Measuring Sample Quality with Stein’s Method

Lester Mackey

Joint work with Jackson Gorham

Stanford University

January 7, 2016
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: \( Y_l \mid v_l, \beta \sim \text{Ber}\left(\frac{1}{1+e^{-\langle \beta, v_l \rangle}}\right) \)

- 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
\[
\mathbb{E}_P[h(Z)] = \int_X p(x) h(x) dx \text{ with asymptotically exact sample estimates } \mathbb{E}_Q[h(X)] = \sum_{i=1}^n q(x_i) h(x_i)
\]

**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 \( \mathbb{E}_P[h(Z)] = \int_X p(x)h(x)dx \) with asymptotically exact sample estimates \( \mathbb{E}_Q[h(X)] = \sum_{i=1}^{n} q(x_i)h(x_i) \)

- **Problem:** Each point \( x_i \) requires iterating over entire dataset!

**Template solution:** Approximate MCMC with subset posteriors


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


- 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
- **Weighted sample** consisting of distinct sample points $x_1, \ldots, x_n \in \mathcal{X}$ and weights $q(x_i)$ encoded in a p.m.f. $q$
  - Defines **discrete distribution** $Q$ with, for any function $h$,
    \[ \mathbb{E}_Q[h(X)] = \sum_{i=1}^n q(x_i)h(x_i) \] used to approximate $\mathbb{E}_P[h(Z)]$
  - We make no assumption about the provenance of the $x_i$

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

1. Detects when a sample sequence is converging to the target
2. Detects when a sample sequence is not converging to the target
3. Is computationally feasible
Integral Probability Metrics

**Goal:** Quantify how well $E_Q$ approximates $E_P$

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

$$d_{\mathcal{H}}(Q, P) = \sup_{h \in \mathcal{H}} |E_Q[h(X)] - E_P[h(Z)]|$$

- 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 $d_{\mathcal{H}}(Q_m, P)$ to zero implies $(Q_m)_{m \geq 1}$ 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 $\mathbb{E}_Q$ approximates $\mathbb{E}_P$

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

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

- 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 $d_{\mathcal{H}}(Q_m, P)$ to zero implies $(Q_m)_{m \geq 1}$ converges weakly to $P$ (*Requirement II*)

**Problem:** Integration under $P$ intractable!

⇒ 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 $Q$!
- 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 for bounding IPMs [Stein, 1972] proceeds in 3 steps:

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

2. **Lower bound** $S(Q, \mathcal{T}, \mathcal{G})$ by reference IPM $d_{\mathcal{H}}(Q, P)$
   \[
   \Rightarrow S(Q_m, \mathcal{T}, \mathcal{G}) \text{ converges to 0 only if } d_{\mathcal{H}}(Q_m, P) \text{ does (Req. II)}
   \]
   - Performed once, in advance, for large classes of distributions

3. **Upper bound** $S(Q, \mathcal{T}, \mathcal{G})$ 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 Characterizing 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]
- Identify a Markov process $(Z_t)_{t \geq 0}$ with stationary distribution $P$
- Under mild conditions, its **infinitesimal generator**
  
  $$(\mathcal{A}u)(x) = \lim_{t \to 0} \frac{\mathbb{E}[u(Z_t) | Z_0 = x] - u(x)}{t}$$

  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:**
  
  $$(\mathcal{T}_P g)(x) \triangleq \langle g(x), \nabla \log p(x) \rangle + \langle \nabla, g(x) \rangle$$

  - 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

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

Goal: Lower bound classical Stein discrepancy \( S(Q, T_P, \mathcal{G}_{\|\cdot\|}) \) by reference IPM \( d_H(Q, P) \)

- In the univariate case \((d = 1)\), known that for many targets \( P \), \( S(Q_m, T_P, \mathcal{G}_{\|\cdot\|}) \to 0 \) only if Wasserstein \( d_{W_{\|\cdot\|}}(Q_m, P) \to 0 \) [Chen, Goldstein, and Shao, 2010a, Chatterjee 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 and Mackey, 2015]

Theorem (Lower Bound for Strongly Log-concave Densities)

*If \( \mathcal{X} = \mathbb{R}^d \) and \( \log p \in C^4 \) is strongly concave with bounded 3rd and 4th derivatives then \( S(Q_m, T_P, \mathcal{G}_{\|\cdot\|}) \to 0 \) \( \Rightarrow \) \( d_{W_{\|\cdot\|}}(Q_m, P) \to 0 \).*

- Example: Bayesian logistic regression with Gaussian priors
- Conditions sufficient, not necessary: Provides template for lower bounding \( S(Q, T_P, \mathcal{G}_{\|\cdot\|}) \) for other large classes of distributions
Question: Under what conditions on a sample sequence \((Q_m)_{m \geq 1}\) will Stein discrepancy \(S(Q_m, T_P, \mathcal{G}_{\|\cdot\|}) \to 0\)?

