Variational autoencoder and friends

Xun Zheng
xzheng1@andrew.cmu.edu

1 Variational autoencoder and friends

- Generative model (decoder):

$$z \sim P_0(Z) = \mathsf{N}(Z; 0, I) \tag{1}$$
$$x \sim P_\theta(X|z) = \mathsf{N}\big(X; \mu_\theta(z), I\big) \tag{2}$$

  Generative mutual information:

$$\mathsf{I}_{dec}(X; Z) = \mathsf{D}\big(P_{dec}(X, Z) \| P_{dec}(X) P_{dec}(Z)\big) \tag{3}$$

  where

$$P_{dec}(X, Z) = P_\theta(X|Z) P_0(Z) \tag{4}$$
$$P_{dec}(X) = \int P_{dec}(X, z)\, \mathrm{d}z = \underbrace{\mathsf{E}_{z \sim P_0(Z)}\big[P_\theta(X|z)\big]}_{\text{marginal likelihood}} \tag{5}$$
$$P_{dec}(Z) = \int P_{dec}(x, Z)\, \mathrm{d}x = P_0(Z) \tag{6}$$

- Variational posterior takes the form of inference network (encoder):

$$Q_\phi(Z|x) = \mathsf{N}\big(Z; \mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x))\big) \tag{7}$$

  Representative mutual information:

$$\mathsf{I}_{enc}(X; Z) = \mathsf{D}\big(P_{enc}(X, Z) \| P_{enc}(X) P_{enc}(Z)\big) \tag{8}$$

  where

$$P_{enc}(X, Z) = Q_\phi(Z|X) P_{data}(X) \tag{9}$$
$$P_{enc}(X) = \int P_{enc}(X, z)\, \mathrm{d}z = P_{data}(X) \tag{10}$$
$$P_{enc}(Z) = \int P_{enc}(x, Z)\, \mathrm{d}x = \underbrace{\mathsf{E}_{x \sim P_{data}(X)}\big[Q_\phi(Z|x)\big]}_{\text{aggregated posterior}} \tag{11}$$

- Evidence lower bound for a single sample:

$$\mathrm{ELBO}(x) = \log P_{dec}(x) - \mathsf{D}\big(Q_\phi(Z|x) \| \underbrace{P_{dec}(Z|x)}_{\text{true posterior}}\big) \tag{12}$$
$$= \mathsf{E}_{z \sim Q_\phi(Z|x)}\left[\log \frac{P_{dec}(x, z)}{Q_\phi(z|x)}\right] \tag{13}$$

Minimize the negative ELBO objective:

$$L(\theta, \phi) = \mathsf{E}_{x \sim P_{data}(X)}\big[-\text{ELBO}(x)\big] \tag{14}$$

$$= \underbrace{\mathsf{E}_{x \sim P_{data}(X)} \mathsf{E}_{z \sim Q_\phi(Z|x)}\big[-\log P_\theta(x|z)\big]}_{\text{negative expected log-likelihood } (\mathit{Nell})} + \underbrace{\mathsf{E}_{x \sim P_{data}(X)} \mathsf{D}\big(Q_\phi(Z|x) \| P_0(Z)\big)}_{\text{posterior/prior divergence } (\mathit{Div})} \tag{15}$$

- In the case of Gaussian prior and posterior, the divergence has a closed form:

$$\mathit{Div} = \mathsf{E}_{x \sim P_{data}(X)}\left[\frac{1}{2} \cdot \mathbf{1}^T\big(\sigma_\phi^2(x) + \mu_\phi^2(x) - \mathbf{1} - \log \sigma_\phi^2(x)\big)\right] \tag{16}$$

- Posterior/prior divergence can be decomposed (Kim and Mnih, 2018):

$$\mathit{Div} = \mathsf{E}_{x \sim P_{data}(X)} \mathsf{E}_{z \sim Q_\phi(Z|x)}\left[\log \frac{Q_\phi(z|x)}{P_0(z)} \frac{P_{enc}(z)}{P_{enc}(z)}\right] \tag{17}$$

$$= \mathsf{E}_{x \sim P_{data}(X)} \mathsf{E}_{z \sim Q_\phi(Z|x)}\left[\log \frac{Q_\phi(z|x)}{P_{enc}(z)} \frac{P_{data}(x)}{P_{data}(x)}\right] + \mathsf{E}_{z \sim P_{enc}(Z)}\left[\log \frac{P_{enc}(z)}{P_0(z)}\right] \tag{18}$$

$$= \mathsf{I}_{enc}(X; Z) + \underbrace{\mathsf{D}\big(P_{enc}(Z) \| P_0(Z)\big)}_{\text{agg-posterior/prior divergence } (\mathit{Aggdiv})} \tag{19}$$

