Appendix for Variational Generative Stochastic Networks with Collaborative Shaping

Philip Bachman and Doina Precup
School of Computer Science, McGill University

Explaining the Variational Free-Energy

Given distributions $p_\theta(x|z)$, $q_\phi(z|x)$, and $p_*(z)$, we can define several *derived distributions*:

$$p_\theta(x;p_*)=\sum_z p_\theta(x|z)p_*(z) \tag{1}$$

$$p_\theta(z|x;p_*)=\frac{p_\theta(x|z)p_*(z)}{p_\theta(x;p_*)} \tag{2}$$

$$p_\theta(x,z;p_*)=p_\theta(x|z)p_*(z) = p_\theta(z|x;p_*)p_\theta(x;p_*) \tag{3}$$

Given these distributions, we now work "backwards" from $\log p_\theta(x;p_*)$:

$$\log \quad p_\theta(x;p_*) = \sum_z q_\phi(z|x)\log p_\theta(x;p_*) \tag{4}$$

$$= \quad \sum_z q_\phi(z|x)\log\frac{p_\theta(z|x;p_*)p_\theta(x;p_*)}{p_\theta(z|x;p_*)} \tag{5}$$

$$= \quad \sum_z q_\phi(z|x)\log\frac{p_\theta(x,z;p_*)}{p_\theta(z|x;p_*)} \tag{6}$$

$$= \quad \sum_z q_\phi(z|x)(\log p_\theta(x,z;p_*) - \log q_\phi(z|x) + \log q_\phi(z|x) - \log p_\theta(z|x;p_*)) \tag{7}$$

$$= \quad \sum_z q_\phi(z|x)\log\frac{p_\theta(x,z;p_*)}{q_\phi(z|x)} + \sum_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x;p_*)} \tag{8}$$

$$= \quad \sum_z q_\phi(z|x)\left(\log p_\theta(x|z) + \log\frac{p_*(z)}{q_\phi(z|x)}\right) + \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x;p_*)) \tag{9}$$

$$\geq \quad \sum_z q_\phi(z|x)\log p_\theta(x|z) - \mathrm{KL}(q_\phi(z|x)||p_*(z)) \tag{10}$$

$$\geq \quad -\mathcal{F}(x;q_\phi,p_\theta,p_*) \tag{11}$$

where Eqns. 10-11 define the variational free-energy:

$$\begin{aligned}
\mathcal{F}(x; q_\phi, p_\theta, p_*) &= -\sum_z q_\phi(z|x) \log p_\theta(x|z) & (12)\\
&\quad\quad + \mathrm{KL}(q_\phi(z|x) || p_*(z)) \\
&\geq -\log p_\theta(x; p_*) & (13)
\end{aligned}$$

These equations follow from simple algebraic manipulation and Eqn. 9-10 comes from non-negativity and definition of KL. In this derivation of the variational free-energy, we treated p_θ, q_ϕ, and p_* simply as computational mechanisms for producing valid distributions over the appropriate spaces. This emphasizes the fact that, for any triplet of distributions (p_θ, q_ϕ, p_*), $\mathcal{F}(x; q_\phi, p_\theta, p_*)$ can be computed and gives a lower-bound on $\log p_\theta(x; p_*)$ for the *derived distribution* $p_\theta(x; p_*)$.

Note that the derived distributions we used result strictly from interactions between $p_\theta(x|z)$ and $p_*(z)$, and are independent of $q_\phi(z|x)$. Therefore, we could change the domain of $q_\phi(z|x)$ to some alternate, arbitrary space $\mathcal{Y}$ such that $q_\phi(z|y)$ produces distributions over $\mathcal{Z}$ given inputs from $\mathcal{Y}$. Plugging such a q_ϕ into Eqn. 12 (and substituting some ys for some xs appropriately) still produces a valid free-energy $\mathcal{F}(x, y; q_\phi, p_\theta, p_*)$, which still upper-bounds the negative log-likelihood of x under $p_\theta(x; p_*)$. We refer to the resulting system comprising $q_\phi(z|y)$, $p_\theta(x|z)$, and $p_*(z)$ as a variational transcoder.

Variational transcoding is a very general mechanism which encompasses standard variational auto-encoders, where $\mathcal{Y} = \mathcal{X}$, and methods for sequence-to-sequence learning, image-to-text generation, Bayesian classification, etc. E.g., the sequence-to-sequence learning method in [**?**] can be interpreted as a variational transcoder in which $p_\theta(x|z)/q_\theta(z|y)$ are constructed from LSTMs, $p_*(z)$ is, e.g., an isotropic Gaussian distribution over $\mathcal{Z}$, and the distributions output by $q_\phi(z|y)$ are fixed to be Dirac deltas over $\mathcal{Z}$.