Fixing a broken ELBO

Alemi, et al. 2018

Motivation

- Progress in VAEs have led to good generative performance.
- Training via ELBO does not necessarily lead to good representation performance.
- This work focuses on developing losses that give good representations.

Mutual information based loss

- “Good” representation is analogous to higher mutual information between the observations and the latents

$$\mathrm{I_e}(X;Z) = \iint dx\, dz\, p_e(x,z) \log \frac{p_e(x,z)}{p^*(x) p_e(z)}. \qquad (1)$$

- Computing the mutual information can be intractable, lower and upper bounds can be developed:

$$H - D \leq \mathrm{I_e}(X;Z) \leq R \qquad (2)$$

Mutual information based loss

Here,

"Decoder", approximation to *p(x|z)*

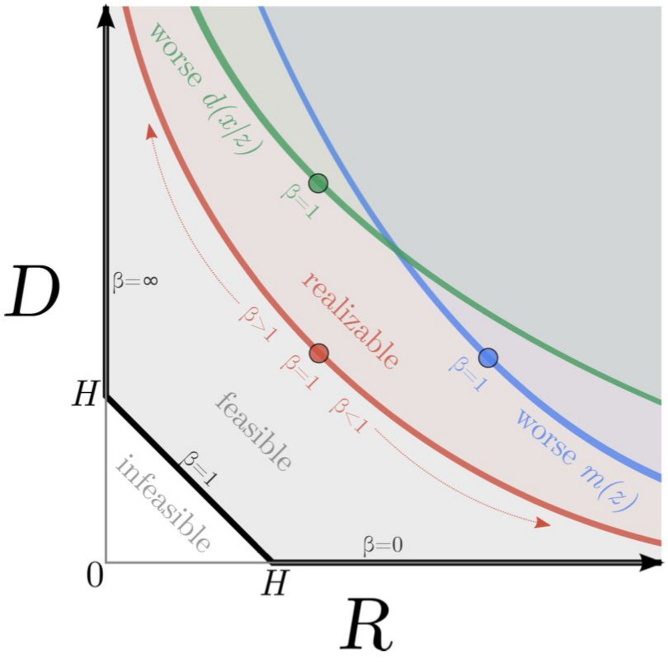
Constant for a given data distribution → $$H \equiv -\int dx\, p^*(x) \log p^*(x) \tag{3}$$

Distortion → $$D \equiv -\int dx\, p^*(x) \int dz\, e(z|x) \log d(x|z) \tag{4}$$

Rate → $$R \equiv \int dx\, p^*(x) \int dz\, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}$$

"Marginal", approximation to *p(z)*

Distortion-Rate phase maps



$$H - D \leq \mathrm{I_e}(X;Z) \leq R$$

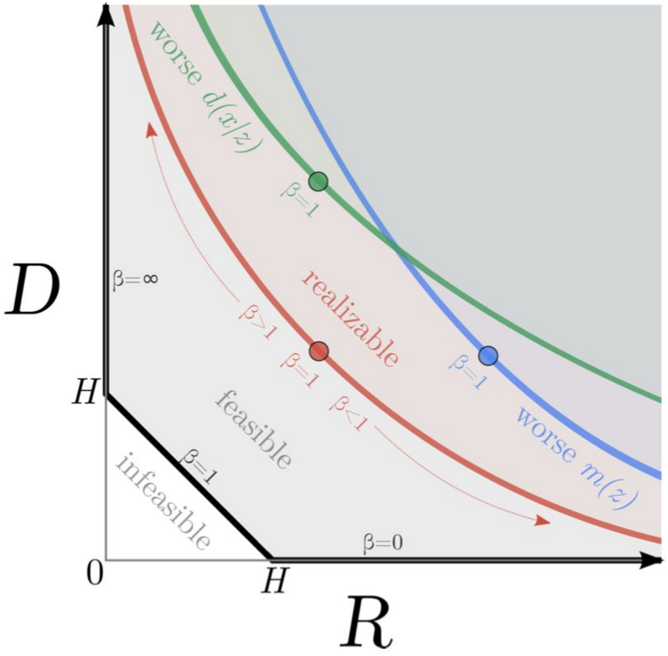
- Assuming that obtaining tighter bounds to MI is good, it is of interest to get closer to the $D + R = H$ line.
- It can be shown that a “perfect” $m(z)$ and a “perfect” $d(x|z)$ will take us to the $D + R = H$ line.

