Operator Variational Inference

Raghav Singhal & Saad Lahlou

March 26, 2018

1 Define Objective

Note, for $\phi(z) = [\phi_1(z_1) \dots \phi_n(z_n)]$, we define the stein operator, $\mathcal{A}_{p(z|x)}$, as follows

$$\mathbb{E}_{q_\lambda(z)}[\mathcal{A}_{p(z|x)}\phi(z)] = \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{1}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_{-i})}\mathbb{E}_{q_\lambda(z_i|z_{-i})}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{2}$$

Now, for a fixed conditional, z_i, we ask the following question,

Proposition 1. *Does $\mathbb{E}_{q_\lambda(z_i)}[f(z_i)] = \mathbb{E}_{q_\lambda(z_{-i})}\mathbb{E}_{q_\lambda(z_i|z_{-i})}[f(z_i)]$ for all f continuos and bounded, here f is a function of only z_i.*

Proof of Proposition 1. Now, note that for any density $p(x, y) = p(x|y)p(y)$ and set $A_x \in \mathcal{B}(X)$, by Fubini's Theorem,

$$\begin{aligned}
\int_Y \int_X \mathbb{I}_{A_x}(x)p(x|y)p(y)dxdy &= \int_Y \int_X \mathbb{I}_{A_x}(x)p(x, y)dydx \\
&= \int_{A_x} \int_Y p(x, y)dydx \text{ (check this step)} \\
&= \int_{A_x} (\int_Y p(x, y)dy)dx \\
&= \int_X \mathbb{I}_{A_x}(x)p(x)dx
\end{aligned}$$

Now, for all $f(x) = \mathbb{I}_{A_x}(x)$ where $A_x \in \mathcal{B}(X)$, we just showed that,

$$\mathbb{E}_{p(y)}\mathbb{E}_{p(x|y)}[f(X)] = \mathbb{E}_{p(x)}[f(X)] \tag{3}$$

As $\forall f : X \to \mathbb{R}$ (Borel measurable & Integrable), there exists an increasing sequence of simple functions, s.t $f_n \to f$ and$f_n \leq f$, Proposition 1 holds for all $f : X \to \mathbb{R}$ Borel measurable and Integrable by the dominated convergence theorem.

$\square$

If Proposition 1 holds, then equation 2 can be simplified to

$$\mathbb{E}_{q_\lambda(z)}[\mathcal{A}_{p(z|x)}\phi(z)] = \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{4}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_{-i})}\mathbb{E}_{q_\lambda(z_i|z_{-i})}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{5}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{6}$$

Now, suppose $\phi(z) = [\phi_1(z_1)\dots\phi_n(z_n)]$, where $\phi_i(z_i) \in \mathcal{H}$ and $\mathcal{H}$ is a scalar-valued Reproducing Kernel Hilbert Space (RKHS).

Questions to tackle now

1. Which Kernel should we select ? So we can detect non-convergence.
2. Is taking the supremum over $\phi(z) \in \prod_{i=1}^{n}\mathcal{H}$ different from taking supremum over individual $\mathcal{H}$, so does the following hold

$$\sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \mathbb{E}[\mathcal{A}\phi(Z)] = \sum_{i=1}^{n}\sup_{\phi_i\in\mathcal{H}} \mathbb{E}[\mathcal{A}^i\phi_i(Z_i)]$$

1.1 Attempt for item 2

Note for a univariate distribution, the optimal ϕ^* is given by $\phi^*(y) = \mathbb{E}_{x\sim q}[\mathcal{A}_p k(x,y)]$. Now, note that

$$\sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \mathbb{E}_{q_\lambda(z)}[\mathcal{A}_{p(z|x)}\phi(z)] = \sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)]$$

$$\leq \sum_{i=1}^{n}\sup_{\phi_i\in\mathcal{H}} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)]$$

Now, all we have to show is that the reverse inequality holds true,

$$\sum_{i=1}^{n}\sup_{\phi_i\in\mathcal{H}} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \leq \sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)]$$

This is probably True. Show Rigorously !

1.2 Final Objective

Assuming the above holds, we take

$$\sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \mathbb{E}_{q_\lambda(z)}[\mathcal{A}_{p(z|x)}\phi(z)] = \sup_{\phi\in\prod_{i=1}^{n}\mathcal{H}} \sum_{i=1}^{n} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{7}$$

$$= \sum_{i=1}^{n}\sup_{\phi_i\in\mathcal{H}} \mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)] \tag{8}$$

