Measuring Sample Quality with Kernels

Lester Mackey*

Joint work with Jackson Gorham†

Microsoft Research*, Opendoor Labs†

June 25, 2018
Motivation: Large-scale Posterior Inference

Example: Bayesian logistic regression

1. Fixed covariate vector: \( v_l \in \mathbb{R}^d \) for each datapoint \( l = 1, \ldots, L \)
2. Unknown parameter vector: \( \beta \sim \mathcal{N}(0, I) \)
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
  \( \mathbb{E}_P[h(Z)] = \int_X p(x) h(x) dx \) 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?

- **MCMC Benefit:** Approximates intractable posterior expectations \( \mathbb{E}_P[h(Z)] = \int_X p(x) h(x) \, dx \) with asymptotically exact sample estimates \( \mathbb{E}_Q[h(X)] = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \)

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

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


- 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


- 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)

**Examples**

- Bounded Lipschitz (or Dudley) metric, $d_{BL_{\|\cdot\|}}$

$$\mathcal{H} = BL_{\|\cdot\|} \triangleq \{ h : \sup_x |h(x)| + \sup_{x \neq y} \frac{|h(x)-h(y)|}{\|x-y\|} \leq 1 \}$$

- 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 $E_P[(\mathcal{T}g)(Z)] = 0$ for all $g \in \mathcal{G}$.

$\mathcal{T}$ and $\mathcal{G}$ together define the **Stein discrepancy** [Gorham and Mackey, 2015]

$$S(Q_n, \mathcal{T}, \mathcal{G}) \triangleq \sup_{g \in \mathcal{G}} |E_{Q_n}[(\mathcal{T}g)(X)]| = d_{\mathcal{G}}(Q_n, P),$$

an IPM-type measure with no explicit integration under $P$

2. **Lower bound** $S(Q_n, \mathcal{T}, \mathcal{G})$ by reference IPM $d_{\mathcal{H}}(Q_n, P)$

$$\Rightarrow S(Q_n, \mathcal{T}, \mathcal{G}) \to 0 \text{ only if } (Q_n)_{n \geq 1} \text{ converges to } P \text{ (Req. II)}$$

- Performed once, in advance, for large classes of distributions

3. **Upper bound** $S(Q_n, \mathcal{T}, \mathcal{G})$ 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 $\mathbb{E}_\mathcal{P}[(\mathcal{T}g)(Z)] = 0$ 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 $\mathcal{P}$
- Under mild conditions, its infinitesimal generator $(\mathcal{A}u)(x) = \lim_{t \to 0} (\mathbb{E}[u(Z_t) | Z_0 = x] - u(x))/t$ satisfies $\mathbb{E}_\mathcal{P}[(\mathcal{A}u)(Z)] = 0$

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 $\mathcal{P}$ only through $\nabla \log p$; computable even if $p$ cannot be normalized!
- Multivariate generalization of **density method** operator
  $(\mathcal{T} g)(x) = g(x) \frac{d}{dx} \log p(x) + g'(x)$ [Stein, Diaconis, Holmes, and Reinert, 2004]
Identifying a Stein Set $\mathcal{G}$

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

**Approach:** Reproducing kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- 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 kernel $k(x, y) = e^{-\frac{1}{2} \|x-y\|^2_2}$
  - Inverse multiquadric kernel $k(x, y) = (1 + \|x-y\|^2_2)^{-1/2}$
- Generates a reproducing kernel Hilbert space (RKHS) $\mathcal{K}_k$
- We define the **kernel Stein set** $\mathcal{G}_{k, \|\cdot\|}$ as vector-valued $g$ with
  - Each component $g_j$ in $\mathcal{K}_k$
  - Component norms $\|g_j\|_{\mathcal{K}_k}$ jointly bounded by 1
- $\mathbb{E}_P[(\mathcal{T}_P g)(Z)] = 0$ for all $g \in \mathcal{G}_{k, \|\cdot\|}$ under mild conditions [Gorham and Mackey, 2017]
Kernel Stein discrepancy (KSD) $S(Q_n, T_P, G_k, \|\cdot\|)$

- Stein operator $(T_P g)(x) \triangleq \langle g(x), \nabla \log p(x) \rangle + \langle \nabla, g(x) \rangle$
- Stein set $G_{k,\|\cdot\|} \triangleq \{ g = (g_1, \ldots, g_d) \mid \|v\|^* \leq 1 \text{ for } v_j \triangleq \|g_j\|_{K_k} \}$

**Benefit:** Computable in closed form [Gorham and Mackey, 2017]

- $S(Q_n, T_P, G_k, \|\cdot\|) = \|w\|$ for $w_j \triangleq \sqrt{\sum_{i,i'=1}^{n} k_{ij}^0(x_i, x_{i'})}$.
  - Reduces to parallelizable pairwise evaluations of Stein kernels
    
    $$k_{ij}^0(x, y) \triangleq \frac{1}{p(x)p(y)} \nabla x_j \nabla y_j (p(x)k(x, y)p(y))$$

