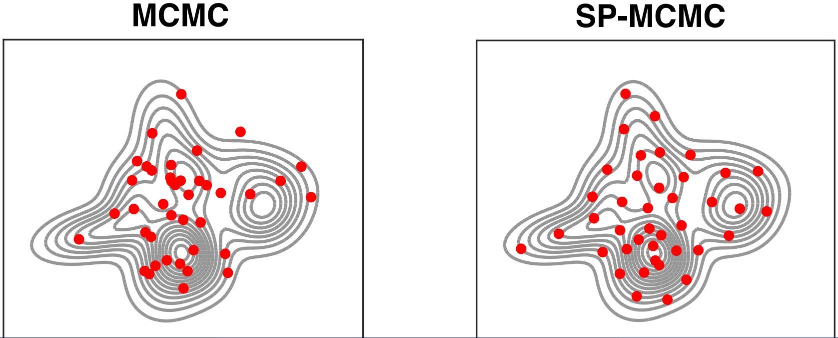
- Greedily minimizes KSD by constructing $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ with

$$x_n \in \operatorname{argmin}_x \mathcal{S}(\tfrac{n-1}{n} Q_{n-1} + \tfrac{1}{n}\delta_x, \mathcal{T}_P, \mathcal{G}_k) = \operatorname{argmin}_x \sum_{j=1}^{d} \frac{k_0^j(x,x)}{2} + \sum_{i=1}^{n-1} k_0^j(x_i, x)$$

  - Sends KSD to zero at $O(\sqrt{\log(n)/n})$ rate

Stein Point MCMC [Chen, Barp, Briol, Gorham, Girolami, Mackey, and Oates, 2019]

- Suffices to optimize over iterates of a Markov chain



Future Directions

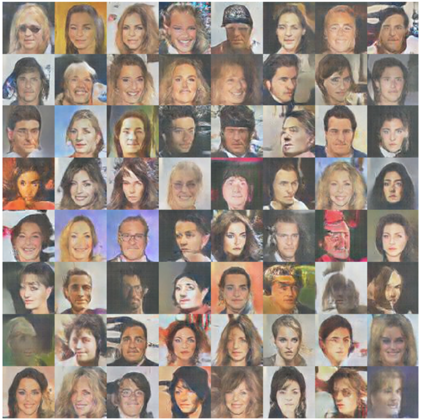
Many opportunities for future development

1. Improving scalability while maintaining convergence control
   - Subsampling of likelihood terms in $\nabla \log p$ [Gorham, Raj, and Mackey, 2020]
   - Linear time, low-rank kernels that distinguish distributions
     - Finite set Stein discrepancies [Jitkrittum, Xu, Szabó, Fukumizu, and Gretton, 2017]
     - Random feature Stein discrepancies [Huggins and Mackey, 2018]
     - **Open question:** When do such discrepancies control convergence?
2. Exploring the impact of Stein operator choice
   - An infinite number of operators $\mathcal{T}$ characterize P
   - **Open questions:** How is discrepancy impacted? How do we select the best $\mathcal{T}$?
   - **Heavy tails:** If $\nabla \log p$ bounded and $k \in C_0^{(1,1)}$, KSD does not control convergence
     - Diffusion Stein operators $(\mathcal{T}g)(x) = \frac{1}{p(x)}\langle \nabla, p(x)a(x)g(x)\rangle$ of Gorham, Duncan, Vollmer, and Mackey [2019] may be appropriate for such heavy-tailed distributions
   - **Isolated modes:** Langevin SDs struggle to detect unexplored modes. Better $\mathcal{T}$?

