Measuring Sample Quality with Stein's Method

Lester Mackey*

Joint work with Jackson Gorham†, Andrew Duncan‡, Sebastian Vollmer**

Microsoft Research*, Opendoor Labs†, University of Sussex‡, University of Warwick**

July 30, 2018

Motivation: Large-scale Posterior Inference

Example: Bayesian logistic regression

1. Unknown parameter vector: $\beta \sim \mathcal{N}(0, I)$
2. Fixed covariate vector: $v_l \in \mathbb{R}^d$ for each datapoint $l = 1, \ldots, L$
3. Binary class label: $Y_l \mid v_l, \beta \overset{\text{ind}}{\sim} \text{Ber}\left(\frac{1}{1+e^{-\langle \beta, v_l \rangle}}\right)$

- 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 $\mathbb{E}_P[h(Z)] = \int_{\mathcal{X}} p(x)h(x)dx$ with asymptotically exact sample estimates $\mathbb{E}_{Q_n}[h(X)] = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$
- **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?

- **MCMC Benefit:** Approximates intractable posterior expectations $\mathbb{E}_P[h(Z)] = \int_{\mathcal{X}} p(x)h(x)dx$ with asymptotically exact sample estimates $\mathbb{E}_{Q_n}[h(X)] = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$
- **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 measure 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 $\mathcal{X} = \mathbb{R}^d$ (will relax to any convex set) and density p
  - p known up to normalization, integration under P is intractable
- **Sample points** $x_1, \dots, x_n \in \mathcal{X}$
  - 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)

Examples

- Total variation distance ($\mathcal{H} = \{h : \sup_x |h(x)| \leq 1\}$)
- Wasserstein (or Kantorovich-Rubenstein) distance, $d_{\mathcal{W}_{\|\cdot\|}}$ ($\mathcal{H} = \mathcal{W}_{\|\cdot\|} \triangleq \{h : \sup_{x \neq y} \frac{|h(x) - h(y)|}{\|x - y\|} \leq 1\}$)

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 (Requirements I and II)?
- 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 $\boldsymbol{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})}$ by reference IPM $\boldsymbol{d_{\mathcal{H}}(Q_n, P)}$**
$\Rightarrow \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}) \to 0$ only if $(Q_n)_{n \geq 1}$ converges to P (Req. II)
    - Performed once, in advance, for large classes of distributions

3. **Upper bound $\boldsymbol{\mathcal{S}(Q_n, \mathcal{T}, \mathcal{G})}$ 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, 2016]

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 $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$

Computing Stein Discrepancies

Question: How do we compute a Stein discrepancy $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) = \sup_{g \in \mathcal{G}} |\mathbb{E}_{Q_n}[(\mathcal{T}_P g)(X)]|$ in practice?

Consider the classical Stein discrepancy optimization problem

$$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) = \sup_g \frac{1}{n} \sum_{i=1}^n \langle g(x_i), \nabla \log p(x_i) \rangle + \langle \nabla, g(x_i) \rangle$$

$$\begin{aligned} \text{s.t. } & \|g(x)\|^* \leq 1, \forall x \in \mathcal{X} \\ & \|\nabla g(x)\|^* \leq 1, \forall x \in \mathcal{X} \\ & \|\nabla g(x) - \nabla g(y)\|^* \leq \|x - y\|, \forall x, y \in \mathcal{X} \end{aligned}$$

- Objective only depends on the values of g and ∇g at the n sample points x_i
- Infinite-dimensional problem with infinitude of constraints

Idea: Find alternative Stein set $\mathcal{G}$ with equivalent convergence properties and only finitely many constraints

Graph Stein Discrepancies

For any graph $G = (V, E)$ with vertices $V = \{x_1, \ldots, x_n\}$, define **graph Stein set** $\mathcal{G}_{\|\cdot\|,Q_n,G}$ of functions $g : \mathcal{X} \to \mathbb{R}^d$ with

- Boundedness constraints imposed only at points x_i
- Smoothness constraints imposed only between pairs $(x_i, x_k) \in E$
- **Benefit:** Optimization problem has order $|V| + |E|$ constraints

Proposition (Equivalence of Classical & Complete Graph Stein Discrepancies)

If $\mathcal{X} = \mathbb{R}^d$, and G_1 is the complete graph on $\{x_1, \ldots, x_n\}$, then

$$\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \le \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q_n,G_1}) \le \kappa_d \mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|})$$

for $\kappa_d > 0$ depending only on the dimension d and the norm $\|\cdot\|$.

