Due Date : April 14th, 2018

Instructions

- *For all questions, show your work!*
- *This part (theory) is to be done individually.*
- *Use a doc prep system such as LaTeX, or scan a* **very neatly** *hand written version.*
- *Submit your answers electronically via the course studium page.*

1. **(10 points) Reparameterization Trick of Variational Autoencoder**
   Consider a generative model that factorizes as follows $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$, where $p(\boldsymbol{x} \mid \boldsymbol{z})$ is mapped through a neural net, i.e. $p(\boldsymbol{x} \mid \boldsymbol{z}) = p(\boldsymbol{x}; \boldsymbol{h}_\theta(\boldsymbol{z}))$, θ being the set of parameters for the generative network (i.e. decoder), a simple distribution parameterized by $h(\cdot)$ such as Gaussian or Bernoulli (i.e. $p(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_j p(x_j \mid \boldsymbol{z})$). In the case of Gaussian, $\boldsymbol{h}_\theta(z)$ refers to the mean and variance, per dimension as it is fully factorized in the common setting. We have $\boldsymbol{z} \in \mathbb{R}^K$, which implies a continuous latent space model, and $p(\boldsymbol{z}) = \mathcal{N}(0, \boldsymbol{I}_K)$, where $\boldsymbol{I}$ is the identity matrix. The framework of auto-encoding variational Bayes considers maximizing the variational lower bound on the log-likelihood $\mathcal{L}(\theta, \phi) \leq \log p(\boldsymbol{x})$, which is expressed as

$$\mathcal{L}(\theta, \phi) = \mathbf{E}_{q_\phi}[\log p(\boldsymbol{x} \mid \boldsymbol{z})] - \mathbf{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \| p_\theta(\boldsymbol{z})) \tag{1}$$

   where ϕ is the set of parameters used for the inference network (i.e. encoder). The reparameterization trick used in the original work rewrites the random variable in the variational distribution as

$$\boldsymbol{z} = \mu(\boldsymbol{x}) + \sigma(\boldsymbol{x}) \odot \epsilon \tag{2}$$

   where $\epsilon \sim \mathcal{N}(\epsilon; 0, \boldsymbol{I})$, so that gradient can be backpropagated through the stochastic bottleneck.

   (a) Prove that the linearly transformed standard Gaussian noise (2) has the same mean and variance as $\mathcal{N}(\boldsymbol{z}; \mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$. What if we write $\boldsymbol{z} = \mu(\boldsymbol{x}) + S(\boldsymbol{x})\epsilon$, where $S(\boldsymbol{x}) \in \mathbb{R}^{K \times K}$ could be a reshaped K^2 dimensional output of a neural net ? What is the new distribution this reparameterization induces.

   (b) If the traditional mean field variational method is used, i.e. if we factorize the variational distribution as a product of distributions : $q^{mf}(z_i) = \prod_j \mathcal{N}(z_{i,j} \mid m_{i,j}, \sigma^2_{i,j})$ for each data instance x_i, and we maximize the lower bound with respect to the variational parameters and model parameters iteratively, can the inference network used in the variational autoencoder q_ϕ (2) outperform the mean field method ? What is the advantage of using an encoder as in VAE ?

2. **(10 points) Importance Weighted Autoencoder**
   When training a variational autoencoder, the standard training objective is to maximize the evidence lower bound (ELBO). Here we consider another lower bound, called the Importance Weighted Lower Bound (IWLB), a tighter bound than ELBO, defined as

$$\mathcal{L}_k = \mathbf{E}_{z_{1:k} \sim q(\boldsymbol{z} \mid \boldsymbol{x})} \left[\log \frac{1}{k} \sum_{j=1}^{k} \frac{p(\boldsymbol{x}, z_j)}{q(z_j \mid \boldsymbol{x})} \right] \tag{3}$$

   for an observed variable $\boldsymbol{x}$ and a latent variable $\boldsymbol{z}$, k being the number of importance samples. The model we are considering has joint that factorizes as $p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$, $\boldsymbol{x}$ and $\boldsymbol{z}$

being the observed and latent variables, respectively. In the following questions, one needs to make use of the Jensen's inequality :

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)] \tag{4}$$

for a convex function f.

(a) Show that IWLB is a lower bound on the log likelihood $\log p(\boldsymbol{x})$.

(b) Given a special case where $k = 2$, prove that $\mathcal{L}_2$ is a tighter bound than the ELBO (with $k = 1$).

3. **(10 points) Maximum Likelihood for Generative Adversarial Networks**
The original GAN objective is the following

$$\max_D \mathbf{E}_{P_{\text{data}}(x)}\big[\log D(x)\big] + \mathbf{E}_{P(z)}\big[\log(1 - D(G(z)))\big]; \quad \max_G \mathbf{E}_{P(z)}\big[\log D(G(z))\big]$$

This generator objective can be generalized by replacing the log with a general function f :

$$\max_G \mathbf{E}_{P(z)}\big[f(D(G(z)))\big]$$

Find a function f such that the objective corresponds to maximum likelihood, assuming the discriminator is optimal.
Hint : Use the optimal discriminator : $D^* = \frac{P_{data}}{(P_{data}+P_{gen})}$.