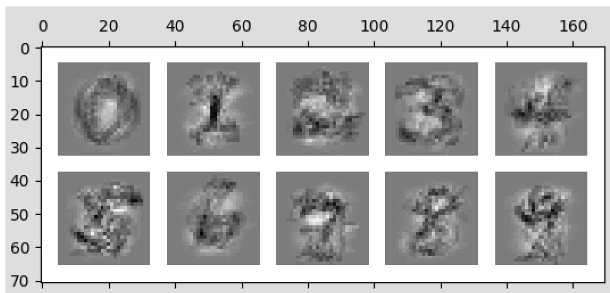
Phileas Hocquard
phileas@cs.toronto.edu

Probabilistic Learning and Reasoning, #3

Problem 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) **Fit a maximum likelihood estimate of logistic regression to the 300 training points, 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:**

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

Write down $\log\left(p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)\right)$, the log-likelihood of the entire training set (X, t) of 300 examples, multiplied by the prior on parameters.

$$\log(p(w|\sigma^2)\,p(t|X, w)) = \log(p(w|\sigma^2)) + \log(p(t|X, w))$$

$$\log(p(w|\sigma^2)\,p(t|X, w)) = \sum_{k,d=1}^{K,D} \frac{-(w_{cd})^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) + \sum_{n=1}^{N} \sum_{k,d=1}^{K,D} P(c|x_d^n, w_{kd})$$

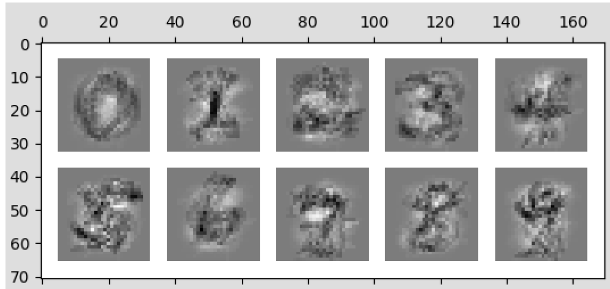
$$\log(p(w|\sigma^2)\,p(t|X, w)) = \sum_{k,d=1}^{K,D} \frac{-(w_{cd})^2}{2\sigma_d^2} + \frac{1}{2}\log(2\pi\sigma_d^2) + \sum_{n=1}^{N} \sum_{k,d=1}^{K,D} \log\frac{\exp(w_{kd}^T x_d)}{\sum_{c'=0}^{9} \exp(w_{cd}^T x_d)}$$

Gradient:

$$\nabla_w \log(p(t|X, w)\,p(w|\sigma^2)) = \nabla_w \log(p(t|X, w)) + \nabla_w \log(p(w|\sigma^2))$$

$$\nabla_w \log(p(t|X, w)\,p(w|\sigma^2)) = \sum_{n=1}^{N} \left(\sum_{k,d=1}^{K,D} x_d^n * (1(t_k^n = 1) - \log\frac{\exp(w_{kd}^T x_d^n)}{\sum_{c'=0}^{9} \exp(w_{cd'}^T x_d^n)}) \right) + \sum_{k,d=1}^{K,D} \frac{-(w_{kd})}{\sigma_d^2}$$

(c) **Fit a MAP estimate of the parameters w on the training set using gradient ascent.**
The accuracy is (only) 0.03% higher with a prior.



Problem 2 (Bayesian Logistic Regression using Stochastic Variational Inference, 20 points)
In this question, we'll avoid choosing a single set of parameters $\hat{\mathbf{w}}$. Instead, we'll approximately *integrate over all possible* $\mathbf{w}$. This will avoid over-fitting by making approximately Bayes-optimal predictions, given the assumptions of our model.

(a) **Code up SVI for this model. That is, use stochastic gradient ascent to find locally optimal variational parameters maximizing the evidence lower bound:**

$$\boldsymbol{\phi}^* = \operatorname{argmax}_{\boldsymbol{\phi}} \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\phi})} \left[\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\sigma^2) - \log q(\mathbf{w}|\boldsymbol{\phi})\right] \tag{3}$$

(b) **Use your code to find $\boldsymbol{\phi}^*$. Compute the average predictive accuracy on the test set using simple Monte Carlo using your approximate posterior and 100 samples (S=100):**

$$p(t_i|\mathbf{x}_i) = \int p(t_i|\mathbf{x}_i, \mathbf{w}) p(\mathbf{w}|\mathbf{t}, \mathbf{X}) d\mathbf{w} \approxeq \frac{1}{S} \sum_{j=1}^{S} p(t_i|\mathbf{x}_i, \mathbf{w}^{(j)}), \qquad \text{each } \mathbf{w}^{(j)} \sim q(\mathbf{w}|\boldsymbol{\phi}^*) \tag{4}$$

Play with the prior variance σ^2 to see if you can get a higher test-set accuracy than MAP inference.

