Description

1 Variational Auto-Encoder

Let us consider a dataset $X = \{x^{(i)}\}_{i=1}^{N}$ consisting of N i.i.d. samples. We assume that the data are generated from parametric family of distributions $p_{\theta^*}(x)$ and we introduce the generative model $p_{\theta^*}(x,z) = p_{\theta^*}(x|z)p_{\theta^*}(z)$ where z is an unobserved random variable. The true parameters θ^* and the values of the latent variables $z^{(i)}$ are unknown to us.

It is worth noting that we are interested in a general algorithm that works efficiently in the case of:

- intractability of the marginal likelihood $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz$ and the true posterior density $p_\theta(z|x) = \dfrac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}$;
- scalability for a large dataset .

Our purpose is to solve the following three problems:

- efficient approximate ML or MAP estimation for the parameters θ;
- efficient approximate posterior inference of the latent variable $p_\theta(z|x)$;
- efficient approximate marginal inference of the variable x.

The algorithm which solves the above problems was proposed by D. Kingma and Prof. Dr. M. Welling in the paper [1]. At first authors introduce a recognition model $q_\varphi(z|x)$: an approximation to the intractable true posterior $p_\theta(z|x)$. After Kingma et al. introduce a method for learning the recognition model parameters φ jointly with the generative model parameters θ.

The key idea is to use the variational lower bound of the marginal likelihood $\ln p_\theta(x)$:

$$\ln p_\theta(x) = D_{KL}(q_\varphi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta,\varphi;x) \quad \Rightarrow$$
$$\Rightarrow \quad \ln p_\theta(x) \geqslant \mathcal{L}(\theta,\varphi;x) = \mathbb{E}_{q_\varphi(z|x)}\left(-\ln q_\varphi(z|x) + \ln p_\theta(x,z)\right) = -D_{KL}(q_\varphi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\varphi(z|x)}(\ln p_\theta(x|z))$$

Our aim is to maximize the lower bound $L(\theta,\varphi;x)$ w.r.t. both the variational parameters φ and the generative parameters θ. However, we have some difficulties with the gradient of the lower bound w.r.t. φ. The usual Monte Carlo gradient estimator is:

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z)}\left[f(z,\varphi)\right] = \mathbb{E}_{q_\varphi(z)}\left[\nabla_\varphi f(z,\varphi)\right] + \mathbb{E}_{q_\varphi(z)}\left[f(z,\varphi)\nabla_\varphi \ln q_\varphi(z)\right] \approx$$
$$\approx \frac{1}{L}\sum_{i=1}^{L}(\nabla_\varphi f(\hat{z}_i,\varphi) + f(\hat{z}_i,\varphi)\nabla_\varphi \ln q_\varphi(\hat{z}_i)), \quad \text{where} \quad \hat{z}_1,\ldots,\hat{z}_L \sim q_\varphi(z)$$

Unfortunately, the term $f(\hat{z}_i,\varphi)\nabla_\varphi \ln q_\varphi(\hat{z}_i)$ in our gradient estimator exhibits very high variance [2] and is impractical for our purposes. Therefore, in this work I consider the variance reduction methods for continuous and discrete variables, compare their performances by training sigmoid belief networks on MNIST.

2 Variance Reduction Techniques

2.1 Reparameterization trick for continuous random variables

In case of the continuous latent variable z with certain mild conditions for a chosen approximate posterior $q_\theta(z|x)$ we can utilize the reparameterization trick whick was proposed in [1]. The idea is simple. If it is possible to express the variable z as a deterministic variable $z = g_\varphi(\varepsilon, x)$ where ε is a random variable with independent marginal $p(\varepsilon)$ and $g_\varphi(.)$ is some vector-valued function parameterized by φ, then the following is true:

$$\int q_\varphi(z|x) f(z,\varphi) dz = \int p(\varepsilon) f(z,\varphi) d\varepsilon = \int p(\varepsilon) f(g_\varphi(\varepsilon, x), \varphi) d\varepsilon$$

Applying this technique we obtain more robust Monte Carlo gradient estimator:

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z|x)}\left[f(z,\varphi)\right] = \nabla_\varphi \mathbb{E}_{p(\varepsilon)}\left[f(g_\varphi(\varepsilon,x),\varphi)\right] = \mathbb{E}_{p(\varepsilon)}\left[\nabla_\varphi f(g_\varphi(\varepsilon,x),\varphi)\right] \approx$$
$$\approx \frac{1}{L}\sum_{i=1}^{L} \nabla_\varphi f(g_\varphi(\hat{\varepsilon}_i, x), \varphi), \quad \text{where} \quad \hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_L \sim p(\varepsilon)$$

3 Examples

3.1 Encoder – gaussian, decoder – gaussian

$$q_\varphi(z|x) = \mathcal{N}(z|\mu_\varphi(x), \sigma^2_\varphi(x) I), \quad p_\theta(z) = \mathcal{N}(z|0, I), \quad p_\theta(x|z) = \mathcal{N}(x|\mu_\theta(z), I),$$
$$z, \mu_\varphi, \sigma^2_\varphi \in \mathbb{R}^d, \quad x, \mu_\theta \in \mathbb{R}^D$$