$$H \equiv -\int dx\, p^*(x) \log p^*(x) \tag{3}$$

$$D \equiv -\int dx\, p^*(x) \int dz\, e(z|x) \log d(x|z) \tag{4}$$

$$R \equiv \int dx\, p^*(x) \int dz\, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}$$

Distortion-Rate based loss



$$H - D \leq \mathrm{I_e}(X; Z) \leq R$$

- Low distortion implies better encoding and decoding of data.
- Therefore, a good loss function must minimize *R* while also keeping *D* as low as possible.

$$H \equiv - \int dx\, p^*(x) \log p^*(x) \tag{3}$$

$$D \equiv - \int dx\, p^*(x) \int dz\, e(z|x) \log d(x|z) \tag{4}$$

$$R \equiv \int dx\, p^*(x) \int dz\, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}$$

Distortion-Rate based loss

Consider the following optimization problem and its equivalent form:

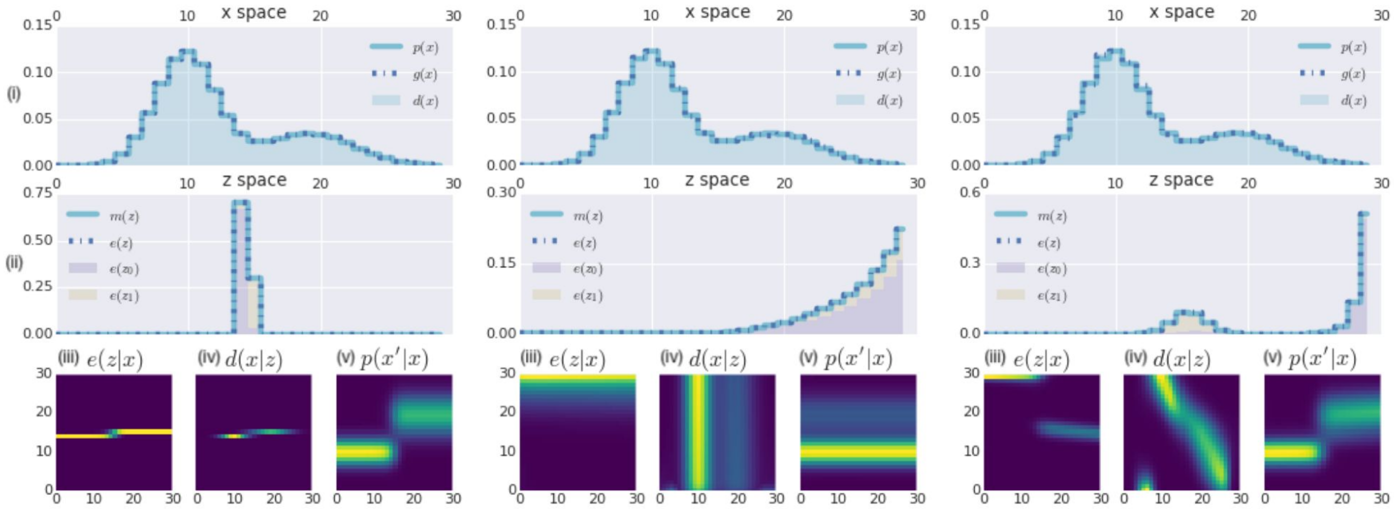
$$\min_{e,d,m} D$$
$$s.t.\, R \leq \epsilon$$

$$\min_{e,d,m} D + \beta R$$

$$\min_{e(z|x),m(z),d(x|z)} \int dx\, p^*(x) \int dz\, e(z|x) \left[-\log d(x|z) + \beta \log \frac{e(z|x)}{m(z)} \right].$$