With σ^2 = 1.0, we obtain a 77.22% accuracy over 100 iterations of training for 300 training examples.

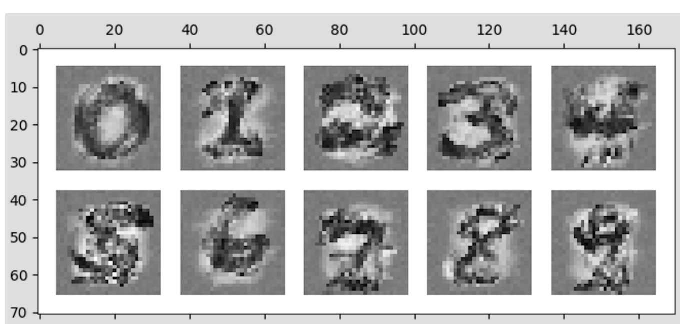


Figure 1: variational posterior means

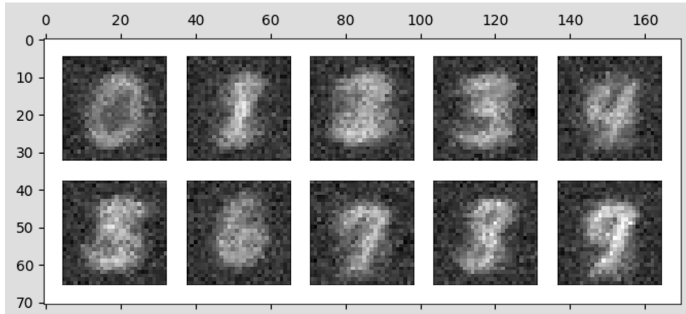


Figure 2: The variational posterior standard deviations

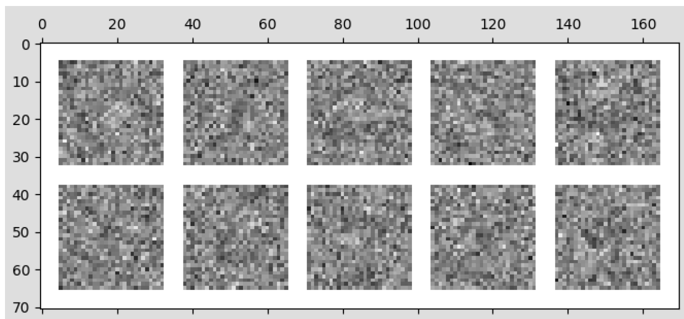


Figure 3: A single sample from the variational posterior

(c) **The above plot for a single sample from $q(\mathbf{w}|\boldsymbol{\phi}^*)$ will be extremely noisy. Consider how our model treats pixels which it never sees 'on' across all training examples. In particular, starting from $\log p(t|w,x)$ show that if $x_d \in B$, the set of pixels which are always off, then the training labels do not effect the optimal variational parameters for those pixels.**

$x_d \epsilon B$, d' signifies that the pixel is not active throughout all of X.

$$\int q(w_{cd'})(\int \prod_{cd,d'\neq d} q(w_{cd}|\phi_{cd})\, p(t|w)\, dw_{cd,d\neq d'})\, dw_{cd'}$$

$f(\phi_{cd,d\neq d'}) = \prod_{cd,d'\neq d} q(w_{cd}|\phi_{cd})\, p(t|w)\, dw_{cd,d\neq d'}$ contains no $w_{cd'}$ at all instances.

. Using fubini theorem: $(\int q(w_{cd'})\, d_{wd'})\, f(\phi_{cd,d\neq d'})$

$Obj = f(\phi_{cd,d\neq d'}) - KL(q(w|\phi)\|p(w))$, where $KL(..) = \int q(w|\mu_{cd'},\sigma_{cd'}) \dfrac{q(w|\mu_{cd'},\sigma_{cd'})}{p(w|0,\epsilon)}$

. The close form of KL is$[\log \dfrac{\sigma_{cd'}}{\sigma} + \dfrac{\sigma_{cd'} + (\mu_{cd'} - 0)^2}{2\sigma^2} - \dfrac{1}{2}]$

$argmax_{\mu_{cd'},\sigma_{cd'}} Obj = 1 - 1 + (g_1(\sigma_{cd'},\mu_{cd'}) + g_2(\sigma_{cd},\mu_{cd})) - 0,$

Where $\sigma_{cd'}, \mu_{cd'}$ will equal to 0 and ϵ respectively. Therefore as demonstrated, the training labels will not have an effect on the optimal variational parameters for the pixels which are off throughout all D.