- Stein set choice inspired by control functional kernels
  
  $k_0 = \sum_{j=1}^{d} k_{0j}^i$ of Oates, Girolami, and Chopin [2016]
  
  - When $\|\cdot\| = \|\cdot\|_2$, recovers the KSD of Chwialkowski, Strathmann, and Gretton [2016], Liu, Lee, and Jordan [2016]

- To ease notation, will use $G_k \triangleq G_{k,\|\cdot\|_2}$ in remainder of the talk
Detecting Non-convergence

**Goal:** Show $S(Q_n, T_P, \mathcal{G}_k) \to 0$ only if $(Q_n)_{n \geq 1}$ converges to $P$

- Let $\mathcal{P}$ be the set of targets $P$ with Lipschitz $\nabla \log p$ and distant strong log concavity $(\frac{\langle \nabla \log p(x)/p(y), y-x \rangle}{\|x-y\|^2_2} \geq k$ for $\|x - y\|_2 \geq r)$
  - Includes Gaussian mixtures with common covariance, Bayesian logistic and Student’s t regression with Gaussian priors, ...
- For a different Stein set $\mathcal{G}$, Gorham, Duncan, Vollmer, and Mackey [2016] showed $(Q_n)_{n \geq 1}$ converges to $P$ if $P \in \mathcal{P}$ and $S(Q_n, T_P, \mathcal{G}) \to 0$

**New contribution** [Gorham and Mackey, 2017]

---

**Theorem (Univarite KSD detects non-convergence)**

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 $S(Q_n, T_P, \mathcal{G}_k) \to 0$ only if $(Q_n)_{n \geq 1}$ converges weakly to $P$.

- Justifies use of KSD with Gaussian, Matérn, or inverse multiquadric kernels $k$ in the univariate case
The Importance of Kernel Choice

Goal: Show $S(Q_n, T_P, G_k) \rightarrow 0$ only if $Q_n$ converges to $P$

- 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 $\|x - y\|_2 \rightarrow \infty$, then $S(Q_n, T_P, G_k) \rightarrow 0$ 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
Goal: Show $S(Q_n, T_P, G_k) \to 0$ 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 $S(Q_n, T_P, G_k) \to 0$, then $(Q_n)_{n \geq 1}$ converges weakly to $P$.

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

**Goal:** Show $S(Q_n, T_P, G_k) \to 0$ only if $Q_n$ converges to $P$

- Consider the inverse multiquadric (IMQ) kernel
  \[ k(x, y) = (c^2 + \|x - y\|^2)^\beta \text{ for some } \beta < 0, c \in \mathbb{R}. \]
- IMQ KSD fails to detect non-convergence when $\beta < -1$
- However, IMQ KSD automatically enforces tightness and detects non-convergence when $\beta \in (-1, 0)$

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

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

- No extra assumptions on sample sequence $(Q_n)_{n \geq 1}$ needed
- Intuition: Slow decay rate of kernel $\Rightarrow$ unbounded (coercive) test functions in $T_PG_k \Rightarrow$ non-tight sequences detected
**Goal:** Show $S(Q_n, T_P, G_k) \rightarrow 0$ when $Q_n$ 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 $S(Q_n, T_P, G_k) \rightarrow 0$ whenever the Wasserstein distance $d_{W_{\|\cdot\|_2}}(Q_n, P) \rightarrow 0$.

- Covers Gaussian, Matérn, IMQ, and other common bounded kernels $k$
A Simple Example

**Left plot:**
- For target \( p(x) \propto e^{-\frac{1}{2}(x+1.5)^2} + e^{-\frac{1}{2}(x-1.5)^2} \), compare an i.i.d. sample \( Q_n \) from \( P \) and an i.i.d. sample \( Q'_n \) from one component.
- Expect \( S(Q_{1:n}, T_P, G_k) \to 0 \) & \( S(Q'_{1:n}, T_P, G_k) \not\to 0 \)
- Compare **IMQ KSD** (\( \beta = -\frac{1}{2}, c = 1 \)) with **Wasserstein distance**
**Right plot:** For $n = 10^3$ sample points,

- (Top) Recovered optimal Stein functions $g$
- (Bottom) Associated test functions $h \triangleq T_P g$ which best discriminate sample $Q_n$ from target $P$
The Importance of Kernel Choice

- 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 ($\beta = -\frac{1}{2}, c = 1$) does not have this deficiency
Selecting Sampler Hyperparameters