Let us find the $\mathcal{L}(\theta,\varphi;x) = -D_{KL}(q_\varphi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\varphi(z|x)}(\ln p_\theta(x|z))$. In this case we can analytically calculate the $D_{KL}(q_\varphi(z|x)||p_\theta(z))$:

$$D_{KL}(q_\varphi(z|x)||p_\theta(z)) = \int q_\varphi(z|x) \ln \frac{q_\varphi(z|x)}{p_\theta(z)} dz = \mathbb{E}_{q_\varphi(z|x)} \ln \frac{q_\varphi(z|x)}{p_\theta(z)} =$$
$$= -\mathbb{E}_{q_\varphi(z|x)}\left(\sum_{i=1}^{d} \ln \sigma^i_\varphi(x) + \frac{1}{2}\sum_{i=1}^{d}\left(\frac{1}{(\sigma^i_\varphi)^2}\left(z_i - \mu^i_\theta(x)\right)^2 + z_i^2\right)\right) =$$
$$= -\sum_{i=1}^{d} \ln \sigma^i_\varphi(x) - \frac{d}{2} - \frac{1}{2}\sum_{i=1}^{d}(\sigma^i_\varphi)^2 - \frac{1}{2}\sum_{i=1}^{d}(\mu^i_\varphi)^2 = -\frac{1}{2}\sum_{i=1}^{d}(1 + \ln(\sigma^i_\varphi)^2 + (\sigma^i_\varphi)^2 + (\mu^i_\varphi)^2)$$

Now let us consider the second term $\mathbb{E}_{q_\varphi(z|x)}(\ln p_\theta(x|z))$:

$$\mathbb{E}_{q_\varphi(z|x)}(\ln p_\theta(x|z)) = \mathbb{E}_{q_\varphi(z|x)}\left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(z)||^2_2\right)$$

We can estimate this expectation in several ways. The first usual approach is:

$$\mathbb{E}_{q_\varphi(z|x)}\left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(z)||^2_2\right) \approx \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(\hat{z})||^2_2\right) \ln q_\varphi(\hat{z}|x), \quad \text{where } \hat{z} \sim q_\varphi(z|x)$$

That is,

$$\mathbb{E}_{q_\varphi(z|x)}(\ln p_\theta(x|z)) \approx \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(\hat{z})||^2_2\right)\left(-\frac{d}{2}\ln(2\pi) - \sum_{i=1}^{d}\ln\sigma^i_\varphi(x) - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{1}{(\sigma^i_\varphi)^2}\left(\hat{z}_i - \mu^i_\theta(x)\right)^2\right)\right) =$$
$$= \frac{1}{4}\left(D\ln(2\pi) + ||x - \mu_\theta(\hat{z})||^2_2\right)\sum_{i=1}^{d}\left(\ln(2\pi) + \ln(\sigma^i_\varphi)^2 + \frac{1}{(\sigma^i_\varphi)^2}\left(\hat{z}_i - \mu^i_\theta(x)\right)^2\right), \quad \text{where } \hat{z} \sim q_\varphi(z|x)$$

The second approach utilizes the reparameterization trick:

$$\mathbb{E}_{q_\varphi(z|x)}\left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(z)||^2_2\right) \approx -\frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(\mu_\varphi + \sigma_\varphi \circ \hat{\varepsilon}))||^2_2, \quad \text{where } \hat{\varepsilon} \sim \mathcal{N}(0, I)$$

So, we have obtained the following estimations of $\mathcal{L}(\theta,\varphi;x)$:

- $\mathcal{L}(\theta,\varphi;x) \approx -D_{KL}(q_\varphi(z|x)||p_\theta(z)) + \frac{1}{4}\left(D\ln(2\pi) + ||x - \mu_\theta(\hat{z})||^2_2\right)\sum_{i=1}^{d}\left(\ln(2\pi) + \ln(\sigma^i_\varphi)^2 + \frac{1}{(\sigma^i_\varphi)^2}\left(\hat{z}_i - \mu^i_\theta(x)\right)^2\right)$ where $\hat{z} \sim q_\varphi(z|x)$;
- $\mathcal{L}(\theta,\varphi;x) \approx -D_{KL}(q_\varphi(z|x)||p_\theta(z)) - \frac{D}{2}\ln(2\pi) - \frac{1}{2}||x - \mu_\theta(\mu_\varphi + \sigma_\varphi \circ \hat{\varepsilon}))||^2_2$ where $\hat{\varepsilon} \sim \mathcal{N}(0, I)$;

References

1. *Kingma D. P.*, *Welling M.* Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. — 2013.
2. *Paisley J.*, *Blei D.*, *Jordan M.* Variational Bayesian inference with stochastic search // arXiv preprint arXiv:1206.6430. — 2012.