TTIC 31230 Fundamentals of Deep Learning
Winter 2018
Midterm Exam

Problem 1. Consider convolving an $N \times N$ filter over a $D \times D$ input image x (with $\lfloor N/2 \rfloor$ padding) stride 1 to produce a $D \times D$ output image y. Assume the input image has C_x channels and the output image image has C_y channels and we have batch size B. How many floating point multiplies are required in computing the convolution on the batch (without any activation function)?

$$BD^2N^2C_xC_y$$

Problem 2. The equations defining a UGRNN are given below.

$$\begin{aligned} h^{t+1} &= f^t \odot h^t + (1 - f^t) \odot d^t \\ f^t &= \sigma(W^f[x^t, h^t] + b^b) \\ d^t &= \tanh(W^d[x^t, h^t] + b^d) \end{aligned}$$

a) Rewrite these as an equivalent set of equations with the vector concatenations replaced with a pair of matrix multiplications where W^f is replaced by two matrices $W^{f,x}$ and $W^{f,h}$ and similarly for W^d.

$$h^{t+1} = f^t \odot h^t + (1 - f^t) \odot d^t$$

$$f^t = \sigma(W^{f,x}x^t + W^{f,h}h^t + b^f)$$

$$d^t = \tanh(W^{d,x}x^t + W^{d,h}h^t + b^d)$$

b) Translate the equations from part (a) into explicit index notation with explicit summations including the batch index.

$$\begin{aligned} h^{t+1}[b,c] &= f^t[b,c]h^t[b,c] + (1 - f^t[b,c])d^t[b,c] \\ f^t[b,c] &= \sigma\left(\left(\sum_{c'} W^{f,x}[c,c']x[b,c']\right) + \left(\sum_{c'} W^{f,h}[c,c']h^t[b,c']\right) + b^h[c]\right) \\ d^t[b,c] &= \sigma\left(\left(\sum_{c'} W^{d,x}[c,c']x[b,c']\right) + \left(\sum_{c'} W^{d,h}[c,c']h^t[b,c']\right) + b^d[c]\right) \end{aligned}$$

c) Give explicit index expression for the backward method for h^{t+1}. Your equations should compute additions to f^t.grad, h^t.grad and d^t.grad.

$$f^t.\text{grad}[b,c] \texttt{ += } f^{t+1}.\text{grad}[b,c](h^t[b,c] - d^t[b,c])$$

$$h^t.\text{grad}[b,c] \texttt{ += } f^{t+1}.\text{grad}[b,c]f^t[b,c]$$

$$d^t.\text{grad}[b,c] \texttt{ += } f^{t+1}.\text{grad}[b,c](1 - f^t[b,c])$$

Problem 3.

a) Calculate the differential entropy of a Gaussian distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}.$$

Use the natural logarithm in your definition of entropy.

$$\begin{aligned} H(p) &= E_{x\sim P} - \ln p(x) \\ &= E_{x\sim P}\, \frac{x^2}{2\sigma^2} + \ln\sigma + \ln\sqrt{2\pi} \\ &= \frac{\sigma^2}{2\sigma^2} + \ln\sigma + \frac{1}{2}\ln 2\pi \\ &= \ln\sigma + \frac{1}{2}(1 + \ln 2\pi) \end{aligned}$$

b) Let the "signal" x be a Gaussian random variable with variance σ_x and let the "noise" ϵ be an independent Gaussian random variable with variance σ_ϵ. Let $z = x + \epsilon$. Use the fact that a sum of independent Gaussians is Gaussian with $\sigma_z^2 = \sigma_x^2 + \sigma_\epsilon^2$ to compute the differential mutual information $I(x, z)$. Express your answer in terms of the signal to noise ratio $\sigma_x^2/\sigma_\epsilon^2$. Hint: select a convenient expression for mutual information and use part (a).

$$
\begin{aligned}
I(z,x) &= H(z) - H(z|x) \\
&= \ln(\sigma_x^2 + \sigma_\epsilon^2) - \ln \sigma_\epsilon^2 \\
&= \ln\left(\frac{\sigma_x^2 + \sigma_\epsilon^2}{\sigma_\epsilon^2}\right) \\
&= \ln\left(1 + \frac{\sigma_x^2}{\sigma_\epsilon^2}\right)
\end{aligned}
$$

c) For both the differential entropy in (a) and the mutual information in (b) say whether the numerical value depends on the choice of units.

The numerical value of differential entropy is sensitive to the units we choose for σ. As long as σ_x and σ_ϵ are measured in the same units, the numerical value of the mutual information is units-independent. This is consistent with the fact that differential entropy is not directly meaningful while mutual information can be written as a KL-divergence and differential KL-divergence is meaningful.

