Homework 4

Deadline: Friday, November 20 at 11:59pm.

Submission: You need to submit through Markus.
Late Submission: 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

Collaboration: Homeworks are individual work. See the course website[1] for detailed policies.

1. **[4pts] Multilayer Perceptron.** Give the weights and biases of a multilayer perceptron which takes as input two scalar values (x_1, x_2) and outputs the values in sorted order, i.e. (y_1, y_2) with $y_1 = \min(x_1, x_2)$ and $y_2 = \max(x_1, x_2)$. The hidden units should all use the ReLU activation function, and the output units should be linear. You should explain why your solution works, but you don't need to provide a formal proof.

2. **[6pts] Backprop.** The deep residual network, or ResNet, is the state-of-the-art architecture for image classification. It's based on a kind of layer called a *residual block*; in this question, you'll figure out how to backprop through a residual block. While the actual ResNet is a convolutional architecture, we'll consider a toy version that's fully connected.

   Consider the following architecture, which takes as input a vector $\mathbf{x}$ and outputs a vector $\mathbf{y}$ of the same size. Its hidden representation $\mathbf{h}$ also has the same size (i.e. number of units). The computations are as follows:

$$\begin{aligned}\mathbf{h} &= \phi(\mathbf{W}\mathbf{x} + \mathbf{b}) + \mathbf{x} \\ \mathbf{y} &= \phi(\mathbf{V}\mathbf{h} + \mathbf{c}) + \mathbf{h}\end{aligned}$$

   The parameters are the weight matrices $\mathbf{W}$ and $\mathbf{V}$ and the bias vectors $\mathbf{b}$ and $\mathbf{c}$. Here, ϕ is the activation function, and you can write its elementwise derivatives as $\phi'(\cdots)$.

   To help with the backprop derivations, it's useful to decompose out these computations in a way that introduces variables to hold some intermediate results:

$$\begin{aligned}\mathbf{z} &= \mathbf{W}\mathbf{x} + \mathbf{b} \\ \mathbf{h} &= \phi(\mathbf{z}) + \mathbf{x} \\ \mathbf{r} &= \mathbf{V}\mathbf{h} + \mathbf{c} \\ \mathbf{y} &= \phi(\mathbf{r}) + \mathbf{h}\end{aligned}$$

   (a) **[2pt]** Draw the computation graph for all the variables ($\mathbf{x}$, $\mathbf{z}$, $\mathbf{h}$, $\mathbf{r}$, $\mathbf{y}$, $\mathbf{W}$, $\mathbf{b}$, $\mathbf{V}$, and $\mathbf{c}$).

   (b) **[4pts]** Determine the backprop rules (in vector form) for computing the gradients with respect to all the parameters ($\mathbf{W}$, $\mathbf{b}$, $\mathbf{V}$, and $\mathbf{c}$).

3. **[10 points] EM for Probabilistic PCA.** In lecture, we covered the EM algorithm applied to mixture of Gaussians models. In this question, we'll look at another interesting example of EM but where the latent variables are continuous: probabilistic PCA. This is a model very similar in spirit to PCA: we have data in a high-dimensional space, and we'd like to

[1] `https://www.cs.toronto.edu/~huang/courses/csc2515_2020f/index.html`

summarize it with a lower-dimensional representation. Unlike ordinary PCA, we formulate the problem in terms of a probabilistic model. We assume the latent code vector $\mathbf{z}$ is drawn from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and that the observations are drawn from a spherical Gaussian whose mean is a linear function of $\mathbf{z}$. We'll consider the slightly simplified case of scalar-valued z (i.e. only one principal component). The probabilistic model is given by:

$$z \sim \mathcal{N}(0, 1) \tag{1}$$

$$\mathbf{x} \mid z \sim \mathcal{N}(z\mathbf{u}, \sigma^2) \tag{2}$$

where σ^2 is the noise variance (which we assume to be fixed) and $\mathbf{u}$ is a parameter vector (which, intuitively, should correspond to the top principal component). Note that the observation model can be written in terms of coordinates:

$$x_j \mid z \sim \mathcal{N}(zu_j, \sigma^2).$$

We have a set of observations $\{\mathbf{x}^{(i)}\}_{i=1}^N$, and z is a latent variable, analogous to the mixture component in a mixture-of-Gaussians model.

