Measuring Sample Quality with Stein's Method

Jackson Gorham and Lester Mackey
{jgorham,lmackey}@stanford.edu

One-minute Summary:

To improve the efficiency of Monte Carlo estimation, practitioners are turning to biased Markov chain Monte Carlo procedures that trade off asymptotic exactness for computational speed. The reasoning is sound: a reduction in variance due to more rapid sampling can outweigh the bias introduced. However, the inexactness creates new challenges for sampler and parameter selection, since standard measures of sample quality like effective sample size do not account for asymptotic bias. To address these challenges, we introduce a new computable quality measure that quantifies the maximum discrepancy between sample and target expectations over a large class of test functions. We use our tool to compare exact, biased, and deterministic sample sequences and illustrate applications to hyperparameter selection, convergence rate assessment, and quantifying bias-variance tradeoffs in posterior inference.

Motivation: Approximate MCMC

Example: Bayesian logistic regression

1. Parameter vector: $\beta \in \mathbb{R}^d, \beta \sim \mathcal{N}(0, I)$
2. Fixed covariate vector: $v_l \in \mathbb{R}^d, l = 1, \dots, L$
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

- **MCMC Benefit:** Approximates intractable posterior expectations $\mathbb{E}_P[h(Z)] = \int_{\mathcal{X}} p(x)h(x)dx$ with asymptotically exact sample estimates $\mathbb{E}_Q[h(X)] = \sum_{i=1}^n q(x_i)h(x_i)$
- **Problem:** Each sample 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 MCMC procedure in a manner that makes use of only a small subset of datapoints per sample
- Introduces asymptotic bias but reduced computational overhead leads to faster sampling and reduced Monte Carlo variance

Introduces new challenges

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

Sample Quality Measures

- **Target distribution** P, support $\mathcal{X} = \mathbb{R}^d$ (can relax to any convex set), density p (known up to normalization)
- **Weighted sample** Q: sample points $x_1, \dots, x_n \in \mathcal{X}$, weights $q(x_i)$
- **Goal:** Quantify how well $\mathbb{E}_Q$ 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

Idea: Consider an **integral probability metric (IPM)**

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

Example: Wasserstein, $d_{\mathcal{W}_{\|\cdot\|}}$ $(\mathcal{H} = \mathcal{W}_{\|\cdot\|} \triangleq \{h : \sup_{x \neq y} \frac{|h(x)-h(y)|}{\|x-y\|} \leq 1\})$

Problem: Typically cannot compute as integration under P intractable!
Idea: Only consider functions with $\mathbb{E}_P[h(Z)]$ known *a priori* to be 0

Stein's Method

- Typically used as an analytical tool to prove distributional convergence
- **Our goal:** Develop into a practical quality measure (Requirement III)

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 $\mathcal{G}$ define the **Stein discrepancy**

$$\boldsymbol{\mathcal{S}(Q, \mathcal{T}, \mathcal{G})} \triangleq \sup_{g \in \mathcal{G}} |\mathbb{E}_Q[(\mathcal{T}g)(X)]| = d_{\mathcal{T}\mathcal{G}}(Q, P)$$

   How to pick $\mathcal{T}$? The **infinitesimal generator** of a Markov process with stationary distribution P yields suitable operator. Example:
   - **Overdamped Langevin diffusion:** $dZ_t = \frac{1}{2}\nabla \log p(Z_t)dt + dW_t$
   - **Stein Operator** does not depend on normalizing constant:

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

   - $\mathbb{E}_P[(\mathcal{T}_P g)(Z)] = 0$ for all $g : \mathcal{X} \to \mathbb{R}^d$ in **classical Stein set**

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

2. **Lower bound $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$ by reference IPM (Req. II)**. New result:
   Theorem 1 (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 $\mathcal{S}(Q_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0 \Rightarrow d_{\mathcal{W}_{\|\cdot\|}}(Q_m, P) \to 0$.*
   - Sufficient, not necessary; covers Bayesian logistic regression example
3. **Upper bound $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$ to demonstrate convergence (Req. I)**
   Proposition 2. *If $X \sim Q$ and $Z \sim P$ with $\nabla \log p(Z)$ integrable, then*