Future Directions

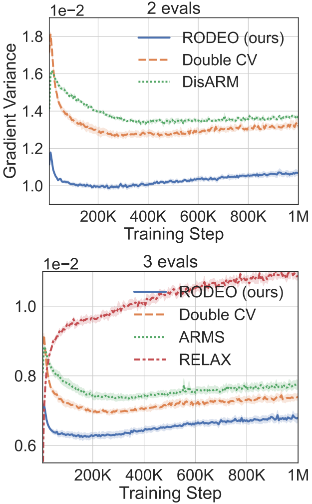
Training generative models

[Wang and Liu, 2016, Pu, Gan, Henao, Li, Han, and Carin, 2017, Shi, Zhou, Hwang, Titsias, and Mackey, 2022b]



DCGAN

SteinGAN

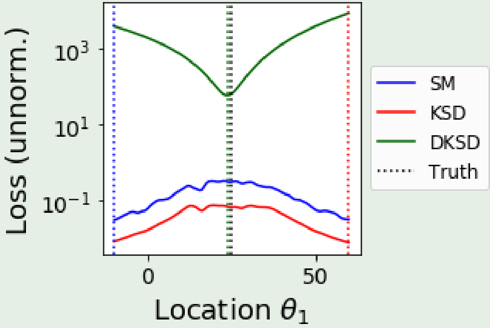


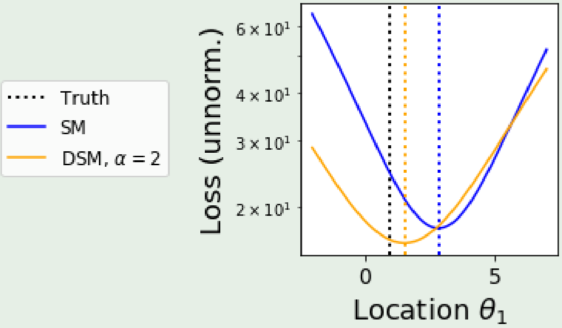
Parameter estimation in unnormalized models

Example (Minimum Stein Discrepancy Estimation [Barp, Briol, Duncan, Girolami, and Mackey, 2019]**)**

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{S}(Q_n, \mathcal{T}_{P_\theta}, \mathcal{G})$$

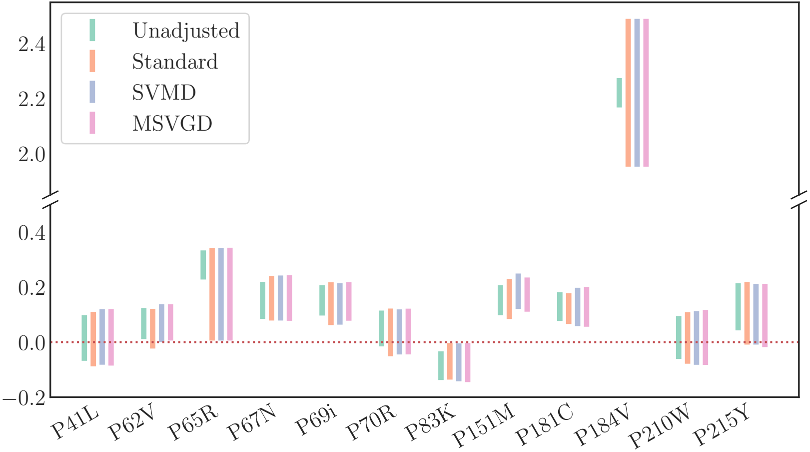
- Unlike maximum likelihood, avoids normalization constant / integration under P_θ!
- Can design diffusion-based discrepancies to deal with heavy tails and outliers





Future Directions

Post-selection inference



- Constrained targets P arise when testing significance after variable selection

  [Tian and Taylor, 2018]

- Stein Variational Mirror Descent and Mirrored SVGD can derive confidence intervals for constrained P

  [Shi, Liu, and Mackey, 2022a]