- Agg-posterior/prior term above can be further decomposed (Chen et al., 2018; Esmaeili et al., 2019):

$$\mathit{Aggdiv} = \mathsf{E}_{z \sim P_{enc}(Z)}\left[\log \frac{P_{enc}(z)}{P_0(z)} \frac{\prod_j P_{enc}(z_j)}{\prod_j P_{enc}(z_j)} \frac{\prod_j P_0(z_j)}{\prod_j P_0(z_j)}\right] \tag{20}$$

$$= \underbrace{\mathsf{D}\big(P_{enc}(Z) \| \prod_j P_{enc}(Z_j)\big)}_{\text{total correlation } (\mathsf{TC})} + \underbrace{\sum_j \mathsf{D}\big(P_{enc}(Z_j) \| P_0(Z_j)\big)}_{\text{dimension-wise KL}} - \underbrace{\mathsf{E}_{z \sim P_{enc}(Z)}\left[\log \frac{P_0(z)}{\prod_j P_0(z_j)}\right]}_{\text{prior factorization}} \tag{21}$$

where $P_0(Z_j), P_{enc}(Z_j)$ are singleton marginals of $P_0(Z), P_{enc}(Z)$ respectively. The last term disappears in (Chen et al., 2018) as it assumes independent prior: $P_0(Z) = \prod_j P_0(Z_j)$.

- Monte Carlo estimate of agg-posterior entropy:

$$\mathsf{H}_{agg}(Z) = \mathsf{E}_{z \sim P_{enc}(Z)}\big[-\log P_{enc}(z)\big] \tag{22}$$

$$= \mathsf{E}_{x \sim P_{data}(X)} \mathsf{E}_{z \sim Q_\phi(Z|x)}\big[-\log \mathsf{E}_{x' \sim P_{data}(X)} Q_\phi(z|x')\big] \tag{23}$$

  - Monte Carlo estimate using full dataset:

$$\mathsf{H}_{agg}(Z) \approx \frac{1}{N} \sum_{n=1}^{N} -\log\left(\frac{1}{N} \sum_{n'=1}^{N} Q_\phi(z^{(n)}|x^{(n')})\right), \quad z^{(n)} \sim Q_\phi(Z|x^{(n)}) \tag{24}$$

    Cross-evaluate N posterior samples against N posterior distributions, hence the complexity is $\mathcal{O}(N^2)$. A mini-batch version is straightforward.

  - Partially stratified sampling (Esmaeili et al., 2019):

$$\mathsf{H}_{agg}(Z) \approx \frac{1}{B} \sum_{b=1}^{B} -\log P_{enc}(z^{(b)}) \tag{25}$$

Suppose the inner term is estimated using the full dataset, and separate out diagonal ($n' = b$) and off-diagonal terms ($n' \neq b$):

$$P_{enc}(z^{(b)}) \approx \frac{1}{N}\sum_{n'=1}^{N} Q_\phi(z^{(b)}|x^{(n')}) \tag{26}$$

$$= \frac{1}{N} Q_\phi(z^{(b)}|x^{(b)}) + \frac{1}{N}\sum_{n'\neq b} Q_\phi(z^{(b)}|x^{(n')}) \tag{27}$$

The second term can be estimated from a mini-batch:

$$\frac{1}{N-1}\sum_{n'\neq b} Q_\phi(z^{(b)}|x^{(n')}) \approx \frac{1}{B-1}\sum_{b'\neq b} Q_\phi(z^{(b)}|x^{(b')}) \tag{28}$$

Hence the final estimator is

$$\mathsf{H}_{agg}(Z) \approx \frac{1}{|B|}\sum_{b\in B} -\log\left(\frac{1}{N} Q_\phi(z^{(b)}|x^{(b)}) + \frac{N-1}{N}\frac{1}{B-1}\sum_{b'\neq b} Q_\phi(z^{(b)}|x^{(b')})\right) \tag{29}$$