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \leq \mathbb{E}[\|X - Z\|] + \mathbb{E}[\|\nabla \log p(X) - \nabla \log p(Z)\|] + \mathbb{E}\left[\left\|\nabla \log p(Z)(X - Z)^\top\right\|\right].$$

Computing Stein Discrepancies

Classical Stein discrepancy optimization problem:

$$\begin{aligned} \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) = \sup_g \sum_{i=1}^n & q(x_i)(\langle g(x_i), \nabla \log p(x_i)\rangle + \langle \nabla, g(x_i)\rangle) \\ \text{s.t. } & \|g(x)\|^* \leq 1, \forall x \in \mathcal{X} \\ & \|\nabla g(x)\|^* \leq 1, \forall x \in \mathcal{X} \\ & \|\nabla g(x) - \nabla g(y)\|^* \leq \|x - y\|, \forall x, y \in \mathcal{X} \end{aligned}$$

Problem: Infinite-dimensional problem with infinitude of constraints!

Solution: Graph Stein Discrepancies
For any graph $G = (V, E)$ with $V \subset \mathcal{X}$, define the **graph Stein set**:

$$\begin{aligned} \mathcal{G}_{\|\cdot\|, Q, G} = \{ & g \mid \forall\, x \in V,\ \max\big(\|g(x)\|^*, \|\nabla g(x)\|^*\big) \leq 1, \\ & \forall\, (x, y) \in E : x \neq y, \max\left(\tfrac{\|g(x)-g(y)\|^*}{\|x-y\|}, \tfrac{\|\nabla g(x)-\nabla g(y)\|^*}{\|x-y\|}\right) \leq 1, \\ & \max\left(\tfrac{\|g(x)-g(y)-\nabla g(x)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2}, \tfrac{\|g(x)-g(y)-\nabla g(y)(x-y)\|^*}{\frac{1}{2}\|x-y\|^2}\right) \leq 1\}, \end{aligned}$$

Proposition 3 (Equivalence of Classical & Complete Graph Stein Discrepancies). *If $\mathcal{X} = \mathbb{R}^d$, and G_1 is the complete graph on $\{x_1, \dots, x_n\}$,*

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

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

Problem: Complete graph introduces order n^2 constraints!
Solution: Spanner Stein Discrepancies

- For a **dilation factor** $t \geq 1$, a **t-spanner** $G = (V, E)$ has
  - weight $\|x - y\|$ on each edge $(x, y) \in E$
  - a path with total weight no larger than $t\|x - y\| \forall x, y \in V$

Proposition 4 (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*

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

- 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];
- We use efficient greedy spanner code of Bouts, ten Brink, and Buchin [2014]

Spanner Stein discrepancy algorithm (recommended)