$\beta = 1$ Corresponds to the standard VAE loss

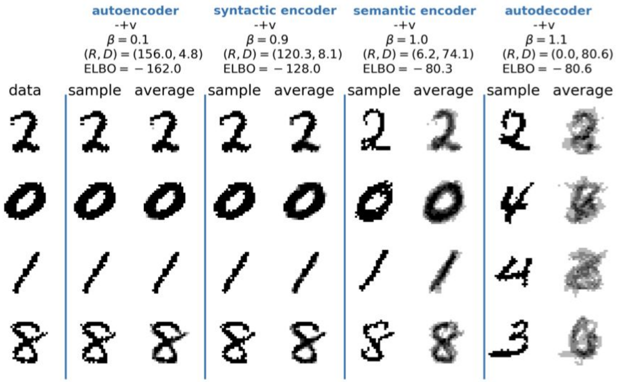
Experiments: Toy problem



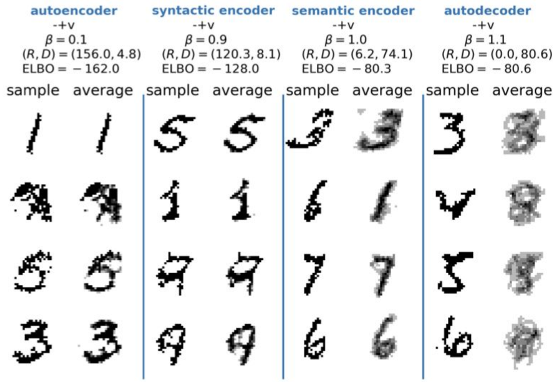
(a) Optimal (hand-constructed) (b) VAE (c) Target Rate

Figure 2. Toy Model illustrating the difference between fitting a model by maximizing ELBO (b) vs minimizing distortion for a fixed rate (c). **Top (i):** Three distributions in data space: the true data distribution, $p^*(x)$, the model's generative distribution, $g(x) = \sum_z m(z)d(x|z)$, and the empirical data reconstruction distribution, $d(x) = \sum_{x'} \sum_z \hat{p}(x')e(z|x')d(x|z)$. **Middle (ii):** Four distributions in latent space: the learned (or computed) marginal $m(z)$, the empirical induced marginal $e(z) = \sum_x \hat{p}(x)e(z|x)$, the empirical distribution over z values for data vectors in the set $\mathcal{X}_0 = \{x_n : z_n = 0\}$, which we denote by $e(z_0)$ in purple, and the empirical distribution over z values for data vectors in the set $\mathcal{X}_1 = \{x_n : z_n = 1\}$, which we denote by $e(z_1)$ in yellow. **Bottom:** Three $K \times K$ distributions: (iii) $e(z|x)$, (iv) $d(x|z)$ and (v) $p(x'|x) = \sum_z e(z|x)d(x'|z)$.

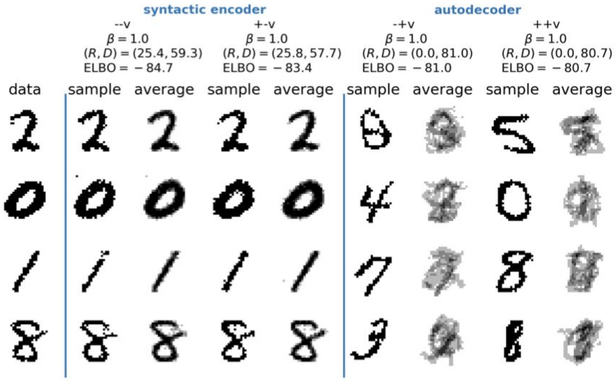
Experiments: Binary MNIST

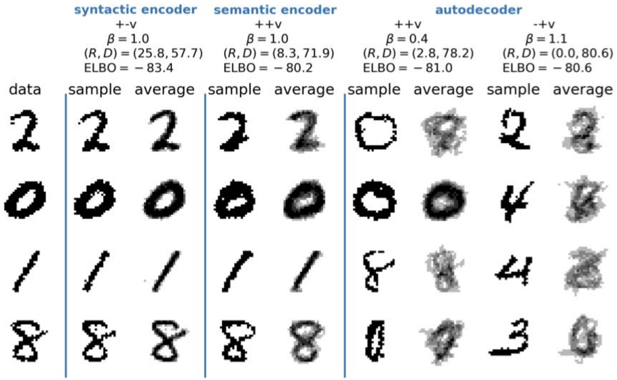


(a) Reconstructions from $-{+}\mathrm{v}$ with $\beta = 0.1 - 1.1$.



(b) Generations from $-{+}\mathrm{v}$ with $\beta = 0.1 - 1.1$





References

- Alemi, Alexander, et al. "Fixing a broken ELBO." *International Conference on Machine Learning*. PMLR, 2018.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).