Instructor: Vishnu Boddeti
CSE 891-001: Deep Learning
Homework 4
December, 08

Instructions: This homework is based on the background material necessary for the course, in terms of linear algebra and basic machine learning concepts. Please refer to the background slides for a helpful material.

- **Filename:** Submit solutions in PDF format titled `written-assignment-4-msunetid.pdf`. Submissions in other formats or with other filenames will not be graded. You can produce your file however you like, LaTeX, Word or scan. Handwritten scans that are not legible will not be graded.
- **Submission:** Only homeworks uploaded to Google Classroom will be graded. Make sure to show all the steps of your derivations in order to receive full credit.
- **Integrity and Collaboration:** You are expected to work on the homeworks by yourself. You are not permitted to discuss them with anyone except the instructor. The homework that you hand in should be entirely your own work. You may be asked to demonstrate how you got any results that you report.
- **Clarifications:** If you have any question, please look at Google Classroom first. Other students may have encountered the same problem, and is solved already. If not, post your question there. We will respond as soon as possible.

1 Reversible Architectures [3pts]: In this section, we will investigate a variant for implementing reversible block with affine coupling layers. Consider the following reversible affine coupling block:

$$\begin{aligned} \boldsymbol{y}_1 &= \exp(\mathcal{G}(\boldsymbol{x}_2)) \circ \boldsymbol{x}_1 + \mathcal{F}(\boldsymbol{x}_2) \\ \boldsymbol{y}_2 &= \exp(\boldsymbol{s}) \circ \boldsymbol{x}_2 \end{aligned} \tag{1}$$

where $\circ$ denotes element-wise multiplication. The each inputs $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{\frac{d}{2}}$. The functions $\mathcal{F}$ and $\mathcal{G}$ maps from $\mathbb{R}^{\frac{d}{2}} \rightarrow \mathbb{R}^{\frac{d}{2}}$. This modified block is identical to the ordinary reversible block, except that the inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are multiplied element-wise by vectors $\exp(\mathcal{F}(\boldsymbol{x}_2))$ and $\exp(\boldsymbol{s})$.

1. (**1pt**) Give the equations for inverting this block, i.e. computing $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from $\boldsymbol{y}_2$ and $\boldsymbol{y}_2$. You may use / to denote element-wise division.
2. (**1pt**) Give a formula for the Jacobian $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$, where $\boldsymbol{y}$ denotes the concatenation of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. You may denote the solution as a block matrix, as long as you clearly define what the matrix for each block corresponds to.
3. (**1pt**) Give a formula for the determinant of the Jacobian from previous part, i.e. compute $det\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)$. Is this a volume preserving transformation? Justify your answer.

2 Variational Free Energy [6pts]: In this question you will derive some expressions related to variational free energy which is maximized to train a VAE. Recall that the VFE is defined as:

$$\mathcal{F}(q) = \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}))$$

where KL divergence is defined as

$$D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z})) = \mathbb{E}_q[\log q(\boldsymbol{z}) - \log p(\boldsymbol{z})]$$

We will assume that the prior $\boldsymbol{z}$ is a standard Gaussian:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}) = \prod_{i=1}^{D} p_i(z_i) = \prod_{i=1}^{D} \mathcal{N}(z_i; 0, 1)$$

Similarly we will assume that the variational approximation $q(\boldsymbol{z})$ is a fully factorized (i.e., diagonal) Gaussian:

$$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{D} q_i(z_i) = \prod_{i=1}^{D} \mathcal{N}(z_i; \mu_i, \sigma_i)$$

1. (**1pt**) Show that:
$$\mathcal{F}(q) = \log p(\boldsymbol{x}) - D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

2. (**1pt**) Show that the KL term decomposes as a sum of KL terms for individual dimensions. In particular,
$$D_{KL}(q(\boldsymbol{z})||p(\boldsymbol{z})) = \sum_i D_{KL}(q_i(z_i)||p_i(z_i))$$

3. (**2pts**) Give an explicit formula for the KL divergence $D_{KL}(q_i(z_i)||p_i(z_i))$. This should be a mathematical expression involving μ_i and σ_i.

4. (**2pts**) One way to do gradient descent on the KL term is to apply the formula from above. Another approach is to compute stochastic gradients using the reparameterization trick:
$$\nabla_{\boldsymbol{\theta}} D_{KL}(q_i(z_i)||p_i(z_i)) = \mathbb{E}_\epsilon[\nabla_{\boldsymbol{\theta}} t_i]$$
, where
$$\boldsymbol{\theta} = \begin{bmatrix} \mu_i \\ \sigma_i \end{bmatrix}$$
and
$$\begin{aligned} z_i &= \mu_i + \sigma_i \epsilon_i \\ r_i &= \log q_i(z_i) \\ s_i &= \log p_i(z_i) \\ t_i &= r_i - s_i \end{aligned} \tag{2}$$
Show how to compute a stochastic estimate of $\nabla_{\boldsymbol{\theta}} D_{KL}(q_i(z_i)||p_i(z_i))$ by doing backpropagation on the above equations. You may find it helpful to draw the computation graph.

3 Feedback (1pt):

1. What aspects of the written and programming homeworks did you enjoy for this course?
2. What aspects of the written and programming homeworks did you hate for this course?
3. Suggestions for what you would like to modify in the homeworks.
4. Suggestions for course content/lecture slides and topics.