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

Approximate slice sampling [DuBois, Korattikara, Welling, and Smyth, 2014]
- Approximate MCMC procedure designed for scalability
  - Uses random subset of datapoints to approximate each slice sampling step
  - Target \( P \) is not stationary distribution
- Tolerance parameter \( \epsilon \) controls number of datapoints evaluated
  - \( \epsilon \) too small \( \Rightarrow \) too few sample points generated
  - \( \epsilon \) 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
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
  \[ Y_l|X \overset{iid}{\sim} \frac{1}{2} \mathcal{N}(X_1, 2) + \frac{1}{2} \mathcal{N}(X_1 + X_2, 2) \]
  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 parameters $\epsilon$, run approximate slice sampling for 148000 datapoint likelihood evaluations and store resulting posterior sample $Q_n$
- Use minimum IMQ KSD ($\beta = -\frac{1}{2}, c = 1$) 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 sequences
ESS (higher is better)  
KSD (lower is better)  

- ESS maximized at tolerance $\epsilon = 10^{-1}$  
- IMQ KSD minimized at tolerance $\epsilon = 10^{-2}$
Selecting Samplers

**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 Fisher Scoring (SGFS)**
[Ahn, Korattikara, and Welling, 2012]
- Approximate MCMC procedure designed for scalability
  - Approximates Metropolis-adjusted Langevin algorithm and continuous-time Langevin diffusion with preconditioner
  - Random subset of datapoints used to select each sample
  - No Metropolis-Hastings correction step
  - Target \( P \) is not stationary distribution
- 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
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]
Selecting Samplers

**Left:** IMQ KSD quality comparison for SGFS Bayesian logistic regression (no surrogate ground truth used)

**Right:** SGFS sample points \((n = 5 \times 10^4)\) with bivariate marginal means and 95\% confidence ellipses (blue) that align best and worst with surrogate ground truth sample (red).

Both suggest small speed-up of SGFS-d (0.0017\(s\) per sample vs. 0.0019\(s\) for SGFS-f) outweighed by loss in inferential accuracy
Goodness-of-fit testing

- Chwialkowski, Strathmann, and Gretton [2016] used the KSD $S(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ 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 $x_i = z_i + u_i e_1$
  - $z_i \sim \mathcal{N}(0, I_d)$ and $u_i \sim \text{Unif}[0, 1]$. 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

<table>
<thead>
<tr>
<th></th>
<th>d=2</th>
<th>d=5</th>
<th>d=10</th>
<th>d=15</th>
<th>d=20</th>
<th>d=25</th>
</tr>
</thead>
<tbody>
<tr>
<td>B&amp;H</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.91</td>
<td>0.57</td>
<td>0.26</td>
</tr>
<tr>
<td>Gaussian</td>
<td>1.0</td>
<td>1.0</td>
<td>0.88</td>
<td>0.29</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>IMQ</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Improving sample quality

- Given sample points \((x_i)_{i=1}^n\), can minimize KSD \(S(\tilde{Q}_n, \mathcal{T}_P, \mathcal{G}_k)\) 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 [2016] do this with Gaussian kernel \(k(x, y) = e^{-\frac{1}{h}\|x-y\|^2}\)
  - Bandwidth \(h\) set to median of the squared Euclidean distance between pairs of sample points.

- We recreate their experiment with the IMQ kernel
  \[k(x, y) = (1 + \frac{1}{h}\|x - y\|^2)^{-1/2}\]
MSE averaged over 500 simulations (±2 standard errors)

Target $P = \mathcal{N}(0, I_d)$

Starting sample $Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ for $x_i \overset{iid}{\sim} P$, $n = 100$. 
Future Directions

Many opportunities for future development

1. Improve KSD scalability while maintaining convergence control
   - Inexpensive approximations of kernel matrix
   - Subsampling of likelihood terms in $\nabla \log p$

2. Addressing other inferential tasks
   - Control variate design
     - [Oates, Girolami, and Chopin, 2016]
   - Variational inference
     - [Liu and Wang, 2016, Liu and Feng, 2016]
   - Training generative adversarial networks
     - [Wang and Liu, 2016] and variational autoencoders
     - [Pu, Gan, Henao, Li, Han, and Carin, 2017]

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 $\mathcal{T}$?
   - **Thm:** If $\nabla \log p$ bounded and $k \in C_0^{(1,1)}$, then
     $S(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ for some $(Q_n)_{n \geq 1}$ not converging to $P$
   - Diffusion Stein operators
     - $(\mathcal{T} g)(x) = \frac{1}{p(x)} \langle \nabla, p(x)m(x)g(x) \rangle$
     - Gorham, Duncan, Vollmer, and Mackey [2016] may be appropriate for heavy tails


Comparing Discrepancies

- **Left:** Samples drawn i.i.d. from either the bimodal Gaussian mixture target $p(x) \propto e^{-\frac{1}{2}(x+1.5)^2} + e^{-\frac{1}{2}(x-1.5)^2}$ or a single mixture component.

- **Right:** Discrepancy computation time using $d$ cores in $d$ dimensions.