- Follows from Whitney-Glaeser extension theorem [Glaeser, 1958]
- $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q_n,G_1})$ inherits convergence properties of classical
- **Problem:** Complete graph introduces order n^2 constraints!

Spanner Stein Discrepancies

Goal: Find equivalent Stein discrepancy with only $O(n)$ constraints

Approach: Geometric spanners [Chew, 1986, Peleg and Schäffer, 1989]

- For a **dilation factor** $t \geq 1$, a t-**spanner** $G = (V, E)$ has
  - The weight $\|x - y\|$ on each edge $(x, y) \in E$
  - Path with total weight $\leq t\|x - y\|$ between each $(x, y) \in V^2$

Proposition (Equivalence of Spanner and Complete Graph Stein Discrepancies)

If $\mathcal{X} = \mathbb{R}^d$, G_1 is the complete graph on $\{x_1, \dots, x_n\}$, and G_t is a t-spanner on $\{x_1, \dots, x_n\}$, then

$$1 \leq \frac{\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|, Q_n, G_t})}{\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|, Q_n, G_1})} \leq 2t^2.$$

- For $t = 2$, can compute spanner with $O(\kappa_d n)$ edges in $O(\kappa_d n \log(n))$ expected time [Har-Peled and Mendel, 2006]
- Fix $t = 2$ and use efficient greedy spanner implementation of Bouts, ten Brink, and Buchin [2014] in our experiments

Decoupled Linear Programs

Norm recommendation: $\|\cdot\| = \|\cdot\|_1$

- Optimization problem decouples across components g_j
  - Can solve d subproblems in parallel
- Each subproblem is a linear program

Recommended spanner Stein discrepancy algorithm

- Compute 2-spanner G_2 on $V = \{x_1, \ldots, x_n\}$
- Solve d finite-dimensional linear programs in parallel

$$\sum_{j=1}^{d} \sup_{\gamma_j \in \mathbb{R}^n, \Gamma_j \in \mathbb{R}^{d \times n}} \frac{1}{n} \sum_{i=1}^{n} \gamma_{ji} \nabla_j \log p(x_i) + \Gamma_{jji}$$

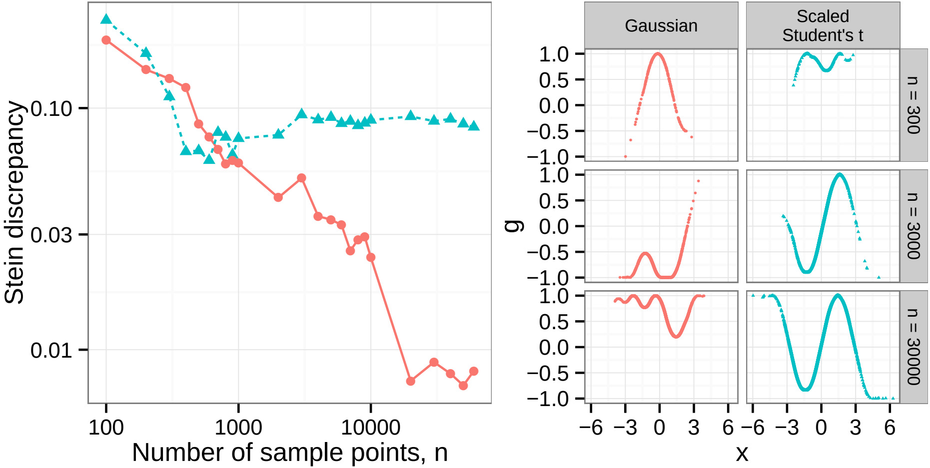
$$\text{s.t. } \|\gamma_j\|_\infty \leq 1, \|\Gamma_j\|_\infty \leq 1, \text{ and } \forall\, i \neq l : (x_i, x_l) \in E,$$

$$\max\left(\frac{|\gamma_{ji} - \gamma_{jl}|}{\|x_i - x_l\|_1}, \frac{\|\Gamma_j(e_i - e_l)\|_\infty}{\|x_i - x_l\|_1}\right) \leq 1,$$

$$\max\left(\frac{|\gamma_{ji} - \gamma_{jl} - \langle \Gamma_j e_i, x_i - x_l \rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}, \frac{|\gamma_{ji} - \gamma_{jl} - \langle \Gamma_j e_l, x_i - x_l \rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}\right) \leq 1.$$