In this question, you'll derive both the E-step and the M-step for the EM algorithm.

(a) **E-step (4 points).** In this step, your job is to calculate the statistics of the posterior distribution $q(z) = p(z \mid \mathbf{x})$ which you'll need for the M-step. In particular, your job is to find formulas for the (univariate) statistics:

$$m = \mathrm{E}[z \mid \mathbf{x}] =$$
$$s = \mathrm{E}[z^2 \mid \mathbf{x}] =$$

Tips:

- First determine the conditional distribution $p(z \mid \mathbf{x})$ using the Gaussian conditioning formulas from the Appendix. To help you check your work: $p(z \mid \mathbf{x})$ is a univariate Gaussian distribution whose mean is a linear function of $\mathbf{x}$, and whose variance does not depend on $\mathbf{x}$.
- Once you've determined the conditional distribution (and hence the posterior mean and variance), use the fact that $\mathrm{Var}(Y) = \mathrm{E}[Y^2] - \mathrm{E}[Y]^2$ for any random variable Y.

(b) **M-step (6 points).** In this step, we need to re-estimate the parameters, which consist of the vector $\mathbf{u}$. (Recall that we're treating σ as fixed.) Your job is to derive a formula for $\mathbf{u}_{\text{new}}$ that maximizes the expected log-likelihood, i.e.,

$$\mathbf{u}_{\text{new}} \leftarrow \arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}_{q(z^{(i)})}[\log p(z^{(i)}, \mathbf{x}^{(i)})].$$

(Recall that $q(z)$ is the distribution computed in part (a).) This is the new estimate obtained by the EM procedure, and will be used again in the next iteration of the E-step. Your answer should be given in terms of the $m^{(i)}$ and $s^{(i)}$ from the previous part. (I.e., you don't need to expand out the formulas for $m^{(i)}$ and $s^{(i)}$ in this step, because if you were implementing this algorithm, you'd use the values $m^{(i)}$ and $s^{(i)}$ that you previously computed.)

Tips:

- First expand out $\log p(z^{(i)}, \mathbf{x}^{(i)})$. You'll find that a lot of the terms don't depend on $\mathbf{u}$ and can therefore be dropped.
- Apply linearity of expectation. You should wind up with terms proportional to $\mathrm{E}_{q(z^{(i)})}[z^{(i)}]$ and $\mathrm{E}_{q(z^{(i)}}[[z^{(i)}]^2]$. Replace these expectations with $m^{(i)}$ and $s^{(i)}$. You should get an equation that does not mention $z^{(i)}$. (If you don't wind up with terms of this form, then see if there's some way you can simplify $\log p(z^{(i)}, \mathbf{x}^{(i)})$.
- In order to find the maximum likelihood parameter $\mathbf{u}_{\text{new}}$, you need to determine the gradient with respect to $\mathbf{u}$, set it to zero, and solve for $\mathbf{u}_{\text{new}}$.

Appendix: Some Properties of Gaussians

Consider a multivariate Gaussian random variable $\mathbf{z}$ with the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. I.e.,

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Now consider another Gaussian random variable $\mathbf{x}$, whose mean is an affine function of $\mathbf{z}$ (in the form to be clear soon), and its covariance $\mathbf{S}$ is independent of $\mathbf{z}$. The conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ is

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mathbf{A}\mathbf{z} + \mathbf{b}, \mathbf{S}).$$

Here the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are of appropriate dimensions.

In some problems, we are interested in knowing the distribution of $\mathbf{z}$ given $\mathbf{x}$, or the marginal distribution of $\mathbf{x}$. One can apply Bayes' rule to find the conditional distribution $p(\mathbf{z} \mid \mathbf{x})$. After some calculations, we can obtain the following useful formulae:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \mathbf{S}\right)$$
$$p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid \mathbf{C}(\mathbf{A}^\top\mathbf{S}^{-1}(\mathbf{x} - \mathbf{b}) + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), \mathbf{C}\right)$$

with

$$\mathbf{C} = (\boldsymbol{\Sigma}^{-1} + \mathbf{A}^\top\mathbf{S}^{-1}\mathbf{A})^{-1}.$$

You may also find it helpful to read Section 2.3 of Bishop.