Future Directions

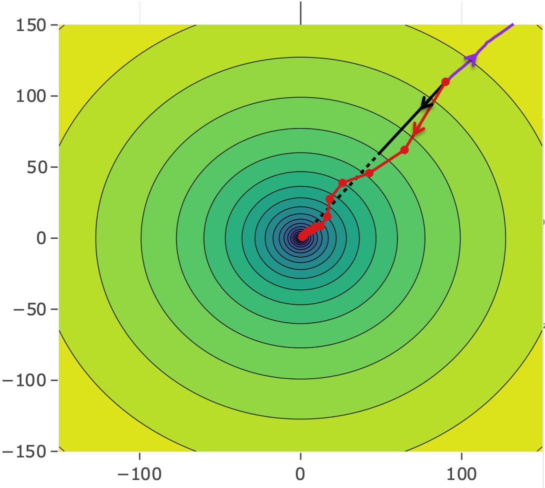
Non-convex optimization

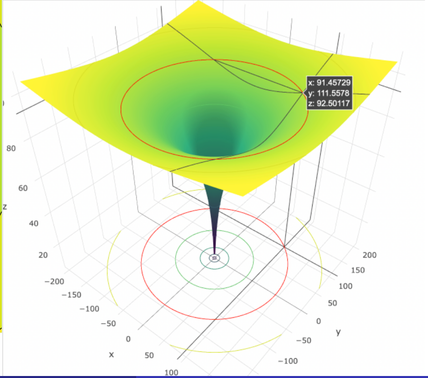
Example (Optimization with Discretized Diffusions [Erdogdu, Mackey, and Shamir, 2018]**)**

- To minimize $f(x)$, choose $a(x) \succcurlyeq cI$ with $a(x)\nabla f(x)$ Lipschitz and distantly dissipative ($\frac{\langle a(x)\nabla f(x) - a(y)\nabla f(y), x-y\rangle}{\|x-y\|_2^2} \geq k$ for $\|x-y\|_2 \geq r$)
- Approximate target sequence $p_n(x) \propto e^{-\gamma_n f(x)}$ using Markov chain $\boldsymbol{x_{n+1} \sim \mathcal{N}(x_n - \frac{\epsilon_n}{2} a(x_n) \nabla f(x_n) + \frac{\epsilon_n}{2\gamma_n}\langle \nabla, a(x_n)\rangle, \frac{\epsilon_n}{\gamma_n} a(x_n))}$
- **Thm:** $\min_{1\leq i\leq n} \mathbb{E} f(x_i) \to \min_x f(x)$ (with explicit error bounds) for appropriate ϵ_n and γ_n when $\nabla f, \nabla a$, and $a^{1/2}$ are Lipschitz

Non-convex optimization [Erdogdu, Mackey, and Shamir, 2018]

$$\min_x f(x) = 5\log(1+\tfrac{1}{2}\|x\|_2^2),\ \ a(x) = (1+\tfrac{1}{2}\|x\|_2^2)I,\ \ a(x)\nabla f(x) = 5x$$



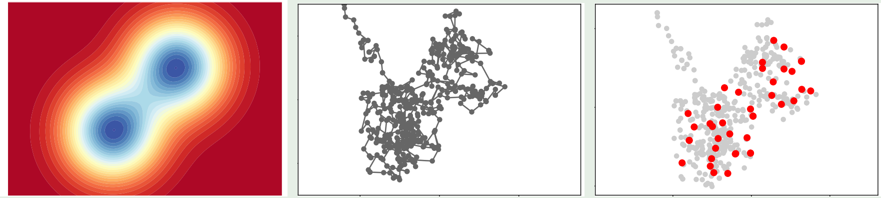


Future Directions

Distribution compression

Example (Stein Kernel Thinning [Li, Dwivedi, and Mackey, 2024])

- **Goal:** Compress sample approximation Q_n to reduce downstream costs
  - e.g., in heart modeling, each sample point can trigger a 1000 CPU hour simulation
- **Stein Thinning:** Greedily minimize KSD using sample points $x_1, \ldots, x_n$



  - **Bonus:** Corrects for biases due to off-target sampling, tempering, approximate MCMC, or burn-in [Riabiz, Chen, Cockayne, Swietach, Niederer, Mackey, and Oates, 2022]
- **Kernel Thinning:** Compress n point summary into $\sqrt{n}$ point summary with comparable KSD [Dwivedi and Mackey, 2024]