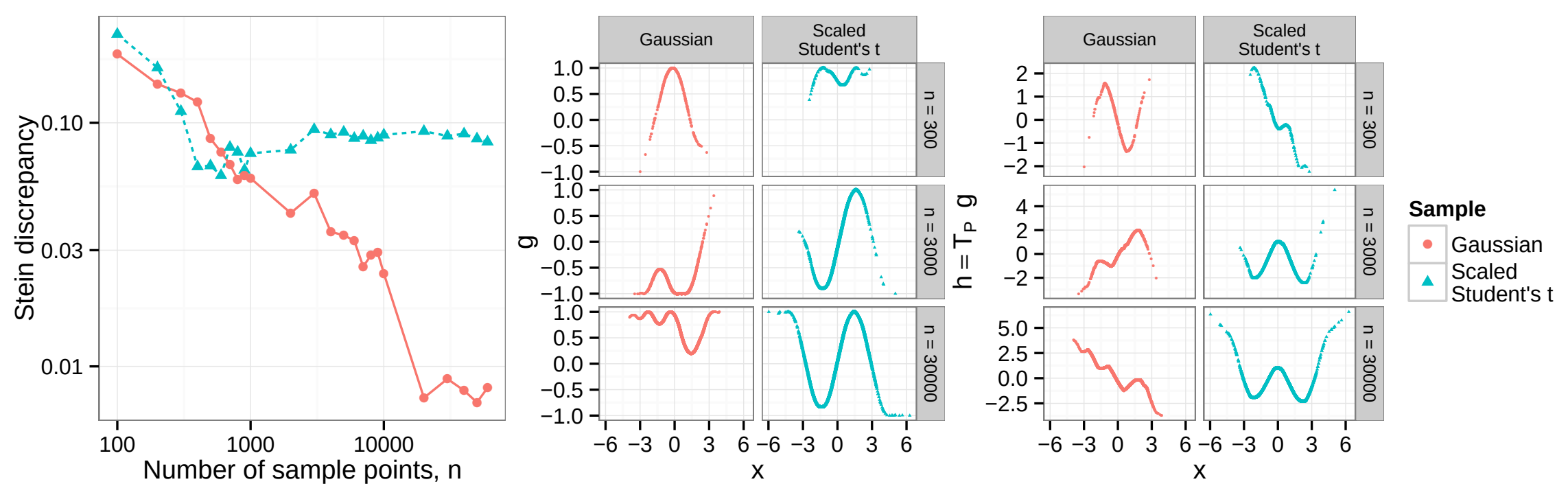
- Choose $\|\cdot\| = \|\cdot\|_1$; Compute 2-spanner G_2 on $V = \{x_1, \dots, x_n\}$
- Solve d finite-dimensional linear programs in parallel

$$\begin{aligned} \sum_{j=1}^d \sup_{\substack{\gamma_j \in \mathbb{R}^n, \\ \Gamma_j \in \mathbb{R}^{d \times n}}} \sum_{i=1}^n & q(x_i)(\gamma_{ji} \nabla_j \log p(x_i) + \Gamma_{jji}) \\ \text{s.t. } & \|\gamma_j\|_\infty \leq 1, \|\Gamma_j\|_\infty \leq 1, \text{ and } \forall\, i \neq l : (x_i, x_l) \in E, \\ & \max\left(\tfrac{|\gamma_{ji} - \gamma_{jl}|}{\|x_i - x_l\|_1}, \tfrac{\|\Gamma_j(e_i - e_l)\|_\infty}{\|x_i - x_l\|_1}\right) \leq 1, \\ & \max\left(\tfrac{|\gamma_{ji} - \gamma_{jl} - \langle \Gamma_j e_i, x_i - x_l\rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}, \tfrac{|\gamma_{ji} - \gamma_{jl} - \langle \Gamma_j e_l, x_i - x_l\rangle|}{\frac{1}{2}\|x_i - x_l\|_1^2}\right) \leq 1. \end{aligned}$$

