Gorham et al. - Measuring sample quality with diffusions

Notes by Joris Bierkens

9th June 2020

1 Introduction

Suppose Q and P are probability distributions on $\mathbb{R}^d$. The Wasserstein distance is defined as $\mathcal{W}_s(Q,P) := \inf_{X\sim Q, Z\sim P} \mathbb{E}[|X-Z|^s]^{1/s}$. Note the analogy to the Total Variation distance, $\mathrm{TV}(Q,P) := \inf_{X\sim Q, Z\sim P} \mathbb{E}[\mathbb{1}_{X\neq Z}]$.

These distances allow a variational characterization. For simplicity restrict to $s=1$.

$$\mathcal{W}_1(Q,P) = \sup_{\substack{f \text{ Lipschitz} \\ [f]_{\mathrm{Lip}}=1}} |\mathbb{E}_Q f(X) - \mathbb{E}_P f(Z)|,$$
$$\mathrm{TV}(Q,P) = \sup_{\|f\|_\infty \leq 1} |\mathbb{E}_Q f(X) - \mathbb{E}_P f(Z)|.$$

Suppose we wish to estimate $d_{\mathcal{H}}(Q,P)$, where

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

2 Stein's method

Now we describe Stein's method (Section 2). Let $(\mathcal{T}, \mathcal{D}(\mathcal{T}) = \mathcal{G})$ denote an operator (the *Stein operator*) that maps functions $g : \mathbb{R}^d \to \mathbb{R}^d$, $g \in \mathcal{G}$ into real-valued mean-zero functions under P, i.e.

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

Then the *Stein discrepancy* is defined as

$$\begin{aligned}\mathcal{S}(Q,\mathcal{T},\mathcal{G}) &:= \sup_{g\in\mathcal{G}} \|\mathbb{E}_Q[\mathcal{T}g(X)]\| \\ &= \sup_{g\in\mathcal{G}} \|\mathbb{E}_Q[\mathcal{T}g(X)] - \mathbb{E}_P[\mathcal{T}g(Z)]\| = d_{\mathcal{T}\mathcal{G}}(Q,P).\end{aligned}$$

The operator $\mathcal{T}$ and domain $\mathcal{G}$ should satisfy (or should be designed such that): for every $h \in \mathcal{H}$, there exists a function $g = g_h \in \mathcal{G}$ such that the *Stein equation*

$$h(x) - \mathbb{E}_P[h(Z)] = \mathcal{T}g_h(x), \quad x \in \mathbb{R}^d \tag{1}$$

is satisfied.

We then immediately have

$$d_{\mathcal{H}}(Q,P) = \sup_{h\in\mathcal{H}} |\mathbb{E}_Q h(X) - \mathbb{E}_P h(Z)| \leq \sup_{g\in\mathcal{G}} |\mathbb{E}_Q \mathcal{T}g(X) - \mathbb{E}_P \mathcal{T}g(Z)| = \sup_{g\in\mathcal{G}} |\mathbb{E}_Q \mathcal{T}g(X)| = \mathcal{S}(Q,\mathcal{T},\mathcal{G}).$$

Note that the right hand side only depends on P through $\mathcal{T}$ and $\mathcal{G}$. Of course this is only useful if we can find good bounds for $\mathcal{S}(Q,\mathcal{T},\mathcal{G})$.

2.1 Using Itô diffusions to identify a Stein operator

In [GDVM19] several different diffusions are proposed, whereas in [GM15] only the Langevin diffusion is used. We will also restrict to the latter case.

Consider the Langevin diffusion,

$$dZ_{t,x} = \frac{1}{2}\nabla \log p(Z_t, x)\, dt + dW_t, \quad t \geq 0, \quad Z_{0,x} = x, \tag{2}$$

where (W_t) is a standard Brownian motion. This diffusion has $p(x)$ as stationary density. Let P denote the corresponding probability distribution. The generator of this diffusion is given by

$$\mathcal{A}u(x) = \tfrac{1}{2}\langle \nabla u(x), \nabla \log p(x)\rangle + \tfrac{1}{2}\Delta u(x).$$

Note that

$$\mathcal{A}u(x) = \mathcal{T}(\tfrac{1}{2}\nabla u)(x),$$

where

$$\mathcal{T}g(x) = \langle g(x), \nabla \log p(x)\rangle + (\nabla \cdot g)(x).$$

