Measuring Sample Quality with Stein's Method

An attempt at quantifying Monte-Carlo efficiency by Gorham and Mackey (2015)

Luis Sierra Muntané
luis.sierra@mail.utoronto.ca

Bayesian Reading Group
University of Toronto DoSS

9th December 2025

Outline

Our goal is to calculate quantities of the form $\mathbb{E}_P[h(X)]$ where P is an unknown target distribution form the samples of Q as

$$\mathbb{E}_Q[h(X)] = \sum_{i=1}^{n} q(x_i)h(x_i).$$

We will be following (Gorham and Mackey, 2015).

- Revisiting Stein's Lemma
- Finding distances between distributions
- How Stein's method provides an answer
- Constructing Stein operators
- Calculating the discrepancies
- Experiments

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longrightarrow \quad \mathrm{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longleftrightarrow \quad \mathrm{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longleftrightarrow \quad \mathrm{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

$$\mathbb{E}[g'(X) - Xg(X)] = 0,$$

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longleftrightarrow \quad \mathrm{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

$$\mathbb{E}[g'(X) - Xg(X)] = 0, \qquad \longrightarrow \qquad \mathbb{E}[\mathcal{A}g(X)] = 0$$

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longleftrightarrow \quad \operatorname{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

$$\mathbb{E}[g'(X) - Xg(X)] = 0, \qquad \longrightarrow \qquad \mathbb{E}[\mathcal{A}g(X)] = 0$$

$$\begin{aligned} X \sim \mathrm{Gamma}(\alpha, \beta) \quad &\Longleftrightarrow \quad \mathbb{E}[Xf'(X) + (\alpha - \beta X)f(X)] = 0 \\ X \sim \mathrm{Poisson}(\lambda) \quad &\Longleftrightarrow \quad \mathbb{E}[\lambda f(X+1) - Xf(X)] = 0 \\ X \sim \mathrm{Binomial}(n, p) \quad &\Longleftrightarrow \quad \mathbb{E}[(1-p)Xf(X) - p(n-X)f(X+1)] \end{aligned}$$

Stein's Lemma

For X mean-zero random variable and g a weakly differentiable and L_1 integrable function, we have that

$$X \text{ standard Gaussian} \quad \Longleftrightarrow \quad \mathrm{Cov}(X, g(X)) = \mathbb{E}[g'(X)],$$

$$\mathbb{E}[g'(X) - Xg(X)] = 0, \qquad \longrightarrow \qquad \mathbb{E}[\mathcal{A}g(X)] = 0$$

$$\begin{aligned} X \sim \mathrm{Gamma}(\alpha, \beta) \quad &\Longleftrightarrow \quad \mathbb{E}[Xf'(X) + (\alpha - \beta X)f(X)] = 0 \\ X \sim \mathrm{Poisson}(\lambda) \quad &\Longleftrightarrow \quad \mathbb{E}[\lambda f(X+1) - Xf(X)] = 0 \\ X \sim \mathrm{Binomial}(n, p) \quad &\Longleftrightarrow \quad \mathbb{E}[(1-p)Xf(X) - p(n-X)f(X+1)] \end{aligned}$$

$$X \sim P \quad \Longleftrightarrow \quad \mathbb{E}_P[(\mathcal{A}f)(X)] = 0.$$

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

- Detects convergence $Q_m \to P$
- Detects Q not converging to P
- Is computationally feasible

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

- Detects convergence $Q_m \to P$
- Detects Q not converging to P
- Is computationally feasible

A possible way to ensure these is by using an *integral probability metric*

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

- Detects convergence $Q_m \to P$
- Detects Q not converging to P
- Is computationally feasible

A possible way to ensure these is by using an *integral probability metric*

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

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

- Detects convergence $Q_m \to P$
- Detects Q not converging to P
- Is computationally feasible

A possible way to ensure these is by using an *integral probability metric*

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

Quality Measures for samples

Target distribution P with open convex support $\mathcal{X} \subseteq \mathbb{R}^d$. We approximate P with Q.

Goal: to have a measure of the quality of the samples Q:

- Detects convergence $Q_m \to P$
- Detects Q not converging to P
- Is computationally feasible

A possible way to ensure these is by using an *integral probability metric*

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

How to ensure that calculating this expression is tractable?