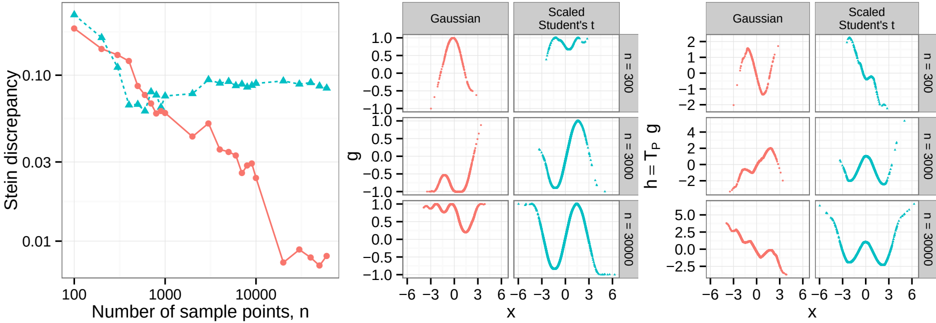
  - Here $\gamma_{ji} = g_j(x_i)$ and $\Gamma_{jki} = \nabla_k g_j(x_i)$

A Simple Example



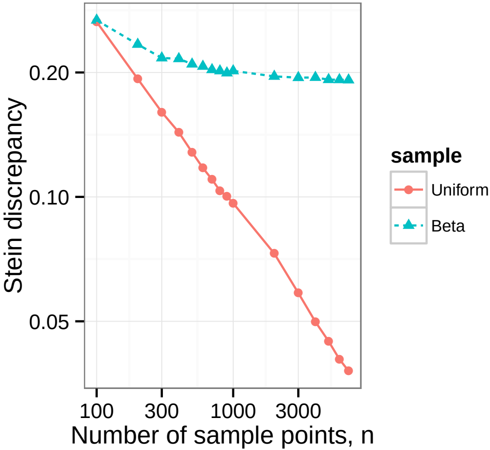
- For target $P = \mathcal{N}(0,1)$, compare i.i.d. $\mathcal{N}(0,1)$ sample Q_n to scaled Student's t sample Q'_n with matching variance
- Expect $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q_n,G_1}) \to 0$ & $\mathcal{S}(Q'_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q_n,G_1}) \not\to 0$

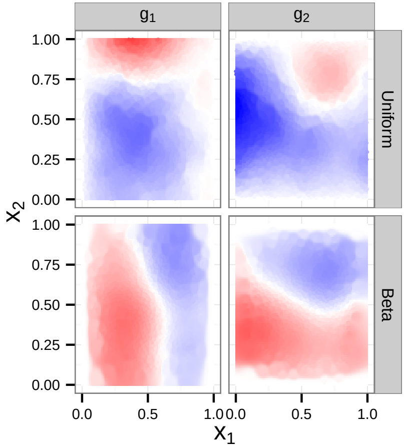
A Simple Example



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

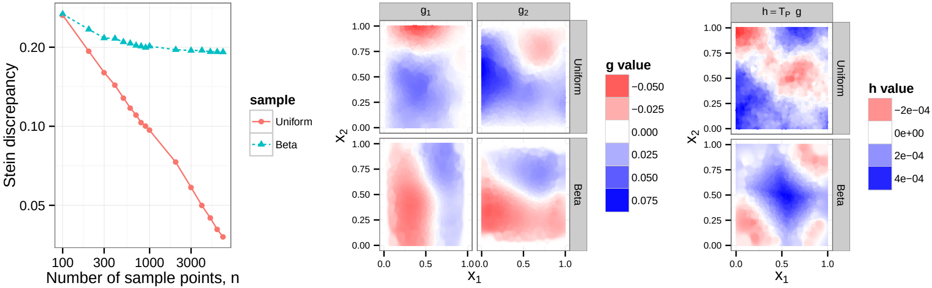
A Simple Constrained Example





- For two-dimensional target $P = \text{Unif}(0,1) \times \text{Unif}(0,1)$, compare i.i.d. $\text{Unif}(0,1) \times \text{Unif}(0,1)$ sample Q_n to i.i.d. $\text{Beta}(3,3) \times \text{Beta}(3,3)$ sample Q'_n