Proposition (Convergence of Stein Discrepancy [Gorham and Mackey, 2015])

If \(X \sim Q\) and \(Z \sim P\) with \(\nabla \log p(Z)\) integrable, then
\[
S(Q, T_P, \mathcal{G}_{\|\cdot\|}) \leq \mathbb{E}[\|X - Z\|] + \mathbb{E}[\|\nabla \log p(X) - \nabla \log p(Z)\|] + \mathbb{E}[\|\nabla \log p(Z)(X - Z)^\top\|] \\
\leq \mathbb{E}[\|X - Z\|] + \mathbb{E}[\|\nabla \log p(X) - \nabla \log p(Z)\|] + \sqrt{\mathbb{E}[\|\nabla \log p(Z)\|^2] \mathbb{E}[\|X - Z\|^2]}.
\]

• One take-away: If \(\|\nabla \log p(Z)\|\) square-integrable, then Stein discrepancy converges whenever \(X_m \sim Q_m\) converges in mean square and \(\nabla \log p(X_m)\) converges in mean.
Computing Stein Discrepancies

**Question:** How do we compute a Stein discrepancy
\[ S(Q, T_P, G) = \sup_{g \in G} |\mathbb{E}_Q[(T_P g)(X)]| \] in practice?

Consider the classical Stein discrepancy optimization problem
\[ S(Q, T_P, G_{\|\cdot\|}) = \sup g \sum_{i=1}^{n} q(x_i)(\langle g(x_i), \nabla \log p(x_i) \rangle + \langle \nabla, g(x_i) \rangle) \]

s.t. \[ \|g(x)\|^{\ast} \leq 1, \forall x \in \mathcal{X} \]
\[ \|\nabla g(x)\|^{\ast} \leq 1, \forall x \in \mathcal{X} \]
\[ \|\nabla g(x) - \nabla g(y)\|^{\ast} \leq \|x - y\|, \forall x, y \in \mathcal{X} \]

- **Objective only depends on** the values of \( g \) and \( \nabla g \) at the \( n \) sample points \( x_i \)
- **Infinite-dimensional problem with infinitude of constraints**

**Idea:** Find alternative Stein set \( G \) with equivalent convergence properties and only finitely many constraints
Graph Stein Discrepancies

For any graph \( G = (V, E) \) with \( V \subset \mathcal{X} \), define the graph Stein set \( \mathcal{G}_{\|\cdot\|, Q, G} \) of functions \( g: \mathcal{X}^d \to \mathbb{R}^d \) with

- Boundedness constraints imposed only at points \( x \in V \)
- Smoothness constraints imposed only between pairs \( (x, y) \in E \)
- **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

\[
S(Q, T_P, \mathcal{G}_{\|\cdot\|}) \leq S(Q, T_P, \mathcal{G}_{\|\cdot\|, Q, G_1}) \leq \kappa_d S(Q, T_P, \mathcal{G}_{\|\cdot\|})
\]

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

- Follows from Whitney-Glaeser extension theorem [Glaeser, 1958]
- \( S(Q, T_P, \mathcal{G}_{\|\cdot\|, Q, G_1}) \) 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

- 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 $\{x_1, \ldots, x_n\}$, and $G_t$ is a $t$-spanner on $\{x_1, \ldots, x_n\}$, then

$$S(Q, T_P, G_{\|\cdot\|,Q,G_1}) \leq S(Q, T_P, G_{\|\cdot\|,Q,G_t}) \leq 2t^2 S(Q, T_P, G_{\|\cdot\|,Q,G_1}).$$

- 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

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

s.t. $\|\gamma_j\|_\infty \leq 1, \|\Gamma_j\|_\infty \leq 1$, 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 $\Gamma_{jki} = \nabla_k g_j(x_i)$
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 $S(Q_{1:n}, T_P, G_{\|\cdot\|, Q, G_1}) \to 0$ & $S(Q'_{1:n}, T_P, G_{\|\cdot\|, Q, G_1}) \not\to 0$
Middle: Recovered optimal functions $g$

Right: Associated test functions $h(x) \triangleq T_P g$ which best discriminate sample $Q$ from target $P$
Comparing Discrepancies

