15-943-178
Lauro Langosco di Langosco dlauro@student.ethz.ch

Relation between Stein and KL-divergence

Let k be a positive definite kernel on $\mathbb{R}^d$ with RKHS H and write $H^d = H \times \cdots \times H$. Let $\phi \in H^d$, and set $T_\varepsilon = \mathrm{id} + \varepsilon\phi$. Further let q and p be distributions on $\mathbb{R}^d$. Then (as we know from the SVGD paper),

$$-\frac{d}{d\varepsilon}\mathrm{KL}(q_{T_\varepsilon} \parallel p)\Big|_{\varepsilon=0} = E\big[\mathcal{A}_p^T[\phi](x)\big],$$

where $\mathcal{A}$ is the Stein operator. The direction $\phi^* \in H^d$ in which the gradient is maximal is given by $\phi^*_{q,p}$ as defined in the SVGD paper. We have the following:

$$\begin{aligned}
-\sup_{\|\phi\|_{H^d}\le 1}\frac{d}{d\varepsilon}\mathrm{KL}(q_{T_\varepsilon} \parallel p)\Big|_{\varepsilon=0} &= E\big[\mathcal{A}_p^T[\phi^*_{q,p}](x)\big]\cdot\frac{1}{\|\phi^*_{p,q}\|_{H^d}} \\
&= \big\|\phi^*_{p,q}\big\|_{H^d} \\
&= \mathrm{KSD}(q \parallel p)
\end{aligned}$$

In other words, one step of SVGD reduces the KL divergence by approximately $\varepsilon \cdot \mathrm{KSD}(q \parallel p)$, where ε is the step size.

Writing k-KSD for the Stein discrepancy computed using kernel k, this means that in the context of SVGD the k-KSD is best understood as the magnitude of the reduction in KL divergence after taking an SVGD step with kernel k.

In particular, this means that k-KSDs computed using different ks are in fact comparable: they all measure the reduction in (a linear approximation of) the KL divergence after one step of SVGD.

15-943-178
Lauro Langosco di Langosco dlauro@student.ethz.ch

Proposed scheme for learning the kernel parameters

This is a proposal for a 'greedy' algorithm that at each step wants to maximize the reduction in KL-divergence. We want to choose the kernel k such that the k-KSD is maximal before each SVGD step. Concretely:

Initialize particles $X_1, \dots, X_n \sim q$ and kernel bandwidth $h_0 = 1$. Write $\hat{q}$ for the empirical distribution of the current particles $X_1, \dots, X_n$. Then repeat:

1. (Maximize KSD) Update the bandwidth:

$$h_{\text{new}} = h_{\text{old}} + \eta \nabla_h \text{KSD}_{h_{\text{old}}}(\hat{q} \parallel p).$$

2. (Minimize KL) Compute SVGD update step using new bandwidth h_{new}:

$$X_i = X_i + \phi^*_{\hat{q},p}(X_i)$$

I'll set up some experiments and see how that goes. Let me know what you think!