Techinique Review

Following is what I learned about how to extend our framework.

Let $g(x)$ represents the probability density of sampling x from the generator G, let $f(x)/c$ be the target probability density at the point x, where $f(x)$ can be evaluated, but the normalization constant, c, may be unknown. The goal is to learn a generator network G to make sample x with probability $g(x)$ as close to $f(x)/c$ as possible. This can be achieved by minimizing the Kullback–Leibler (KL) divergence.

$$\begin{aligned} \underset{g}{\operatorname{argmin}} D_{\mathbb{KL}}(g, f) &= \underset{g}{\operatorname{argmin}} \mathbb{E}_g \ln \frac{g(x)}{f(x)/c} && (1) \\ &= \underset{g}{\operatorname{argmin}} \mathbb{E}_g \ln g(x) - \mathbb{E}_g \ln f(x) + \ln c && (2) \\ &= \underset{g}{\operatorname{argmin}} \mathbb{E}_g \ln g(x) - \mathbb{E}_g \ln f(x) && (3) \\ &= \underset{g}{\operatorname{argmin}} \; -\mathbb{H}(g) + \mathbb{H}(g, f) && (4) \end{aligned}$$

where $\mathbb{H}(g)$ is entropy of g and $\mathbb{H}(g, f)$ is cross-entropy between g and f. The cross-entropy represents that the cost using g to estimate f; when g is closer to f, the cost becomes lower.

Minimizing $D_{\mathbb{KL}}(g, f)$ means to force g closer to f. Kim and Bengio [KB16] says that it will make the generated sample converge toward one or more local minima on the energy surface; However, the using $\mathbb{H}(g)$ as a regularizer can force the generator G to generate samples that cover even more local minima on the energy surface.

$\mathbb{E}_g \ln f(x)$ can be calculate easily, while $\mathbb{E}_g \ln g(x)$ is difficult to handle. In our paper, we estimate $g(x)$ using KDE but proved not accurate. Fortunately, there are many papers to handle this problem. The following material is quoted from Murphy's book (the download link: https://probml.github.io/pml-book/book2.html)

> the density of $q_{\boldsymbol{\phi}}(\mathbf{x})$ is unknown. Kim and Bengio [KB16] and Zhai et al. [Zha+16] propose several heuristics to approximate this entropy function. Kumar et al. [Kum+19c] propose to estimate the entropy through its connection to mutual information: $H(q_{\boldsymbol{\phi}}(\mathbf{z})) = I(g_{\boldsymbol{\phi}}(\mathbf{z}), \mathbf{z})$, which can be estimated from samples with variational lower bounds [NWJ10b; NCT16b]. Dai et al. [Dai+19a] noticed that when defining $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_0(\mathbf{x}) e^{-E_{\boldsymbol{\theta}}(\mathbf{x})}/Z_{\boldsymbol{\theta}}$, with $p_0(\mathbf{x})$ being a fixed base distribution, the entropy term $-H(q_{\boldsymbol{\phi}}(\mathbf{x}))$ in Equation (25.68) equates $\mathbb{KL}(q_{\boldsymbol{\phi}}(\mathbf{x}) \parallel p_0(\mathbf{x}))$, which can also be approximated with variational lower bounds using samples from $q_{\boldsymbol{\phi}}(\mathbf{x})$ and $p_0(\mathbf{x})$, without requiring the density of $q_{\boldsymbol{\phi}}(\mathbf{x})$.

Reference

[KB16] T. Kim and Y. Bengio. “Deep directed generative models with energy-based probability estimation”. In: arXiv preprint arXiv:1606.03439 (2016).

[Zha+16] S. Zhai, Y. Cheng, R. Feris, and Z. Zhang. “Generative adversarial networks as variational training of energy based models”. In: arXiv preprint arXiv:1611.01799 (2016).

[Kum+19c] R. Kumar, S. Ozair, A. Goyal, A. Courville, and Y. Bengio. “Maximum entropy generators for energy-based mod- els”. In: arXiv preprint arXiv:1901.08508 (2019).

[NWJ10b] X. Nguyen, M. J. Wainwright, and M. I. Jordan. “ Es- timating divergence functionals and the likelihood ratio by con- vex risk minimization”. In: IEEE Transactions on Information Theory 56.11 (2010), pp. 5847–5861.

[NCT16b] S. Nowozin, B. Cseke, and R. Tomioka. “f-gan: Train- ing generative neural samplers using variational divergence minimization”. In: NIPS. 2016, pp. 271–279.

[Dai+19a] B. Dai, H. Dai, A. Gretton, L. Song, D. Schuurmans, and N. He. “Kernel exponential family estimation via doubly dual embedding”. In: AISTATS. PMLR. 2019, pp. 2321–2330.