A Simple Constrained Example



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

Comparing Discrepancies

Setup

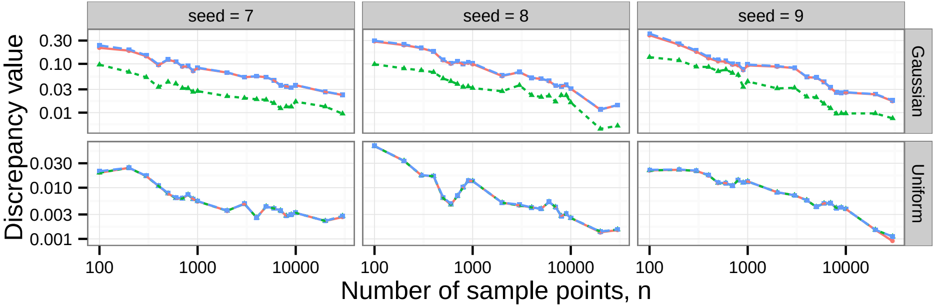
- Draw $n = 30,000$ points i.i.d. from $\mathcal{N}(0,1)$ or Unif$[0,1]$
  - Yields sample Q_n
- Compare behavior of classical and graph Stein discrepancy
  - When $d = 1$ classical Stein discrepancy solves finite-dimensional convex quadratically constrained quadratic program with $O(n)$ variables, $O(n)$ constraints, and linear objective [Gorham and Mackey, 2015]
- Compare to Wasserstein distance

$$d_{\mathcal{W}_{\|\cdot\|}}(Q_n, P) = \int_{\mathbb{R}} |Q_n(t) - P(t)| dt$$

  - Can adjust smoothness constants (**Stein factors**) so that Stein discrepancies directly lower bounded by Wasserstein distance
  - For uniform target, classical Stein discrepancy equals Wasserstein distance

Comparing Discrepancies

Orange = Classical Stein, Blue = Graph Stein, Green = Wasserstein



Selecting Sampler Hyperparameters

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

- Prior $\pi(x)$, Likelihood $\pi(y \mid x)$

Stochastic Gradient Langevin Dynamics (SGLD)

[Welling and Teh, 2011]

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

- Approximate MCMC procedure designed for scalability
  - Approximates Metropolis-adjusted Langevin algorithm and continuous-time Langevin diffusion
  - Random subset $\mathcal{B}_k$ of datapoints used to select each sample
  - No Metropolis-Hastings correction step
  - Target P is not stationary distribution
- Choice of step size ϵ critical for accurate inference
  - Too small $\Rightarrow$ slow mixing
  - Too large $\Rightarrow$ sampling from very different distribution
  - Standard MCMC selection criteria like **effective sample size** (ESS) and asymptotic variance do not account for this bias

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

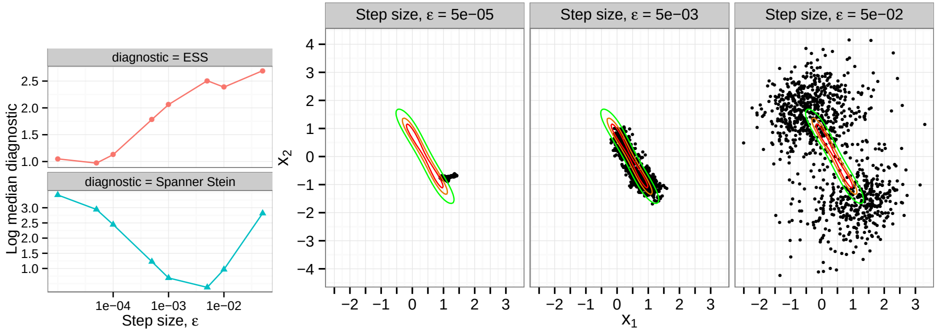
$$Y_l | X \overset{\text{iid}}{\sim} \tfrac{1}{2}\mathcal{N}(X_1, 2) + \tfrac{1}{2}\mathcal{N}(X_1 + X_2, 2)$$

  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 step sizes ϵ, use SGLD with batch size 10 to draw approximate posterior sample Q_n of size $n = 1000$
- Use minimum Stein discrepancy 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 SGLD sequences

Selecting Sampler Hyperparameters



- ESS maximized at step size $\epsilon = 5 \times 10^{-2}$
- Stein discrepancy minimized at step size $\epsilon = 5 \times 10^{-3}$
- **Right:** ESS: $2.6, 12.3, 14.8$; Stein discrepancies: $19.0, 1.5, 16.7$

Quantifying a Bias-Variance Trade-off

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

- Prior $\pi(x)$, Likelihood $\pi(y \mid x)$

Approximate Random Walk Metropolis-Hastings (ARWMH)