Future Directions

Many opportunities for future development

1. Improving scalability while maintaining convergence control
    - Subsampling of likelihood terms in $\nabla \log p$ [Gorham, Raj, and Mackey, 2020]
    - Linear time, low-rank kernels that distinguish distributions [Jitkrittum, Xu, Szabó, Fukumizu, and Gretton, 2017, Huggins and Mackey, 2018]
2. Exploring the impact of Stein operator choice
    - An infinite number of operators $\mathcal{T}$ characterize P
    - How is discrepancy impacted? How do we select the best $\mathcal{T}$?
    - Diffusion Stein operators [Gorham, Duncan, Vollmer, and Mackey, 2019]; Best $\mathcal{T}$ for unexplored modes?
3. Addressing other inferential tasks
    - Generative models [Wang and Liu, 2016, Pu, Gan, Henao, Li, Han, and Carin, 2017, Shi, Zhou, Hwang, Titsias, and Mackey, 2022b]
    - Parameter estimation [Barp, Briol, Duncan, Girolami, and Mackey, 2019]
    - Post-selection inference [Shi, Liu, and Mackey, 2022a]
    - Non-convex optimization [Erdogdu, Mackey, and Shamir, 2018]
    - Distribution compression [Li, Dwivedi, and Mackey, 2024]
    - Control variates [Assaraf and Caffarel, 1999, Mira, Solgi, and Imparato, 2013, Oates, Girolami, and Chopin, 2016, Shi, Zhou, Hwang, Titsias, and Mackey, 2022b]

References I

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proc. 29th ICML*, ICML'12, 2012.

R. Assaraf and M. Caffarel. Zero-variance principle for monte carlo algorithms. *Phys. Rev. Lett.*, 83:4682–4685, Dec 1999. doi: 10.1103/PhysRevLett.83.4682. URL https://link.aps.org/doi/10.1103/PhysRevLett.83.4682.

K. Balasubramanian, S. Banerjee, and P. Ghosal. Improved finite-particle convergence rates for stein variational gradient descent. *arXiv preprint arXiv:2409.08469*, 2024.

A. D. Barbour. Stein's method and Poisson process convergence. *J. Appl. Probab.*, (Special Vol. 25A):175–184, 1988. ISSN 0021-9002. A celebration of applied probability.

A. D. Barbour. Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3):297–322, 1990. ISSN 0178-8051. doi: 10.1007/BF01197887.

L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008. ISSN 1980-0436.

S. Chatterjee and Q. Shao. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. Appl. Probab.*, 21 (2):464–483, 2011. ISSN 1050-5164. doi: 10.1214/10-AAP712.

L. Chen, L. Goldstein, and Q. Shao. *Normal approximation by Stein's method*. Probability and its Applications. Springer, Heidelberg, 2011. ISBN 978-3-642-15006-7. doi: 10.1007/978-3-642-15007-4.

W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. Oates. Stein points. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 844–853, Stockholmsmassan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

W. Y. Chen, A. Barp, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. Oates. Stein point Markov chain Monte Carlo. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1011–1021, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/chen19b.html.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proc. 33rd ICML*, ICML, 2016.

References II

R. Dwivedi and L. Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25(152):1–77, 2024.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global non-convex optimization with discretized diffusions. In *Advances in Neural Information Processing Systems*, pages 9694–9703, 2018.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Adv. NIPS 28*, pages 226–234. Curran Associates, Inc., 2015.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301. PMLR, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, 10 2019. doi: 10.1214/19-AAP1467. URL https://doi.org/10.1214/19-AAP1467.

J. Gorham, A. Raj, and L. Mackey. Stochastic stein discrepancies. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17931–17942. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf.

F. Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2):724–739, 1991.

J. Huggins and L. Mackey. Random feature stein discrepancies. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1903–1913. Curran Associates, Inc., 2018.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A Linear-Time Kernel Goodness-of-Fit Test. In *Advances in Neural Information Processing Systems*, 2017.

A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proc. of 31st ICML*, ICML'14, 2014.

L. Li, R. Dwivedi, and L. Mackey. Debiased distribution compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 203 of *Proceedings of Machine Learning Research*. PMLR, 21–27 Jul 2024.

Q. Liu and J. Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

References III

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proc. of 33rd ICML*, volume 48 of *ICML*, pages 276–284, 2016.

L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.*, 21:14 pp., 2016. doi: 10.1214/16-ECP15.

E. Meckes. On Stein's method for multivariate normal approximation. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, pages 153–178. Inst. Math. Statist., Beachwood, OH, 2009. doi: 10.1214/09-IMSCOLL511.

A. Mira, R. Solgi, and D. Imparato. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

A. Müller. Integral probability metrics and their generating classes of functions. *Ann. Appl. Probab.*, 29(2):pp. 429–443, 1997.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12185.