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

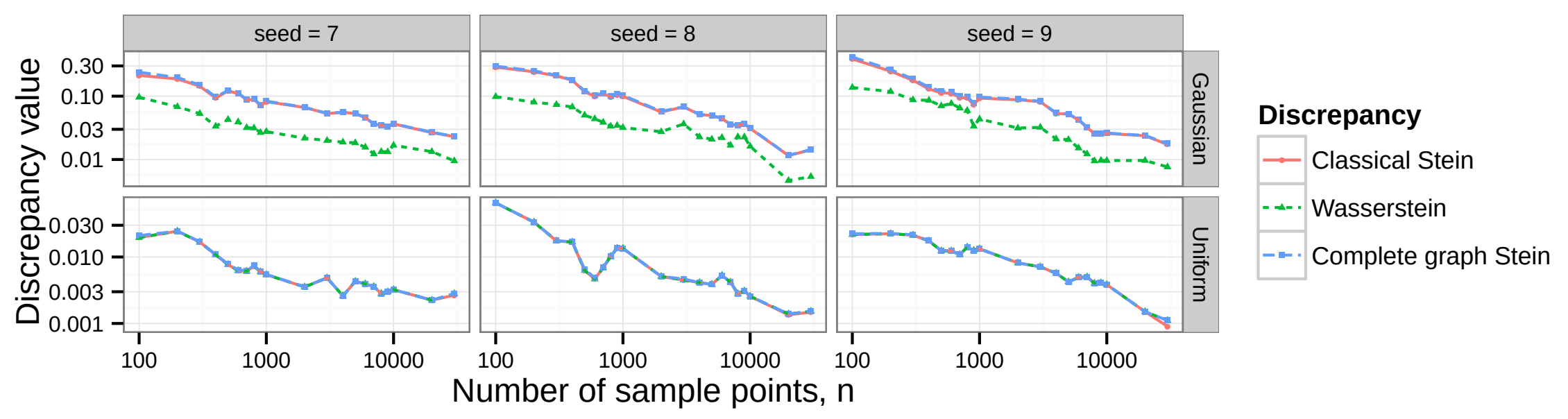
Experiments

1. A Sanity Check



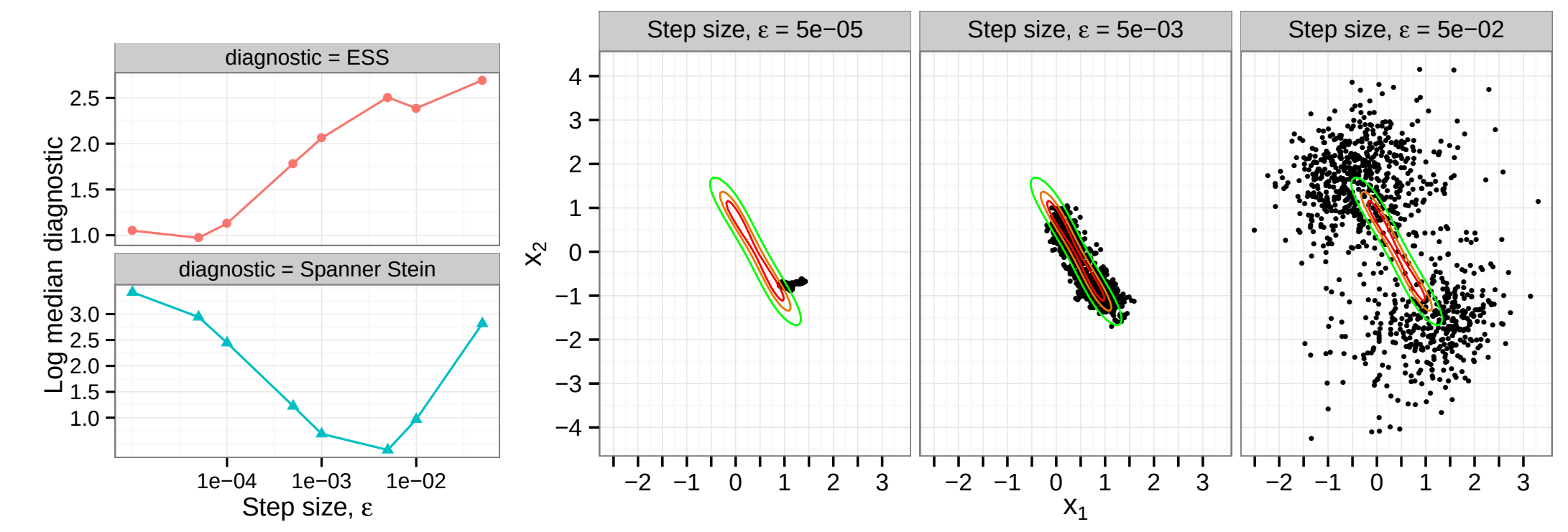
- $P = \mathcal{N}(0, 1)$, Sample Q from P or scaled Student's t (same variance)