Problem 4. Consider a graphical model with N nodes numbered 1 through N and where each node can take on one of the values 0 or 1. We let $\hat{x}$ be an assignment of a value to every node. We define the score of $\hat{x}$ by

$$f(\hat{x}) = \sum_{i=1}^{N-1} \mathbb{1}[\hat{x}[i] = \hat{x}[i+1]]$$

The probability distribution over assignments is defined by a softmax.

$$Q_f(\hat{x}) = \operatorname*{softmax}_{\hat{x}} f(\hat{x})$$

What is the **Pseudoliklihood** of the all ones assignment?

$$\tilde{P}_f(\hat{x}) = \Pi_i P_f(\hat{x}[i] \mid \hat{x}/i)$$

where $\hat{x}/i$ consists of all components of $\hat{x}$ other than i. In a graphical model $P_f(\hat{x}[i] \mid \hat{x}/i)$ is determined by the neighbors of i and we can consider only how a value is scored against it neighbors. For $\hat{x}$ equal to all ones we have

$$
\begin{aligned}
f(\hat{x}) &= N - 1 \\
f(\hat{x}[i=0]) &= \begin{cases} N-3 & \text{for } 1 < i < N \\ N-2 & \text{for } i = 1 \text{ or } i = N \end{cases}
\end{aligned}
$$

For $1 < i < N$ we get

$$\begin{aligned} Q_f(\hat{x}[i=1] \mid \hat{x}/i) &= \frac{e^{N-1}}{e^{N-1}+e^{N-3}} \\ &= \frac{1}{1+e^{-2}} \end{aligned}$$

and for $i = 1$ or $i = N$ we get

$$Q_f(\hat{x}[i=1] \mid \hat{x}/i) = \frac{1}{1+e^{-1}}$$

This gives

$$\tilde{Q}(\hat{x}) = (1+e^{-1})^{-2}(1+e^{-2})^{-(N-2)}$$

Problem 5. This problem is on the initialization of Resnet filters. Consider the following residual skip connection where the diversion is a convolution with an $N \times N$ filter.

$$x^{\ell+1} = x^{\ell} + \text{Conv}(W^{\ell}, x^{\ell})$$

Here we have omitted an activation function that would be present in practice. This omission allows an analysis that seems to provide insight into the more complex case of adding a relu activation around the convolution.

Consider a network of L layers of this equation, all of which are done stride 1 so that the image dimensions are preserved and with a different weight matrix W^{ℓ} at each layer. Assume that each image x^{ℓ} has C channels (ignore the fact that input images have only three channels). Assume that each channel of each pixel of each image is independent of the other channels and assume that each channel value of each pixel of the input image is drawn independently with zero mean and unit variance. Also suppose that the weights of the matrices W^{ℓ} are each drawn at random from a Gaussian distribution with zero mean and variance σ_W. Also assume that the the two terms in the sum of the residual skip equation above are independent (although they aren't). Just assume everything is independent and recall that the variance σ^2 of a sum of independent random is the sum of the variances and the variance of a product of independent random variables is the product of the variances.

a) Give an expression for the variance $\sigma^2_{\ell+1}$ of each channel value at layer $\ell+1$ as a function of the variance σ^2_{ℓ} at layer ℓ and the weight variance σ_W.

Assuming everything is independent we have

$$\begin{aligned}\sigma_{\ell+1}^2 &= \sigma_\ell^2 + N^2 C \sigma_w^2 \sigma_\ell^2 \\ &= \sigma_\ell^2 (1 + N^2 C \sigma_w^2)\end{aligned}$$

b) Assume that the input layer x^0 has independent channel values each with variance 1. Give an expresison for the variance σ_ℓ^2 directly as a function of σ_w and ℓ.

$$\sigma_\ell^2 = (1 + N^2 C \sigma_w^2)^\ell$$

c) Using $(1+\epsilon)^N \approx e^{\epsilon N}$ solve for the value of σ_W such that $\sigma_L^2 = 2$.

$$\begin{aligned}\sigma_L^2 &= (1 + N^2 C \sigma_w^2)^L \\ &\approx e^{L N^2 C \sigma_W^2}\end{aligned}$$

setting

$$e^{L N^2 C \sigma_W^2} = 2$$

gives

$$\sigma_w \approx \sqrt{\frac{\ln 2}{L N^2 C}}$$

d) Assuming top level gradient x^L.grad has zero mean and unit variance, and that all components of $x^{\ell+1}$.grad are independent, give an expression for the variance $\sigma_{\ell,\text{grad}}^2$ of the components of x^ℓ.grad as a function of ℓ and σ_W.

$$\sigma_{\ell,\text{grad}}^2 = (1 + N^2 C \sigma_W^2)^{L-\ell}$$

e) Consider the limit of $\sigma_W \to 0$. Give an explicit index expression for the limit as $\sigma_W \to 0$ of W^ℓ.grad. Your expression should be in terms of x^L.grad. If we add an activation function, does W^ℓ.grad have a well defined limit as $\sigma_W \to 0$?

In the limit of $\sigma_W^2 \to 0$ we have $x^\ell = x^0$ for all ℓ. By the swap rule for W^ℓ.grad we have

$$W^\ell.\text{grad}[\Delta i, \Delta j, c', c] \texttt{ += } x^L.\text{grad}[i, j, c] x^\ell[i + \Delta i, j + \Delta j, c']$$

Hence in the limit as $\sigma_W^2 \to 0$ we have that $W^\ell.\text{grad} = W^0.\text{grad}$ for all ℓ. This last equations holds for any activation function and the limits still exist when we add an activation function.

