Homework 5

Deadline: Thursday, April 4, at 11:59pm.

Submission: You must submit your solutions as a PDF through MarkUs. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner) as long as it is readable.

Late Submission: MarkUs will remain open until 3 days after the deadline, after which no late submissions will be accepted. The late penalty is 10% per day, rounded up.

Weekly homeworks are individual work. See the Course Information handout[1] for detailed policies.

Due to the shortened time period, this assignment has only one question, worth 6 points. You get the remaining 4 points for free.

1. **Variational Free Energy [6pts]** Here, your job is to derive some of the formulas relating to the variational free energy (VFE) which we maximize when we train a VAE. Recall that the VFE is defined as:

$$\mathcal{F}(q) = \mathbb{E}_q[\log p(\mathbf{x} \,|\, \mathbf{z})] - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) \,\|\, p(\mathbf{z})),$$

and KL divergence is defined as

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) \,\|\, p(\mathbf{z})) = \mathbb{E}_q[\log q(\mathbf{z}) - \log p(\mathbf{z})].$$

We assume the prior $p(\mathbf{z})$ is a standard Gaussian:

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

And the variational approximation $q(\mathbf{z})$ is a fully factorized (i.e. diagonal) Gaussian:

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

For reference, here are the formulas for the univariate and multivariate Gaussian distributions:

$$\mathcal{N}(z; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$
$$\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right)$$

(a) **[1pt]** Show that

$$\mathcal{F}(q) = \log p(\mathbf{x}) - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) \,\|\, p(\mathbf{z} \,|\, \mathbf{x})).$$

(Hint: expand out definitions and apply Bayes' Rule.)

(b) **[1pt]** Show that the KL term decomposes as a sum of KL terms for individual dimensions. In particular,

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}) \,\|\, p(\mathbf{z})) = \sum_i \mathrm{D}_{\mathrm{KL}}(q_i(z_i) \,\|\, p_i(z_i)).$$

[1] http://www.cs.toronto.edu/~rgrosse/courses/csc421_2019/syllabus.pdf

(c) **[2pts]** Give an explicit formula for the KL divergence $\mathrm{D}_{\mathrm{KL}}(q_i(z_i) \,\|\, p_i(z_i))$. This should be a mathematical expression involving μ_i and σ_i. If you like, you may suppress the i subscripts in your solution.

(d) **[2pts]** One way to do gradient descent on the KL term is to apply the formula from part (c). Another approach is to compute stochastic gradients using the reparameterization trick:

$$\nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q_i(z_i) \,\|\, p_i(z_i)) = \mathbb{E}_{\epsilon}[\nabla_{\boldsymbol{\theta}} t_i],$$

where

$$\boldsymbol{\theta} = \begin{pmatrix} \mu_i \\ \sigma_i \end{pmatrix}$$

and

$$\begin{aligned} z_i &= \mu_i + \sigma_i \epsilon_i \\ r_i &= \log q_i(z_i) \\ s_i &= \log p_i(z_i) \\ t_i &= r_i - s_i \end{aligned}$$

Show how to compute a stochastic estimate of $\nabla_{\boldsymbol{\theta}} \mathrm{D}_{\mathrm{KL}}(q_i(z_i) \,\|\, p_i(z_i))$ by doing backprop on the above equations. You may find it helpful to draw the computation graph. If you like, you may suppress the i subscripts in your solution.