2. Comparing Discrepancies



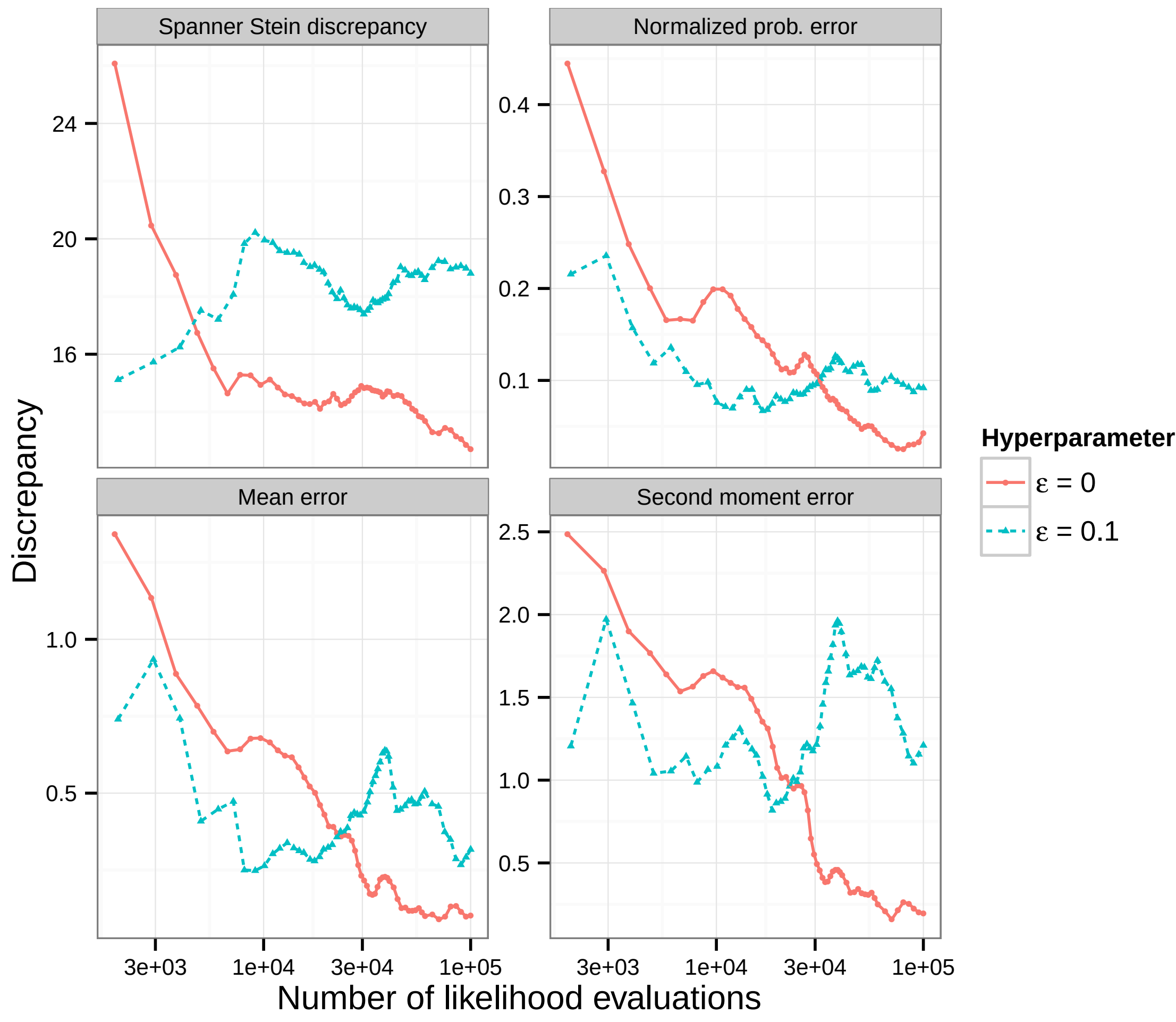
- Draw samples from target ($P = \mathcal{N}(0, 1)$ or $P = \text{Unif}[0, 1]$)
- Compare classical / graph Stein discrepancies and Wasserstein metric