Stein's Method

Characterizing convergence in distribution (Stein, 1972):

Stein's Method

Characterizing convergence in distribution (Stein, 1972):

1. Find a real-valued operator $\mathcal{T} : \mathcal{G} \to \mathbb{R}$ characterizing P in the sense that

$$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \quad \forall g \in \mathcal{G}.$$

   Together, $\mathcal{T}, \mathcal{G}$ define the *Stein discrepancy*

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

Stein's Method

Characterizing convergence in distribution (Stein, 1972):

1. Find a real-valued operator $\mathcal{T} : \mathcal{G} \to \mathbb{R}$ characterizing P in the sense that
$$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \quad \forall g \in \mathcal{G}.$$
Together, $\mathcal{T}, \mathcal{G}$ define the *Stein discrepancy*
$$\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) := \sup_{g \in \mathcal{G}} |\mathbb{E}_Q[(\mathcal{T}g)(X)]| = d_{\mathcal{G}}(Q, P).$$
2. Lower bound $\mathcal{S}$ by some familiar IPM $d_{\mathcal{H}}$. **Reliability**: for $\{\mu_m\}_{m \geq 1}$
$$d_{\mathcal{H}}(\mu_m, P) \to 0 \quad \Longrightarrow \quad \mathcal{S}(\mu_m, \mathcal{T}, \mathcal{G}) \to 0.$$

Stein's Method

Characterizing convergence in distribution (Stein, 1972):

1. Find a real-valued operator $\mathcal{T} : \mathcal{G} \to \mathbb{R}$ characterizing P in the sense that
$$\mathbb{E}_P[(\mathcal{T}g)(Z)] = 0 \quad \forall g \in \mathcal{G}.$$
Together, $\mathcal{T}, \mathcal{G}$ define the *Stein discrepancy*
$$\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) := \sup_{g \in \mathcal{G}} |\mathbb{E}_Q[(\mathcal{T}g)(X)]| = d_{\mathcal{G}}(Q, P).$$
2. Lower bound $\mathcal{S}$ by some familiar IPM $d_{\mathcal{H}}$. **Reliability**: for $\{\mu_m\}_{m \geq 1}$
$$d_{\mathcal{H}}(\mu_m, P) \to 0 \quad \implies \quad \mathcal{S}(\mu_m, \mathcal{T}, \mathcal{G}) \to 0.$$
3. Upper bound $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$ to demonstrate convergence to zero (**Consistency**).

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

$$\int \frac{\mathcal{A}_t f(x) - f(x)}{t} \, d\mu(x) = 0,$$

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

$$\lim_{t \downarrow 0} \int \frac{\mathcal{A}_t f(x) - f(x)}{t} \, d\mu(x) = 0, ,$$

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

$$\lim_{t \downarrow 0} \int \frac{\mathcal{A}_t f(x) - f(x)}{t} \, d\mu(x) = 0, ,$$

$$(\mathcal{A}u)(x) = \lim_{t \to 0} \frac{1}{t} \left(\mathbb{E}[u(x) \,|\, Z_0 = x] - u(x) \right).$$

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

$$\lim_{t \downarrow 0} \int \frac{\mathcal{A}_t f(x) - f(x)}{t} \, d\mu(x) = 0, ,$$

$$(\mathcal{A}u)(x) = \lim_{t \to 0} \frac{1}{t} \left(\mathbb{E}[u(x) \,|\, Z_0 = x] - u(x) \right).$$

With this we can take for a diffusion $dZ_t = \frac{1}{2} \nabla \log p(Z_t) dt + dW_t$ the *Stein operator*

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

How to construct $\mathcal{T}$

Construct a generator for a Markov process $(Z_t)_{t\geq 0} \to P$ (Barbour, 1988). Consider a semigroup of operators $(\mathcal{A}_t f)(x) = \mathbb{E}[f(X_t) \mid X_0 = x]$. P is a limiting distribution if

$$\int (\mathcal{A}_t f) d\mu = \int f d\mu \quad \forall t \geq 0,$$

$$\lim_{t \downarrow 0} \int \frac{\mathcal{A}_t f(x) - f(x)}{t} \, d\mu(x) = 0, ,$$

$$(\mathcal{A}u)(x) = \lim_{t \to 0} \frac{1}{t} \left(\mathbb{E}[u(x) \,|\, Z_0 = x] - u(x) \right).$$

With this we can take for a diffusion $dZ_t = \frac{1}{2} \nabla \log p(Z_t) dt + dW_t$ the *Stein operator*

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

Stein Set

$$\mathcal{G}_{\|\cdot\|} = \Big\{ g : \mathcal{X} \to \mathbb{R}^d \mid \sup_{x,y \in \mathcal{X}} \left(\|g\|^*, \|\nabla g\|^*, \frac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x - y\|} \right) \leq 1, \\ \langle g(x), \hat{n}(x) \rangle = 0,\ \forall x \in \partial \mathcal{X} \Big\},$$

where

$$\|w\|^* = \sup_{\|v\|=1} \langle w, v \rangle .$$

Stein Set

$$\mathcal{G}_{\|\cdot\|} = \left\{ g : \mathcal{X} \to \mathbb{R}^d \mid \sup_{x,y \in \mathcal{X}} \left(\|g\|^*, \|\nabla g\|^*, \frac{\|\nabla g(x) - \nabla g(y)\|^*}{\|x - y\|} \right) \leq 1, \right.$$
$$\left. \langle g(x), \hat{n}(x) \rangle = 0,\ \forall x \in \partial \mathcal{X} \right\},$$

where

$$\|w\|^* = \sup_{\|v\|=1} \langle w, v \rangle .$$

Observation

This imposes conditions for all pairs of points in $\mathcal{X}$.