[Korattikara, Chen, and Welling, 2014]

- Approximate MCMC procedure designed for scalability
  - Uses Gaussian random walk proposals: $x_{k+1} \sim \mathcal{N}(x_k, \sigma^2 I)$
  - Approximates Metropolis-Hastings correction using random subset of datapoints to accept or reject proposal
    - Exact MH accepts w.p. $\min\left(1, \frac{\pi(x_{k+1}) \prod_{l=1}^{L} \pi(y_l|x_{k+1})}{\pi(x_k) \prod_{l=1}^{L} \pi(y_l|x_k)}\right)$
- Tolerance parameter ϵ controls number of datapoints considered
  - Larger $\epsilon \Rightarrow$ fewer datapoints considered, fewer likelihood computations, more rapid sampling, more rapid variance reduction
  - Smaller $\epsilon \Rightarrow$ closer approximation to true MH correction, less bias in stationary distribution

Question: Can we quantify this "bias-variance" trade-off explicitly?

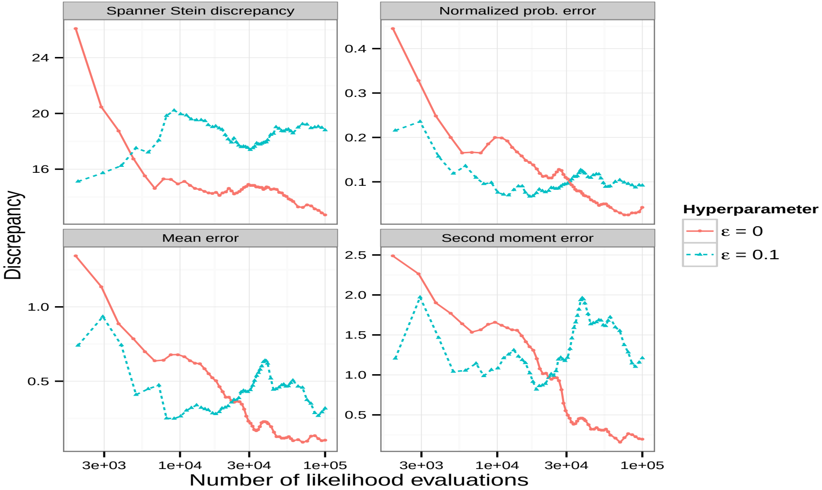
Quantifying a Bias-Variance Trade-off

Setup

- **Nodal dataset** [Canty and Ripley, 2015]
  - 53 patients, 6 predictors, binary response indicating whether cancer spread from prostate to lymph nodes
- Bayesian logistic regression posterior P
  - L independent observations $(y_l, v_l) \in \{1, -1\} \times \mathbb{R}^d$ with

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

  - Gaussian prior on the parameters $X \in \mathbb{R}^d$: $X \sim \mathcal{N}(0, I)$
- Compare ARWMH ($\epsilon = 0.1$ and batch size 2) to exact RWMH
  - Ran each chain until 10^5 likelihood evaluations computed
  - Computed spanner Stein discrepancy after burn-in of 10^3 likelihood computations and thinning down to 1,000 samples
  - Expect ARWMH quality as a function of likelihood evaluations to dominate initially and RWMH quality to overtake eventually
- For external support, also compute deviation between various expectations under Q_n and under a MALA chain with 10^7



- Non-Stein measures based on additional, long-running chain used as surrogate for the target distribution
- Stein discrepancy computed from sample Q_n alone

Assessing Convergence Rates

An observation

- The approximating distribution Q_n in $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|,Q_n,G})$ need not be based on a *random* sample
- Stein discrepancy meaningful even for *deterministic* pseudosamples (e.g., from **quasi-Monte Carlo** or **herding**)

Independent sampling

- $\mathbb{E}[|\mathbb{E}_{Q_n}[h(X)] - \mathbb{E}_P[h(Z)]|] = O(1/\sqrt{n})$ for bounded variance h

Sobol sequence [Sobol, 1967]

- $d_{\mathcal{H}}(Q_n, P) = O(\log^{d-1}(n)/n)$ for bounded total variation h

Kernel herding [Chen, Welling, and Smola, 2010]

