KERNELISED STEIN DISCREPANCY

LIAM HODGKINSON

Objective: Suppose that p is a target density with an unknown normalising constant (e.g. posterior, model). Determine the degree to which a sample $X_1, \dots, X_n$ represents p, that is, how close is the approximation

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \mathbb{E}_p f(X) = \int f(x)\, p(x)\, dx.$$

Can anyone think of a way (outside of what I'm going to talk about) to do this? I would expect not. This is perhaps one of the biggest problems facing computational statisticians.

- What about kernel density estimation? Take a KDE of the points and compare the resulting density with π. **Problem:** π is unnormalised, so we have no means to compare them.

The existing forms of comparing distributions, for example, Kullback-Leibler divergence:

$$d_{\mathrm{KL}}(\mu|\pi) = \int \mu(x) \log\left(\frac{\mu(x)}{\pi(x)}\right) dx,$$

are not computable. The kernelised Stein discrepancy is inspired by an incredibly powerful technique from analytical probability known as *Stein's method*. In our situation, the method builds upon the following fact: let s_p denote the score function

$$s_p(x) = \nabla \log p(x),$$

Observe that s_p can be computed without knowledge of the normalising constant. Indeed, if $\pi(x) = cp(x)$, then

$$s_\pi(x) = \nabla[\log c + \log p(x)] = \nabla \log p(x) = s_p(x).$$

If $f : \mathbb{R}^d \to \mathbb{R}^d$ is any (differentiable) function, then

$$\mathbb{E}_p[\nabla \cdot f(X) + s_p(X) \cdot f(X)] = 0.$$

This is simply integration by parts. Expanding it out gives

$$\sum_{i=1}^{d} \int p(x) \frac{\partial}{\partial x_i} f(x) + f(x) \frac{\partial}{\partial x_i} p(x)\, dx = \sum_{i=1}^{d} \int \frac{\partial}{\partial x_i}[f(x)\, p(x)]\, dx = 0.$$

Therefore, by defining the *Stein discrepancy*

$$\mathbb{S}(q|p) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_q[\nabla \cdot f(X) + s_p(X) \cdot f(X)]\right)^2,$$

we note that $\mathbb{S}(q|p) = 0$ if $q = p$. Actually, if $\mathcal{F}$ is large enough, you can show that $\mathbb{S}(q|p) = 0$ if and only if $q = p$, and so $\mathbb{S}$ is a valid form of discrepancy. We still cannot compute $\mathbb{S}$ however, since it involves a supremum over a class of test functions. However, if we take $\mathcal{F}$ to be the unit ball $\{h : \|h\|_H \le 1\}$ of a reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel k, then by defining the *Stein kernel*

$$k_p(x, y) = \nabla_x \cdot \nabla_y k(x, y) + s_p(x) \cdot \nabla_y k(x, y) + s_q(x) \cdot \nabla_x k(x, y) + [s_p(x) \cdot s_p(y)]\, k(x, y)$$

there is the *kernelised Stein discrepancy*

$$\mathbb{S}(q|p) = \mathbb{E} k_p(X, Y),$$

where X and Y are iid random variables from q. Therefore, the Stein discrepancy between (the empirical distribution of) a sample $X_1, \ldots, X_n$ and the normalised density p is

$$\mathbb{S}\left(X_1, \ldots, X_n | p\right) = \frac{1}{n^2} \sum_{i,j=1}^{n} k_p\left(X_i, X_j\right),$$

which is easily computable. Gorham and Mackey determined that a good kernel k that ensures the discrepancy is convergence-determining is the IMQ kernel

$$k\left(x, y\right) = \frac{1}{\sqrt{1 + \left\|x - y\right\|^2}}.$$

Even better, the mean squared error for test functions in $\mathcal{H}$ is bounded above by this discrepancy. Indeed,

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} h\left(X_i\right) - \int h\left(x\right) p\left(x\right) dx\right|^2 \leq \mathbb{S}\left(X_1, \ldots, X_n | p\right), \qquad \text{for any } h \in \mathcal{H}.$$