3. Selecting Sampler Hyperparameters



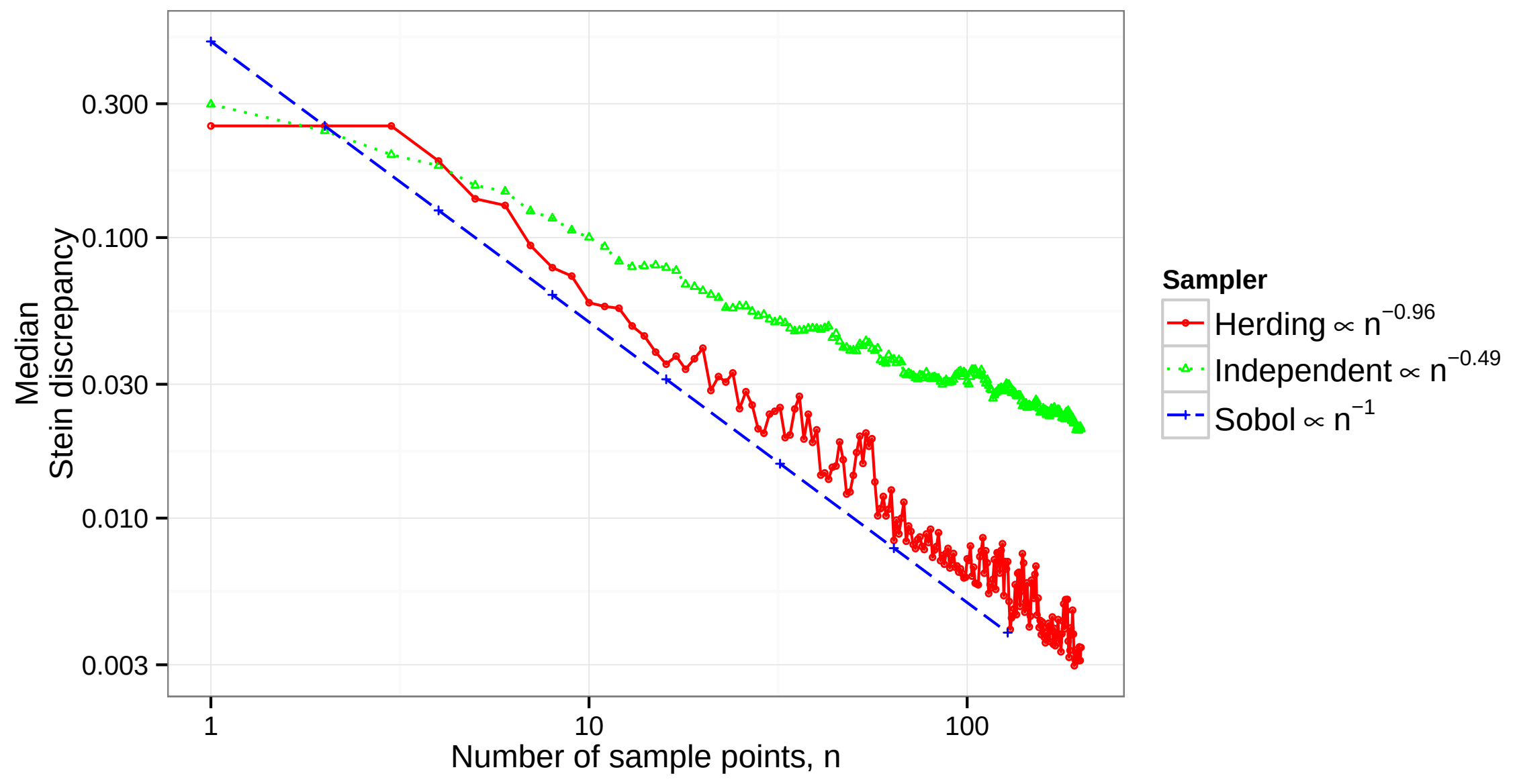
- Target P is bimodal GMM
- Run Stochastic Gradient Langevin Dynamics for many step sizes ϵ
- ESS chooses $\epsilon = 5 \times 10^{-2}$, Stein discrepancy chooses $\epsilon = 5 \times 10^{-3}$

4. Quantifying a Bias-Variance Trade-off



- **Nodal dataset**: 53 patients, 6 predictors of cancer spread
- Model=*Bayesian Logistic Regression with Gaussian priors*
- Random Walk MH ($\epsilon = 0$) vs. Approximate RWMH ($\epsilon = 0.1$)

5. Assessing Sampler Convergence Rates



- For target $P = \text{Unif}(0, 1)$, compare Sobol ($O(\log(n)/n)$) vs. iid ($O(1/\sqrt{n})$) vs. kernel herding (best known bound $O(1/\sqrt{n})$) sequence