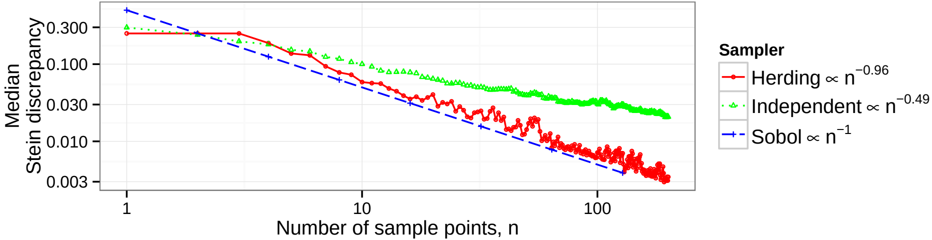
- $d_{\mathcal{H}}(Q_n, P) = O(1/n)$ for finite-dimensional Hilbert space $\mathcal{H}$
- $d_{\mathcal{H}}(Q_n, P) = O(1/\sqrt{n})$ for infinite-dimensional Hilbert space $\mathcal{H}$
  - Rate often better in practice (without theoretical explanation)

Assessing Convergence Rates

Setup [Bach, Lacoste-Julien, and Obozinski, 2012]

- Target $P = \text{Unif}[0, 1]$
- Draw $n = 200$ points
  - i.i.d. from Unif$[0, 1]$ (repeated 50 times)
  - From a Sobol sequence
  - From a Herding sequence with Hilbert space $\mathcal{H}$ defined by the norm $\|h\|_{\mathcal{H}} = \int_0^1 (h'(x))^2 dx$
- Compare median Stein discrepancy decay across three samplers
- Assess convergence rate with best fit line to log-log plot

Assessing Convergence Rates



- Stein discrepancy convergence for **deterministic sequences**, kernel herding [Chen, Welling, and Smola, 2010] and Sobol [Sobol, 1967], versus i.i.d. sample sequence for $P = \text{Unif}(0, 1)$
- Estimated rates for i.i.d. and Sobol accord with expected $O(1/\sqrt{n})$ and $O(1/n)$ rates from literature
- Herding rate outpaces its best known $O(1/\sqrt{n})$ bound [Bach, Lacoste-Julien, and Obozinski, 2012]: opportunity for sharper analysis?

Future Directions

Many opportunities for future development

1. Developing tailored Stein program solvers that exploit problem structure for greater scalability
   - LP constraint matrices are very sparse and, at times, banded
   - Leverage stochastic optimization to avoid expensive summations in Stein program objective
     - e.g., $\nabla \log p(x_i) = \nabla \log \pi(x_i) + \sum_{l=1}^{L} \nabla \log \pi(y_l \mid x_i)$
   - Improve scalability with first order methods?
2. Establishing reference IPM lower bounds for Stein discrepancy
   - For what other families of distributions P does $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ imply $d_{\mathcal{W}_{\|\cdot\|}}(Q_n, P) \to 0$?
3. 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 $\mathcal{T}$?
4. Addressing other inferential tasks
   - Design of control variates [Oates, Girolami, and Chopin, 2014, Oates and Girolami, 2015]
   - One-sample testing [Chwialkowski, Strathmann, and Gretton, 2016, Liu, Lee, and Jordan, 2016]

References I

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

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proc. 29th ICML*, ICML'12, 2012.

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.

Q. W. Bouts, A. P. ten Brink, and K. Buchin. A framework for Computing the Greedy Spanner. In *Proc. of 30th SOCG*, pages 11:11–11:19, New York, NY, 2014. ACM.

A. Canty and B. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2015. R package version 1.3-15.

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.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *UAI*, 2010.

P. Chew. There is a Planar Graph Almost As Good As the Complete Graph. In *Proc. 2nd SOCG*, pages 169–177, New York, NY, 1986. ACM.

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

G. Glaeser. Étude de quelques algèbres tayloriennes. *J. Analyse Math.*, 6:1–124; erratum, insert to 6 (1958), no. 2, 1958.

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, A. Duncan, S. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *arXiv:1611.06972*, Nov. 2016.

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

References II

S. Har-Peled and M. Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.*, 35 (5):1148–1184, 2006.

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

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. *arXiv:1512.07392*, 2015.

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

C. Oates and M. Girolami. Control functionals for Quasi-Monte Carlo integration. *arXiv:1501.03379*, 2015.

C. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *arXiv:1410.2392*, Oct. 2014. To appear in JRSS, Series B.

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

D. Peleg and A. Schäffer. Graph spanners. *J. Graph Theory*, 13(1):99–116, 1989.

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.

I. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. and Math. Phys*, (7):86–112, 1967.

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.

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