Problem 6. We consider SGD on an arbitrary loss function.

$$\Phi^* = \operatorname*{argmin}_\Phi E_{(x,y)\sim \text{Pop}}\ \text{loss}(\Phi, x, y)$$

SGD defines a random walk in parameter space — a Markov process in parameter space. Let P_t be a probability distribution over parameters for time t in this Markov process. We are interested in understanding the stationary distribution of this process as a function of the learning rate η. The update equation is

$$\begin{aligned} \Phi_{t+1} &= \Phi_t - \eta \hat{g}_t \\ \hat{g}_t &= \nabla_\Phi \text{loss}(\Phi_t, x_t, y_t) \end{aligned}$$

We let P_η be the stationary distribution of this process at learning rate η. Intuitively, smaller values of η should lead to sharper distributions more focused on Φ^*. To simplify the notation we assume WLOG that $\Phi^* = 0$. For a distribution P consider the following definition of the spread of the distribution around the optimum $\Phi^* = 0$.

$$\text{Spread}(P) \doteq E_{\Phi\sim P}||\Phi||^2$$

a) Let P_t be the probability distribution over parameter vectors for time step t of the Markov process defined by SGD. Give an expression for the expected change in spread $\text{Spread}(\Phi_{t+1}) - \text{Spread}(\Phi_t)$. Your expression should be given as an expectation over Φ drawn from P and over a draw of $\hat{g}$ conditioned on Φ. Leave your expression in terms of a random draw over $\hat{g}$ and do not expand this into a draw over (x, y).

$$\begin{aligned} & E_{\Phi\sim P_t,\hat{g}}\ ||\Phi - \eta\hat{g}||^2 - E_{\Phi\sim P_t}\ ||\Phi||^2 \\ = \ & E_{\Phi\sim P_t,\hat{g}}\ ||\Phi - \eta\hat{g}||^2 - ||\Phi||^2 \\ = \ & E_{\Phi\sim P_t,\hat{g}}\ -2\eta\Phi^\top\hat{g} + \eta^2||\hat{g}||^2 \end{aligned}$$

b) Use your answer to (a) to give a condition on the stationary distribution by setting the change in spread to be zero. (A simple rewrite of (a)).

$$E_{\Phi\sim P_t,\hat{g}}\ \Phi^\top\hat{g} \quad = \quad \frac{\eta}{2} E_{\Phi\sim P_t,\hat{g}}\ ||\hat{g}||^2$$

c) For η small, and for a positive definite total Hessian, the distribution will be tightly focused on the optimum parameter value. In this case we can use the following second order Taylor expansion of the loss function around $\Phi^* = 0$.

$$\begin{aligned} \text{loss}(\Phi, x, y) &\approx L(x,y) + g(x,y)\Phi + \frac{1}{2}\Phi^\top H(x,y)\Phi \\ \\ L(x,y) &= \text{loss}(0,x,y) \\ g(x,y) &= \nabla_\Phi \text{loss}(0,x,y) \\ H(x,y) &= \nabla_\Phi \nabla_\Phi \text{loss}(0,x,y) \end{aligned}$$

Your condition for (b) should express a balance between two expressions. Insert the above second order approximation for $\hat{g}$ into each of these expressions to get a balance condition between the two leading terms of the two expressions.

$$\begin{aligned} E_{\Phi\sim P_t,(x,y)\sim\text{Pop}}\ \Phi^\top(g(x,y) + H(x,y)\Phi) &= \frac{\eta}{2} E_{(x,y)\sim\text{Pop}}\ ||g(x,y)||^2 \\ E_{\Phi\sim P_t}\ \Phi^\top H\Phi &= \frac{\eta}{2} E_{(x,y)\sim\text{Pop}}\ ||g(x,y)||^2 \\ H &\doteq E_{(x,y)\sim\text{Pop}}\ H(x,y) \end{aligned}$$

d) Recall that P_η is the stationary distribution of the Markov process at learning rate η. Show that under the second order approximations, your solution to (c) implies that the loss gap

$$E_{\Phi\sim P_\eta,\ (x,y)\sim\text{Pop}}\ \text{loss}(\Phi,x,y) - E_{(x,y)\sim\text{Pop}}\ \text{loss}(0,x,y)$$

is proportional to η.

Inserting the second order approximation into the loss gap, and observing that $E_{(x,y)\sim\text{Pop}}\ g(x,y) = 0$ we have that the loss gap can be written as

$$\begin{aligned} \text{gap} &= E_{\Phi\sim P_\eta,\ (x,y)\sim\text{Pop}}\ \frac{1}{2}\Phi^\top H(x,y)\Phi \\ &= E_{\Phi\sim P_\eta}\ \frac{1}{2}\Phi^\top H\Phi \\ &= \frac{\eta}{4} E_{(x,y)\sim\text{Pop}}\ ||g(x,y)||^2 \end{aligned}$$