Bounding the Stein Discrepancy

Bounding the Stein Discrepancy

Lower bound:

Theorem (*Theorem 2*)

If $\mathcal{X} = \mathbb{R}^d$ and $\log p$ is strongly concave with continuous and bounded 3rd and 4th derivatives then for any measures $(\mu_m)_{m \geq 1}$, $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ only if $d_{\mathcal{W}}(\mu_m, P) \to 0$.

Bounding the Stein Discrepancy

Lower bound:

Theorem (*Theorem 2*)

If $\mathcal{X} = \mathbb{R}^d$ *and* $\log p$ *is strongly concave with continuous and bounded 3rd and 4th derivatives then for any measures* $(\mu_m)_{m \geq 1}$, $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ *only if* $d_{\mathcal{W}}(\mu_m, P) \to 0$.

- Sufficient but not necessary conditions for convergence.

Bounding the Stein Discrepancy

Lower bound:

Theorem (*Theorem 2*)

If $\mathcal{X} = \mathbb{R}^d$ *and* $\log p$ *is strongly concave with continuous and bounded 3rd and 4th derivatives then for any measures* $(\mu_m)_{m \geq 1}$, $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ *only if* $d_{\mathcal{W}}(\mu_m, P) \to 0$.

- Sufficient but not necessary conditions for convergence.

Upper bound:

Theorem (*Proposition 3*)

If $X \sim Q$ *and* $Z \sim P$ *with* $\nabla \log p(Z)$ *integrable, then*

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

Bounding the Stein Discrepancy

Lower bound:

Theorem (*Theorem 2*)

If $\mathcal{X} = \mathbb{R}^d$ and $\log p$ is strongly concave with continuous and bounded 3rd and 4th derivatives then for any measures $(\mu_m)_{m\geq 1}$, $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) \to 0$ only if $d_{\mathcal{W}}(\mu_m, P) \to 0$.

- Sufficient but not necessary conditions for convergence.

Upper bound:

Theorem (*Proposition 3*)

If $X \sim Q$ and $Z \sim P$ with $\nabla \log p(Z)$ integrable, then

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

Implies convergence of $\mathcal{S} \to 0$ whenever $X_m \sim Q_m \xrightarrow{L^2} Z \sim P$ and $\nabla \log p(X_m) \xrightarrow{L^1} \nabla \log p(Z)$.

Computing Stein Discrepancies

For observed sample values $\{x_i\}_{i=1}^n$, we want to solve the optimization problem

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) = \sup_{g \in \mathcal{G}_{\|\cdot\|}} \sum_{i=1}^{n} q(x_i)(\langle g(x_i), \nabla \log g(x_i) \rangle + \nabla \cdot g_{x_i}),$$

such that g satisfies the conditions to be inside $\mathcal{G}_{\|\cdot\|}$.

Computing Stein Discrepancies

For observed sample values $\{x_i\}_{i=1}^n$, we want to solve the optimization problem

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) = \sup_{g \in \mathcal{G}_{\|\cdot\|}} \sum_{i=1}^{n} q(x_i)(\langle g(x_i), \nabla \log g(x_i) \rangle + \nabla \cdot g_{x_i}),$$

such that g satisfies the conditions to be inside $\mathcal{G}_{\|\cdot\|}$.

To make it feasible, Gorham and Mackey (2015) propose constraining the problem only on the values of g on $\{x_i\}_{i=1}^n$.

Computing Stein Discrepancies

For observed sample values $\{x_i\}_{i=1}^n$, we want to solve the optimization problem

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_{\|\cdot\|}) = \sup_{g \in \mathcal{G}_{\|\cdot\|}} \sum_{i=1}^{n} q(x_i)(\langle g(x_i), \nabla \log g(x_i) \rangle + \nabla \cdot g_{x_i}),$$

such that g satisfies the conditions to be inside $\mathcal{G}_{\|\cdot\|}$.