Setup

- Draw \( n = 30,000 \) points i.i.d. from \( \mathcal{N}(0, 1) \) or \( \text{Unif}[0, 1] \)
  - Yields sample sequence \( Q_{1:n} \)

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

- 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

Discrepancy value vs. Number of sample points, n for different seeds and distributions.
Selecting Sampler Hyperparameters

Target posterior density: \( p(x) \propto \pi(x) \prod_{l=1}^{L} \pi(y_l | x) \)
- Prior \( \pi(x) \), Likelihood \( \pi(y | x) \)

Stochastic Gradient Langevin Dynamics (SGLD)
[Welling and Teh, 2011]
\[
x_{k+1} \sim \mathcal{N}(x_k + \frac{\epsilon}{2} (\nabla \log \pi(x_k) + \frac{L}{|B_k|} \sum_{l \in B_k} \nabla \log \pi(y_l | x_k)), \epsilon)
\]
- Approximate MCMC procedure designed for scalability
  - Approximates Metropolis-adjusted Langevin algorithm and continuous-time Langevin diffusion
  - Random subset \( B_k \) of datapoints used to select each sample
  - No Metropolis-Hastings correction step
  - Target \( P \) is not stationary distribution
- Choice of step size \( \epsilon \) 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

\[
Y_l | X \sim \text{iid} \frac{1}{2} \mathcal{N}(X_1, 2) + \frac{1}{2} \mathcal{N}(X_1 + X_2, 2)
\]

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 \( \epsilon \), use SGLD with batch size 10 to draw approximate posterior sample \( Q \) of size \( n = 1000 \)

- Use minimum Stein discrepancy to select appropriate \( \epsilon \)
  - 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
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: \( p(x) \propto \pi(x) \prod_{l=1}^{L} \pi(y_l | x) \)
- Prior \( \pi(x) \), Likelihood \( \pi(y | x) \)

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 \( \epsilon \) controls number of datapoints considered
  - Larger \( \epsilon \) ⇒ fewer datapoints considered, fewer likelihood computations, more rapid sampling, more rapid variance reduction
  - Smaller \( \epsilon \) ⇒ 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
    \[ P(Y_l = 1|v_l, X) = 1/(1 + \exp(-\langle v_l, X \rangle)) \]

- 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 $Q$ and under a MALA chain with $10^7$ samples
• Non-Stein measures based on additional, long-running chain used as surrogate for the target distribution
• Stein discrepancy computed from sample $Q$ alone
Assessing Convergence Rates

An observation

- The approximating distribution $Q$ in $\mathcal{S}(Q, T_P, G_{\|\cdot\|,Q,G})$ 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

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

Sobol sequence [Sobol, 1967]

$$d_H(Q_n, P) = O(\log^d(n)/n)$$ for bounded total variation functions

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

- $d_H(Q_n, P) = O(1/n)$ for finite-dimensional Hilbert space $\mathcal{H}$
- $d_H(Q_n, P) = O(1/\sqrt{n})$ 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 $\text{Unif}[0, 1]$ (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
Stein discrepancy convergence for deterministic sequences, kernel herding [Chen, Welling, and Smola, 2010b] 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?
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 sequence \( Q_{1:n} \) to i.i.d. \( \text{Beta}(3, 3) \times \text{Beta}(3, 3) \) sequence \( Q'_{1:n} \).
A Simple Constrained Example

- **Middle:** Recovered optimal functions $g$
- **Right:** Associated test functions $h(x) \triangleq \mathcal{T}_P g$ which best discriminate sample $Q$ from target $P$
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., \( \nabla \log p(x_i) = \nabla \log \pi(x_i) + \sum_{l=1}^{L} \nabla \log \pi(y_l | x_i) \)
   - Improve scalability with first order methods?

2. Establishing reference IPM lower bounds for Stein discrepancy
   - For what other families of distributions \( P \) does \( S(Q_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \rightarrow 0 \) imply \( d_{\mathcal{W}_{\|\cdot\|}}(Q_m, P) \rightarrow 0 \)?

3. Exploring the impact of Stein operator choice
   - An infinite number of operators \( \mathcal{T} \) characterize a given target distribution
   - How is diagnostic impacted by choice of operator? How do we select the best \( \mathcal{T} \)?
References I


P. Chew. There is a planar graph almost as good as the complete graph. In *Proceedings of the Second Annual Symposium on Computational Geometry*, SCG ’86, pages 169–177, New York, NY, 1986. ACM.