As *Stein set* the authors choose the '*classical Stein set*',

$$\mathcal{G} := \left\{ g : \mathbb{R}^d \to \mathbb{R}^d : \sup_{x \neq y} \max\left(\|g(x)\|, \|\nabla g(x)\|, \frac{\|\nabla g(x) - \nabla g(y)\|}{\|x - y\|} \right) \leq 1 \right\}.$$

Proposition 2.1. *If $\mathbb{E}_P\|\nabla \log p(Z)\| < \infty$, then $\mathbb{E}_P \mathcal{T}g(Z) = 0$ for all $g \in \mathcal{G}$.*

Proof. Follows by integration by parts. □

3 Lower bounding the Stein discrepancy

In [GDVM19] the authors consider quite general conditions, but here, as in [GM15, MG16] we just assume log concavity of $p(x)$.

Theorem 3.1. *Suppose $\log p \in C^4(\mathbb{R}^d)$ is k-strongly concave, i.e.*

$$v^\top \nabla^2 \log p(x) v \leq -k\|v\|^2, \quad x, v \in \mathbb{R}^d.$$

with

$$\sup_{x \in \mathbb{R}^d} \|\nabla^3 \log p(z)\| < \infty \quad \textit{and} \quad \sup_{x \in R^d} \|\nabla^4 \log p(z)\| < \infty.$$

Let $(Z_{t,x})_{t \geq 0, x \in \mathbb{R}^d}$ be the Langevin diffusion process as in (2). Then, for each Lipschitz $h \in C^3(\mathbb{R}^d)$,

$$u(x) := \int_0^\infty \mathbb{E}_P[h(Z)] - \mathbb{E}[h(Z_{t,x})]\, dt, \quad x \in \mathbb{R}^d,$$

solves the Poisson equation

$$h(x) - \mathbb{E}_P[h(Z)] = \mathcal{A}u(x)$$

and the first three moments of u can be bounded in terms of the first three moments of h.

Proof. See [MG16, proof of Theorem 2.1]. □

Remark 3.2. One of the achievements of [GDVM19] is a substantial weakening of the condition of strong concavity. △

It follows that $g = \frac{1}{2}\nabla u$ is a solution to the Stein equation (1), and we can bound g and the first two moments of g in terms of the first three moments of h. So *if* h satisfies the conditions of Theorem 3.1, then we can bound

$$\|\mathbb{E}_Q h(X) - \mathbb{E}_P h(Z)\| \precsim \|\mathbb{E}_Q \mathcal{T} g(X)\|.$$

But that is not goed enough: What if h does not satisfy these conditions?

Theorem 3.3. *We have*

$$\mathcal{W}_1(Q,P) \precsim \max\left(\mathcal{S}(Q,\mathcal{T},\mathcal{G}),\, \sqrt[3]{\mathcal{S}(Q,\mathcal{T},\mathcal{G})}\right).$$

Sketch of proof: Suppose h is 1-Lipschitz and $t > 0$. Define

$$h_t(x) = \int_{\mathbb{R}^d} h(x + tz)\phi(z)\, dz,$$

where ϕ is the standard normal CDF. Then for any measure Q,

$$\|\mathbb{E}_Q[h(X) - h_t(X)]\| = \|\mathbb{E}_Q[h(X) - h(X + tG)]\| \le t\mathbb{E}\|G\|,$$

using the Lipschitz property of h, where G is a d-dimensional Gaussian. We can now find useful expressions for $\nabla h_t(x)$, $\nabla^2 h_t(x)$ and $\nabla^3 h_t(x)$, and we can bound (details in [MG16])

$$\begin{aligned}
\sup_{x\in\mathbb{R}^d} \|\nabla h_t(x)\| &\le 1,\\
\sup_{x\in\mathbb{R}^d} \|\nabla^2 h_t(x)| &\le \frac{1}{t}\sqrt{\frac{2}{\pi}},\\
\sup_{x,y\in\mathbb{R}^d, x\neq y} \frac{\|\nabla^2 h_t(x) - \nabla^2 h_t(y)\|}{\|x-y\|} &\le \frac{\sqrt{2}}{t^2}.
\end{aligned}$$

