Assignment #3

Due: 1 April, 1 pm

In this assignment, we'll look at various approaches to dealing with having small amounts of data. You can use automatic differentiation in your code, but must still answer the gradient questions.

Data preparation Binarize the MNIST dataset. In this assignment, we'll use only **30 examples** in our training set. We'll keep the test set the same size, at 10000 examples.

Question 1 (L2-Regularized Logistic Regression, 10 points)

In this question, we'll attempt to regularize logistic regression to deal with having such a small dataset. Recall that the likelihood given by this model is:

$$p(c|\mathbf{x}, \mathbf{w}) = \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\sum_{c'=0}^{9} \exp(\mathbf{w}_{c'}^T \mathbf{x})} \tag{1}$$

(a) Using your code from assignment 2's question 3(d) and (e), fit a maximum likelihood estimate of logistic regression to the 30 training points, and report the training and test-set error. Also plot the learned parameters as a set of 10 images.

(b) Next, let's define a prior distribution on parameters, so that we can fit a *maximum a posteriori* (MAP) estimate. Let's consider a spherical Gaussian prior on the parameters:

$$p(\mathbf{w}|\sigma^2) = \prod_{c=0}^{9} \prod_{d=1}^{784} \mathcal{N}(w_{cd}|0, \sigma^2) \tag{2}$$

For observed target classes $\mathbf{t}$, write down $\log \left[p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\sigma^2) \right]$, the log of the likelihood of the entire dataset $(\mathbf{X}, \mathbf{t})$ multiplied by the prior on parameters. Also write down its gradient w.r.t. $\mathbf{w}$. You do not need to show the derivation. Hint: The gradient should resemble that of assignment 2's question 3(d) but with a term added that mainly depends on $\mathbf{w}$.

(c) Fit a MAP estimate of the parameters $\mathbf{w}$ on the training set using gradient ascent. Try different values of hyperparameter σ^2 across several orders of magnitude. For the value of σ^2 with the highest test-set log-likelihood, plot the optimized $\mathbf{w}_{MAP}$ as 10 images. Also print the training and test accuracy, and average predictive log-likelihood:

$$\frac{1}{N} \sum_{i=1}^{N} \log p(t_i|\mathbf{x}_i, \mathbf{w}) \tag{3}$$

Question 2 (Markov-chain Monte Carlo, 10 points)

Let p($\mathbf{w}$) correspond to a Gaussian mixture model with $\boldsymbol{\pi} = \left\{\frac{3}{4}, \frac{1}{4}\right\}$, $\boldsymbol{\mu} = \{0, 8\}$, and $\boldsymbol{\sigma} = \{1, 1\}$. In this question, you will estimate E($\mathbf{w}$) using 3000 Metropolis-Hastings iterations, initialized at $x = 0$.

(a) Estimate the expectation E($\mathbf{w}$) using a proposal distribution $Q(x'|x) \sim \mathcal{N}(x, 1)$. Plot your samples on the same graph as the true distribution p($\mathbf{w}$), as a histogram. How many of the 3000 iterations resulted in successful samples? Hint: You may wish to run your code a few times but you need only report on one run.

(b) Repeat using a *mixture proposal*: for each iteration, a proposal distribution $Q(x'|x) \sim \mathcal{N}(x, 10^2)$ is used with 50% probability, and with 50% probability the proposal in (a) is used instead.

(c) Compare your two estimates for E($\mathbf{w}$) to one another and to the true answer. State which of (a) or (b) is best for the current task, and justify your answer.

Wilfred Hastings was a U of T student and prof (until 1971), who passed away last May. He is pictured below on the left next to Nicholas Metropolis.