Now, note the optimal test function is $\frac{\phi_i^*()}{||\phi_i^*||_{\mathcal{H}}}$, where $\phi_i^*()$ is given by

$$\begin{aligned}\phi_i^*(y) &= \mathbb{E}_{z_i^*\sim q_\lambda(z_i^*)}[\mathcal{A}^i_{p(z_i^*|z^*_{-i},x)}k(z_i^*,y)]\\ &= \mathbb{E}_{z_i^*\sim q_\lambda(z_i^*)}[k(z_i^*,y)\nabla_{z_i^*}\log p(z_i^*|z^*_{i-1},x)+\nabla_{z_i^*}k(z_i^*,y)]\end{aligned}$$

which implies that

$$\begin{aligned}\sup_{\phi\in\prod_{i=1}^n\mathcal{H}}\mathbb{E}_{q_\lambda(z)}[\mathcal{A}_{p(z|x)}\phi(z)] &= \sum_{i=1}^n\sup_{\phi_i\in\mathcal{H}}\mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i(z_i)]\\ &= \sum_{i=1}^n\mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i^*(z_i)]\\ &= \sum_{i=1}^n\mathbb{E}_{q_\lambda(z_i)}\left[\mathcal{A}^i_{p(z_i|z_{-i},x)}\mathbb{E}_{z_i^*\sim q_\lambda(z_i^*)}[\mathcal{A}^i_{p(z_i^*|z^*_{-i},x)}k(z_i^*,z_i)]\right]\end{aligned}$$

Now, for one conditional the final objective is as follows:

$$\mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}^i_{p(z_i|z_{-i},x)}\phi_i^*(z_i)] = \mathbb{E}_{q_\lambda(z_i)}[\phi_i^*(z_i)\nabla_{z_i}\log p(z_i|z_{-i},x)+\nabla_{z_i}\phi_i^*(z_i)] \tag{9}$$

2 Take Unbiased Gradient

Now, we write the gradient of the Langevin-Stein Operator with respect to the variational parameters,

$$\begin{aligned}\nabla_\lambda\mathbb{E}_{q_\lambda(z_i)}[\mathcal{A}_{p(z_i|z_{-i},x)}\phi_i^*(z_i)] &= \nabla_\lambda\mathbb{E}_{q_\lambda(z_i)}[\phi_i^*(z_i)\nabla_{z_i}\log p(z_i|z_{-i},x)+\nabla_{z_i}\phi_i^*(z_i)]\\ &= \nabla_\lambda\mathbb{E}_{q_\lambda(z_i)}[\phi_i^*(z_i)\nabla_{z_i}\log p(z_i|z_{-i},x)]+\nabla_\lambda\mathbb{E}_{q_\lambda(z_i)}[\nabla_{z_i}\phi_i^*(z_i)]\\ &:= \mathcal{L}_\lambda+\mathcal{I}_\lambda\end{aligned}$$

Now, note that integrals of the form above can be evaluated as follows:

$$\begin{aligned}\mathcal{J}_\lambda &= \nabla_\lambda\int_{z_i\in\Omega_i}\int_{z_i^*\in\Omega_i}q_\lambda(z_i)q_\lambda(z_i^*)f(z_i,z_i^*)dz_idz_i^*\\ &= \int_{z_i\in\Omega_i}\int_{z_i^*\in\Omega_i}\nabla_\lambda\big(q_\lambda(z_i)q_\lambda(z_i^*)\big)f(z_i,z_i^*)dz_idz_i^*\\ &= \int_{z_i\in\Omega_i}\int_{z_i^*\in\Omega_i}q_\lambda(z_i)\nabla_\lambda q_\lambda(z_i^*)f(z_i,z_i^*)dz_idz_i^*+\int_{z_i\in\Omega_i}\int_{z_i^*\in\Omega_i}q_\lambda(z_i^*)\nabla_\lambda q_\lambda(z_i)f(z_i,z_i^*)dz_idz_i^*\\ &= \int_{z_i\in\Omega_i}\int_{z_i^*\in\Omega_i}q_\lambda(z_i)q_\lambda(z_i^*)f(z_i,z_i^*)\nabla_\lambda(\log q_\lambda(z_i)+\log q_\lambda(z_i^*))dz_idz_i^*\end{aligned}$$

which implies that

$$\mathcal{I}_\lambda = \nabla_\lambda\mathbb{E}_{q_\lambda(z_i)}\mathbb{E}_{q_\lambda(z_i^*)}[\mathcal{A}^i_{p(z_i^*|z^*_{-i},x)}\nabla_{z_i}k(z_i^*,z_i)] \tag{10}$$

$$= \mathbb{E}_{q_\lambda(z_i)}\mathbb{E}_{q_\lambda(z_i^*)}[\nabla_\lambda(\log q_\lambda(z_i)+\log q_\lambda(z_i^*))\mathcal{A}^i_{p(z_i^*|z^*_{-i},x)}\nabla_{z_i}k(z_i^*,z_i)] \tag{11}$$

And we can similarly evaluate $\mathcal{L}_\lambda$.