ICML Notes

1 TVO + IWAE proposal

We can use the connection with renyi divergences to make an iwae variant of the TVO. In appendix C Rob showed

$$\log Z_\beta = \beta(\log p(\mathbf{x}) - D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \tag{1}$$

where $D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})])$ is a Renyi divergence defined

$$D_\alpha[p||q]) = \frac{1}{1-\alpha}\log\int p^\alpha q^{1-\alpha} d\omega. \tag{2}$$

If we consider a discrete path $\gamma = \{\beta_0, \beta_1, ..., \beta_K\}$ with $\beta_0 = 0$ and $\beta_1 = 1$, we observe the sum of the corresponding partition functions $\psi(\beta) := \log Z_\beta$ forms a "dual" bound to $\log p(\mathbf{x})$

$$\sum_{i=0}^{K}\psi(\beta) = \sum_{i=0}^{K}\beta_i(\log p(\mathbf{x}) - D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \tag{3}$$

$$= \log p(\mathbf{x})\left(\sum_{i=0}^{K}\beta_i\right) - \sum_{i=0}^{K}\beta_i D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \tag{4}$$

Defining $\bar{\beta}_\gamma := \sum_{i=0}^{K}\beta_i$ and $\mathcal{L}_{\text{TVO}_{IW}} := \sum_{i=0}^{K}\psi(\beta)$, we therefore have

$$\log p(\mathbf{x}) = \frac{1}{\bar{\beta}_\gamma}\left[\mathcal{L}_{\text{TVO}_{IW}} + \sum_{i=0}^{K}\beta_i D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})])\right] \tag{5}$$

Each partition function can be approximated using an iwae-like estimator

$$\psi(\beta) = \log Z_\beta = \log \mathbb{E}_q\left[\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right)^\beta\right] \approx \log\left[\frac{1}{S}\sum_{s=1}^{S}\left(\frac{p_\theta(\mathbf{x},\mathbf{z}_s)}{q_\phi(\mathbf{z}_s\,|\,\mathbf{x})}\right)^\beta\right] \tag{6}$$

Thus the dual bound to $\log p(\mathbf{x})$, formed by the sum of π_{β_i}'s partition functions, corresponds to moving the summation over samples from outside the log in $\mathcal{L}_{\text{TVO}_L}$, to inside the log in $\mathcal{L}_{\text{TVO}_{IW}}$.

Comments:

- The tightness of $\mathcal{L}_{\text{TVO}_{IW}}$ bound clearly depends on $\bar{\beta}_\gamma$, the sum of the betas along the path γ. Maybe this is something we can exploit to choose the beta path?
- These terms should be reparameterizable in a similar fashion to the iwae estimator. Perhaps we can get around having to show which bound is tighter than which other bound, and instead sell this approach as a technique to use the reparam. trick?
- This loss might be more conducive to using the doubly-reparameterizable gradient estimator (which overcomes the SNR issue for inference networks) than the above reparam approach.

2 Proof that TVO IW is a tighter bound than the elbo

We now show

$$\frac{1}{\bar{\beta}_\gamma}\sum_{i=0}^{K}\beta_i D_{1-\beta}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \leq D_{KL}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \tag{7}$$

Proof. Let $\beta_0, \beta_1, ..., \beta_K$ be an increasing sequence with $\beta_0 = 0$, $\beta_1 = 1$, $\forall \beta_i \in [0, 1]$. Let $d_i = D_{1-\beta_i}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})])$ for notational convenience. Because all d_i, β_i are non-negative, we can write the sum $\sum_{i=0}^{K}\beta_i d_i$ as an L1 norm

$$\begin{aligned}
\sum_{i=0}^{K}\beta_i d_i &= \|\beta^T d\|_1 && (8)\\
&\leq \|\beta\|_1\|d\|_\infty \quad \text{From holder's inequality} && (9)\\
&= \left(\sum_{i=0}^{K}\beta_i\right)\max_i(d_i) && (10)\\
&= \bar{\beta}_\gamma D_{KL}[q(\mathbf{z}\,|\,\mathbf{x})||p(\mathbf{z}\,|\,\mathbf{x})] && (11)
\end{aligned}$$

Where the last line follows from non-decreasing property of renyi divergences Li and Turner (2016) and the identity $D_1[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})] = D_{KL}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]$. Therefore

$$\frac{1}{\bar{\beta}_\gamma}\sum_{i=0}^{K}\beta_i D_{1-\beta_i}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})]) \leq D_{KL}[q_\phi(\mathbf{z}\,|\,\mathbf{x})||p_\theta(\mathbf{z}\,|\,\mathbf{x})] \tag{12}$$

□

References

Yingzhen Li and Richard E. Turner. Renyi Divergence Variational Inference. *arXiv:1602.02311 [cs, stat]*, October 2016. URL `http://arxiv.org/abs/1602.02311`. arXiv: 1602.02311.