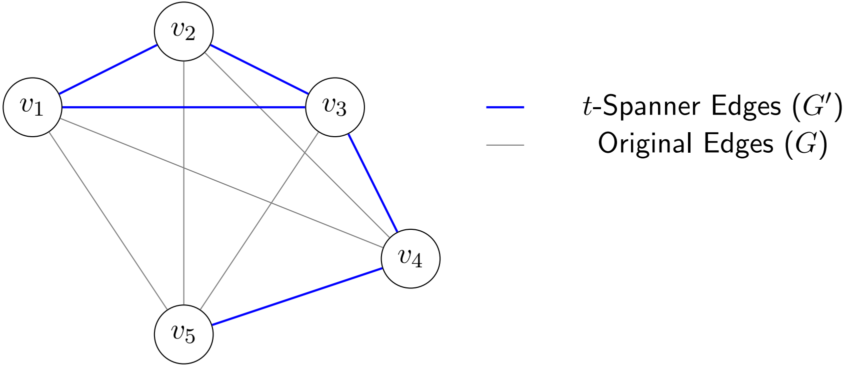
To make it feasible, Gorham and Mackey (2015) propose constraining the problem only on the values of g on $\{x_i\}_{i=1}^n$.

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

for $\gamma_{ji} = g_j(x_i)$, $\Gamma_{jki} = \nabla_k g_j(x_i)$. An efficient way to define the constraints involves using *graph t-spanners* and an ℓ_1 norm.

Graph Spanners

Graph $G = K_n$ (Original)



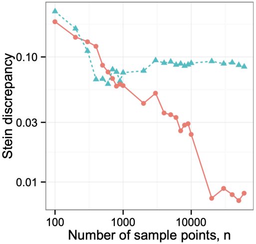
Graph G' (t-Spanner for $t \geq 1$)

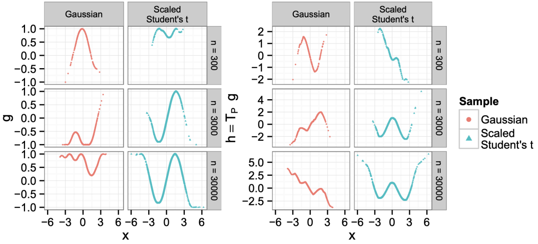
Algorithm Multivariate Spanner Stein Discrepancy (`Algorithm 1` in Gorham & Mackey 2015)

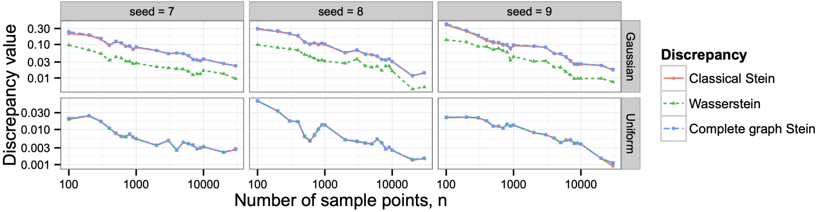
1: **input:** Q, coordinate bounds $(\alpha_1, \beta_1), \ldots, (\alpha_d, \beta_d)$
2: $G_2 \leftarrow$ Compute sparse 2-spanner of $\mathrm{supp}(Q)$
3: **for** $j = 1$ to d **do** (**parallelizable**)
4: $\quad r_j \leftarrow$ Solve the j-th coordinate from linear program $(\star)$
5: **end for** $\sum_{j=1}^{d} r_j$

Experiments

Target distribution $P = \mathcal{N}(0, 1)$.







Extensions

- Change the diffusion process to generate $\mathcal{T}_P$.
- Replacing the calculation of $\mathcal{S}$ with a Kernel Approach (Gorham and Mackey, 2017).
- Consider general diffusion operators (Gorham et al., 2019).
- If $\mathcal{T}_P$ is too expensive to calculate, use *stochastic Stein discrepancies* (SSDs) (Gorham et al., 2020).
- And many others. . . (Anastasiou et al., 2023)

References

A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, L. Mackey, C. J. Oates, G. Reinert, and Y. Swan. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023. doi: 10.1214/22-STS863. URL https://projecteuclid.org/journals/statistical-science/volume-38/issue-1/Steins-Method-Meets-Computational-Statistics--A-Review-of-Some/10.1214/22-STS863.full.

A. D. Barbour. Stein's method and poisson process convergence. *Journal of Applied Probability*, 25A:175–184, 1988.

J. Gorham and L. Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301. PMLR, Aug 06–11 2017. URL https://proceedings.mlr.press/v70/gorham17a.html.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. W. Mackey. Measuring sample quality with diffusions. *Annals of Applied Probability*, 29(5):2884–2928, 2019. doi: 10.1214/19-AAP1467.

J. Gorham, A. Raj, and L. W. Mackey. Stochastic stein discrepancies. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 17931–17942. Curran Associates, Inc., Dec. 2020. URL https://proceedings.neurips.cc/paper/2020/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume II: Probability Theory*, pages 583–602. University of California Press, 1972.