Stein Variational Gradient Descent

Alex Loomis

September 27, 2022

Stein’s Method

Stein's Lemma for Normal Distributions

Lemma
Given $X \sim \mathcal{N}(\mu, \sigma^2)$*,*

$$\mathrm{E}[(X-\mu)\phi(X)] = \sigma^2 \mathrm{E}[\phi'(X)]$$

for every g for which both sides exist. The converse holds as well; if X *satisfies the relation for all* ϕ*, then* $X \sim \mathcal{N}(\mu, \sigma^2)$*.*

This is proven in [Ste86].

Stein Operators

Fix P a probability distribution. An operator A is a Stein operator for P, if for all ϕ,

$$\mathrm{E}[(A\phi)(X)] = 0 \quad \text{if and only if} \quad X \sim P.$$

Example

A Stein operator for the standard normal distribution is $(A\phi)(x) = \phi'(x) - x\phi(x)$.

Stein's Equation

Given a function h, choose a function ϕ_h satisfying

$$(A\phi_h)(x) = h(x) - \mathrm{E}[h(X)],$$

where $X \sim P$. This is the Stein equation for the Stein operator A.

Example

The Stein equation for the given operator for the standard normal distribution is

$$\phi_h'(x) - x\phi_h(x) = h(x) - \mathrm{E}[h(X)].$$

This can be solved explicitly for the Stein solution ϕ_h,

$$\phi_h(x) = e^{\frac{1}{2}x^2} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} \left(h(t) - \mathrm{E}[h(X)]\right) \, dt.$$

Absolutely Continuous Densities

Example

A distribution with an absolutely continuous density f has a Stein operator

$$(A\phi)(x) = \phi'(x) + (\log f)'(x)\phi(x)$$

since

$$\int_{-\infty}^{\infty} \big(\phi'(x) + (\log f)'(x)\phi(x)\big) f(x)\, dx = \int_{-\infty}^{\infty} (f\phi)'(x)\, dx = 0.$$

Absolutely Continuous Densities *(cont.)*

This Stein operator has Stein equation

$$\phi_h'(x) + (\log f)'(x)\phi_h(x) = h(x) - \mathrm{E}[h(X)].$$

and Stein solution

$$\phi_h(x) = \frac{1}{f(x)} \int_{-\infty}^{x} f(t)(h(t) - \mathrm{E}[h(X)])\, dt.$$

Difference in Expectations

Suppose P is a distribution we wish to approximate by the distribution Q. Let $X \sim P$ and $Y \sim Q$. Plugging in $x = Y$ into the Stein equation for P, and taking the expectation of both sides yeilds

$$\mathrm{E}[h(Y)] - \mathrm{E}[h(X)] = \mathrm{E}[(A\phi_h)(Y)].$$

Metrics

By restricting h to different classes of functions, we derive an expression for various metrics.

Example
If $\mathcal{H}$ is the set of half-line indicator functions on $\mathbb{R}$,

$$d_{\mathrm{Kol}}(P,Q) = \sup_{h\in\mathcal{H}} \mathrm{E}[h(Y)] - \mathrm{E}[h(X)] = \sup_{h\in\mathcal{H}} \mathrm{E}[(A\phi_h)(Y)].$$

Metrics

By restricting h to different classes of functions, we derive an expression for various metrics.

Example
If $\mathcal{H}$ is the set of indicator functions on $\mathbb{R}$,

$$d_{\mathrm{TV}}(P, Q) = \sup_{h \in \mathcal{H}} \mathrm{E}[h(Y)] - \mathrm{E}[h(X)] = \sup_{h \in \mathcal{H}} \mathrm{E}[(A\phi_h)(Y)].$$

Metrics

By restricting h to different classes of functions, we derive an expression for various metrics.

Example

If $\mathcal{H}$ is the set of 1-Lipschitz functions on $\mathbb{R}$,

$$d_{\mathrm{Was}}(P,Q) = \sup_{h\in\mathcal{H}} \mathrm{E}[h(Y)] - \mathrm{E}[h(X)] = \sup_{h\in\mathcal{H}} \mathrm{E}[(A\phi_h)(Y)].$$

Choices

Choice of Stein Operator

Let p be a C^1 density on a subset of $\mathbb{R}^d$. We will choose

$$\mathcal{A}_p\phi(x) = \nabla_x \log p(x)\phi(x)^{\mathsf{T}} + \nabla_x\phi(x)$$

as our Stein operator. If $d = 1$, this is the choice

$$(A\phi)(x) = \phi'(x) + (\log f)'(x)\phi(x)$$

mentioned earlier.