Y. Pu, Z. Gan, R. Henao, C. Li, S. Han, and L. Carin. Vae learning via stein variational gradient descent. In *Advances in Neural Information Processing Systems*, pages 4237–4246, 2017.

G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.*, 37(6): 2150–2173, 2009. ISSN 0091-1798. doi: 10.1214/09-AOP467.

M. Riabiz, W. Y. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal thinning of mcmc output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a), 2022. doi: https://doi.org/10.1111/rssb.12503. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12503.

J. Shi, C. Liu, and L. Mackey. Sampling with mirrored stein operators. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/pdf?id=eMudnJsb1T5.

J. Shi, Y. Zhou, J. Hwang, M. Titsias, and L. Mackey. Gradient estimation with discrete stein operators. *Advances in neural information processing systems*, 35: 25829–25841, 2022b.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602. Univ. California Press, Berkeley, Calif., 1972.

C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 1–26. Inst. Math. Statist., Beachwood, OH, 2004.

X. Tian and J. Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

D. Wang and Q. Liu. Learning to Draw Samples: With Application to Amortized MLE for Generative Adversarial Learning. *arXiv:1611.01722*, Nov. 2016.

M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

Selecting Sampler Hyperparameters

Setup [Welling and Teh, 2011]

- Consider the posterior distribution P induced by L datapoints y_l drawn i.i.d. from a Gaussian mixture likelihood

$$Y_l | X \overset{\text{iid}}{\sim} \tfrac{1}{2}\mathcal{N}(X_1, 2) + \tfrac{1}{2}\mathcal{N}(X_1 + X_2, 2)$$

under Gaussian priors on the parameters $X \in \mathbb{R}^2$

$$X_1 \sim \mathcal{N}(0, 10) \perp\!\!\!\perp X_2 \sim \mathcal{N}(0, 1)$$

  - Draw $m = 100$ datapoints y_l with parameters $(x_1, x_2) = (0, 1)$
  - Induces posterior with second mode at $(x_1, x_2) = (1, -1)$
- For range of parameters ϵ, run approximate SGLD for 1000 steps and store resulting posterior sample Q_n
- Use minimum GSD to select appropriate ϵ
  - Compare with standard MCMC parameter selection criterion, effective sample size (ESS), a measure of Markov chain autocorrelation
  - Compute median of diagnostic over 50 random sequences

Selecting Samplers

Setup

- **MNIST handwritten digits** [Ahn, Korattikara, and Welling, 2012]
  - 10000 images, 51 features, binary label indicating whether image of a 7 or a 9
- Bayesian logistic regression posterior P
  - L independent observations $(y_l, v_l) \in \{1, -1\} \times \mathbb{R}^d$ with

$$\mathbb{P}(Y_l = 1 | v_l, X) = 1/(1 + \exp(-\langle v_l, X\rangle))$$

  - Flat improper prior on the parameters $X \in \mathbb{R}^d$
- Use IMQ KSD ($\beta = -\frac{1}{2}, c = 1$) to compare SGFS-f to SGFS-d drawing 10^5 sample points and discarding first half as burn-in
- For external support, compare bivariate marginal means and 95% confidence ellipses with surrogate ground truth Hamiltonian Monte chain with 10^5 sample points [Ahn, Korattikara, and Welling, 2012]

The Importance of Tightness

Goal: Show $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ only if Q_n converges to P

- A sequence $(Q_n)_{n \geq 1}$ is **uniformly tight** if for every $\epsilon > 0$, there is a finite number $R(\epsilon)$ such that $\sup_n Q_n(\|X\|_2 > R(\epsilon)) \leq \epsilon$
  - Intuitively, no mass in the sequence escapes to infinity

Theorem (KSD detects tight non-convergence [Gorham and Mackey, 2017])

Suppose that $P \in \mathcal{P}$ and $k(x, y) = \Phi(x - y)$ for $\Phi \in C^2$ with a non-vanishing generalized Fourier transform. If $(Q_n)_{n \geq 1}$ is uniformly tight and $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$, then $(Q_n)_{n \geq 1}$ converges weakly to P.

- Good news, but, ideally, KSD would detect non-tight sequences automatically...