- Bounds for $\mathsf{I}_{enc}(X;Z)$:

$$\mathsf{H}_{data} - Nell \leq \mathsf{I}_{enc}(X;Z) \leq Div \tag{30}$$

Lower bound (Alemi et al., 2018, Appendix D.1):

$$\mathsf{I}_{enc}(X;Z) \geq \mathsf{I}_{enc}(X;Z) - \mathsf{E}_{z\sim P_{enc}(Z)}\mathsf{D}\big(P_{enc}(X|z)\|P_\theta(X|z)\big) \tag{31}$$

$$= \mathsf{E}_{x,z\sim P_{enc}(X,Z)}\left[\log\frac{P_{enc}(x|z)}{P_{enc}(x)}\right] - \mathsf{E}_{x,z\sim P_{enc}(X,Z)}\left[\log\frac{P_{enc}(x|z)}{P_\theta(x|z)}\right] \tag{32}$$

$$= \mathsf{H}\big[P_{data}(X)\big] + \mathsf{E}_{x\sim P_{data}(X)}\mathsf{E}_{z\sim Q_\phi(Z|x)}\big[\log P_\theta(x|z)\big] \tag{33}$$

$$= \mathsf{H}_{data} - Nell \tag{34}$$

Upper bound (Alemi et al., 2018, Appendix D.2):

$$\mathsf{I}_{enc}(X;Z) \leq \mathsf{I}_{enc}(X;Z) + Aggdiv = Div \tag{35}$$

- Viewing marginal likelihood as expectation, one can substitute the prior with other distributions using importance sampling trick:

$$P_{dec}(x) = \mathsf{E}_{z\sim P_0(Z)}\big[P_\theta(x|z)\big] \tag{36}$$

$$= \mathsf{E}_{z\sim Q_\phi(Z|x)}\big[P_\theta(x|z)R(x,z)\big], \quad R(x,z) = \frac{P_0(z)}{Q_\phi(z|x)} \tag{37}$$

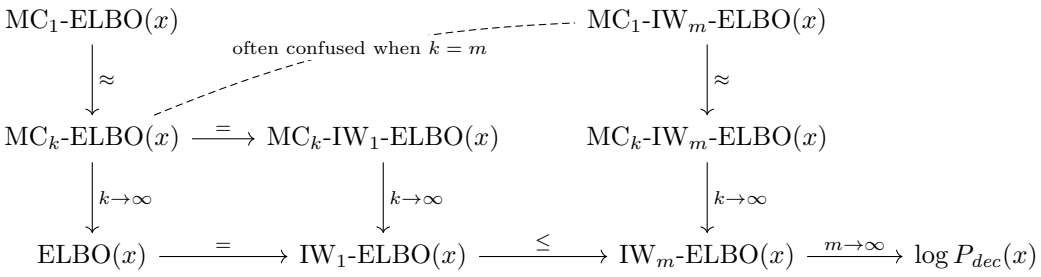
$$= \mathsf{E}_{z^{(1)},\ldots,z^{(m)}\sim Q_\phi(Z|x)}\left[\frac{1}{m}\sum_{i=1}^{m} P_\theta(x|z^{(i)})R(x,z^{(i)})\right] \tag{38}$$

The last line may appear to be an unnecessary complication of the second line, however it makes a difference when Jensen's inequality is applied:

$$\log P_{dec}(x) \geq \tag{39}$$

$$\begin{cases} \underbrace{\mathsf{E}_{z\sim Q_\phi(Z|x)}\big[\log\big(P_\theta(x|z)R(x,z)\big)\big]}_{\text{ELBO}(x)} \approx \underbrace{\frac{1}{k}\sum_{j=1}^{k}\log\big(P_\theta(x|z^{(j)})R(x,z^{(j)})\big)}_{\text{MC}_k\text{-ELBO}(x)} \\ \underbrace{\mathsf{E}_{z^{(1)},\ldots,z^{(m)}\sim Q_\phi(Z|x)}\big[\log\big(\frac{1}{m}\sum_{i=1}^{m} P_\theta(x|z^{(i)})R(x,z^{(i)})\big)\big]}_{\text{IW}_m\text{-ELBO}(x)} \approx \underbrace{\frac{1}{k}\sum_{j=1}^{k}\log\big(\frac{1}{m}\sum_{i=1}^{m} P_\theta(x|z^{(j,i)})R(x,z^{(j,i)})\big)}_{\text{MC}_k\text{-IW}_m\text{-ELBO}(x)} \end{cases} \tag{40}$$

The quantities are related as follows:



Also notice that one can no longer use closed form of *Div* in IW$_m$-ELBO(x) as typically done in VAEs, since log is taken after the averaging over importance samples.

References

A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy. Fixing a Broken ELBO. In *International Conference on Machine Learning*, 2018.

R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent. Structured Disentangled Representations. In *International Conference on Artificial Intelligence and Statistics*, 2019.

H. Kim and A. Mnih. Disentangling by Factorising. In *International Conference on Machine Learning*, 2018.