Choice of Metric

In order for the optimization algorithm to have a closed form solution, we will choose

$$d(p, q) = \max_{\phi \in \mathcal{H}^d} \left\{ \mathsf{E}[\mathrm{Tr}(\mathcal{A}_p \phi(Y))] \mid \|\phi\|_{\mathcal{H}^d} \leq 1 \right\},$$

where $\mathcal{H}^d$ is a reproducing kernel Hilbert space with kernel $k(\cdot, \cdot)$.

Reproducing Kernel Hilbert Space

Given a positive definite kernel $k : A \times A \to \mathbb{R}$, the reproducing kernel Hilbert space $\mathcal{H}$ of k is the closure of the span of k, along with a certain inner product:

$$\text{Span}(k) = \left\{ \sum_{i=1}^{n} a_i k(\cdot, x_i) \right\}$$

$$\mathcal{H} = \overline{\text{Span}(k)}$$

$$<f, g>_{\mathcal{H}} = \sum_{i,j} a_i b_j k(x_i, x_j).$$

Results

Summary of Result

Variational inference approximates a target distribution p using a distribution q from a set Q of simpler distributions that minimizes the KL divergence $\mathrm{KL}(q \mid p)$.

We will choose Q to be the set of distributions of random variables $T(Y)$ where T is a smooth injection, and $Y \sim q$ for some known tractable q.

We will present an algorithm which performs an analog to gradient descent to iteratively choose transformations T that make q more similar to p.

Computing $d(\cdot,\cdot)$

For fixed p, q, the distance

$$d(p,q) = \max_{\phi \in \mathcal{H}^d} \left\{ \mathrm{E}[\mathrm{Tr}(\mathcal{A}_p \phi(Y))] \mid \|\phi\|_{\mathcal{H}^d} \leq 1 \right\}$$

is attained by $\phi(x) = \frac{\phi^*_{q,p}(x)}{\|\phi^*_{q,p}\|_{\mathcal{H}^d}}$, where

$$\phi^*_{q,p}(x) = \mathrm{E}[\mathcal{A}_p k(Y,x)],$$

giving that $d(p,q) = \|\phi^*_{q,p}\|_{\mathcal{H}^d}$

KL Divergence

Theorem
Let $T(x) = x + \varepsilon\phi(x)$ *and* $q_{[T]}$ *be the density of* $T(Y)$*. We have that*

$$\nabla_\varepsilon \operatorname{KL}(q_{[T]} \,|\, p) \,|_{\varepsilon=0} = -\mathrm{E}[\operatorname{Tr}(A_p \phi(Y))].$$

This means that in order to minimize the KL divergence between a target distribution p and a chosen distribution q, we want to move in the direction of $\phi^*_{q,p}$.

Working Pointwise

Instead of working with q directly, we can sample points from q, and apply our transformation to those points. This works because given $Y, Y_j \sim q$,

$$\begin{aligned}\phi_{q,p}(x) &\propto \mathrm{E}[\mathcal{A}_p k(Y,x)] \\ &= \mathrm{E}\left[\nabla_x \log p(x) k(Y,x)^\mathsf{T} + \nabla_x k(Y,x)\right] \\ &\approx \frac{1}{n}\sum_{j=1}^{n} k(Y_j,x)\nabla_{Y_j}\log p(Y_j) + \nabla_{Y_j} k(Y_j,x).\end{aligned}$$

This gives an update function that does not rely on q, except through the choice of points.

An Algorithm

This idea results in the following algorithm.

1. Choose a target density p, and a collection of points $\{x_i^0\}_{i=1}^n$.
2. Let $\hat{\phi}_\ell^*(x) = \frac{1}{n}\sum_{j=1}^n k(x_j^\ell, x)\nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x)$.
3. Define recursively $x_i^{\ell+1} = x_i^\ell + \varepsilon_\ell \hat{\phi}_\ell^*(x_i^\ell)$.

References i

Qiang Liu and Dilin Wang. “Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: `https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf`.

Charles Stein. "Approximate Computation of Expectations". In: *Lecture Notes-Monograph Series* 7 (1986), pp. i–164. ISSN: 07492170. URL: `http://www.jstor.org/stable/4355512`.