It turns out that

$$b_t := \max(1, (1/t)\sqrt{2/\pi}, \sqrt{2}/t^2) = \max(1, \sqrt{2}/t^2).$$

Now let $X \sim Q$ and $Z \sim P$, such that $\mathcal{W}_1(Q,P) = \mathbb{E}\|X - Z\|$. Then

$$\begin{aligned}
\mathcal{W}_1(Q,P) &\le \sup_{[h]_{\mathrm{Lip}}\le 1} \|\mathbb{E}_Q h(X) - \mathbb{E}_P h(Z)\|\\
&\le \inf_{t>0} \sup_{[h]_{\mathrm{Lip}}\le 1} \{\mathbb{E}_Q\|h(X) - h_t(X)\| + \mathbb{E}_P\|h(Z) - h_t(Z)\| + \|\mathbb{E}_Q h_t(X) - \mathbb{E}_P h_t(Z)\|\}\\
&\le \inf_{t>0} 2t\mathbb{E}\|G\| + b_t \mathcal{S}(Q,\mathcal{T},\mathcal{G}).
\end{aligned}$$

Now optimize over $t > 0$. □

4 Computing the Stein discrepancy

Suppose Q is the empirical distribution of a collection of samples (x_i) with weights $q(x_i)$,

$$Q(dx) = \sum_{i=1}^n q(x_i)\delta_{x_i}(dx).$$

Then

$$\begin{aligned}\mathcal{S}(Q,\mathcal{T},\mathcal{G}) = \sup_{g} \quad & \sum_{i=1}^{n} q_i \left(\langle g(x_i), \nabla \log p(x_i)\rangle + (\nabla \cdot g(x_i))\right) \\ \text{s.t.} \quad & g \in \mathcal{G},\end{aligned}$$

where $\mathcal{G}$ is the classical Stein set,

$$\begin{aligned}\mathcal{G} = \big\{ g : \mathbb{R}^d \to \mathbb{R}^d : & \|g(x)\| \le 1, \\ & \|\nabla g(x)\| \le 1, \\ & \|\nabla g(x) - \nabla g(y)\| \le \|x-y\|, \quad \text{for all} \quad x, y \in \mathbb{R}^d \big\}.\end{aligned}$$

This optimization problem requires optimizing over functions subject to infinitely many constraints.

We can simplify this by introducing the *graph Stein set*,

$$\begin{aligned}\mathcal{G}_G := \big\{ g : \mathbb{R}^d \to \mathbb{R}^d : & \|g(x)\| \le 1, \|\nabla g(x)\| \le 1, \\ & \|g(x_i) - g(x_j)\| \le \|x_i - x_j\|, \\ & \|\nabla g(x_i) - \nabla g(x_j)\| \le \|x_i - x_j\|, \\ & \|g(x_i) - g(x_j) - \nabla g(x_i)(x_i - x_j)\| \le \tfrac{1}{2}\|x_i - x_j\|^2, \\ & \|g(x_i) - g(x_j) - \nabla g(x_j)(x_i - x_j)\| \le \tfrac{1}{2}\|x_i - x_j\|^2, \quad \text{for all} \quad i = 1, \dots, n \big\}.\end{aligned}$$

Then $\mathcal{G} \subset \mathcal{G}_G$, and so $\mathcal{S}(Q,\mathcal{T},\mathcal{G}) \le \mathcal{S}(Q,\mathcal{T},\mathcal{G}_G)$. Remarkably, there exists a constant κ_d depending only upon dimension (and the norms used) such that

$$\mathcal{S}(Q,\mathcal{T},\mathcal{G}_G) \le \kappa_d \mathcal{S}(Q,\mathcal{T},\mathcal{G}).$$

Finally we may write $g_k(x_i) = \gamma_{ik}$ and $\partial_k g_j(x_i) = \Gamma_{jki}$ and optimize over $(\gamma_{ik}, \Gamma_{jki})$. This results in a linear programming problem.

References

[GDVM19] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, 2019.

[GM15] Jackson Gorham and Lester Mackey. Measuring Sample Quality with Stein's Method. *Advances in Neural Information Processing Systems 28*, pages 226–234, 2015.

[MG16] Lester Mackey and Jackson Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21:1–14, 2016.