TTIC 31230 Fundamentals of Deep Learning
Midterm Exam Practice Questions

Problem 1. We consider a "Karpathy" normalization Layer applied to a layer in a convolutional nerual network.

$$L' = \mathtt{Karpathy(L, A, B)}$$

```
class Karpathy:

    def __init__(self,L,A,B)
        ...

    def forward(self)
        self.value = self.A.value*(self.L.value + self.B.value)

    def backward(self)
        ...
```

Assume that A and B have shape (C) where C is the number of channels. Assume that L has shape (B, H, W, C) where B is the minibatch size, H is the height dimension, W is the width dimention, C is the number of channels.

a. write the forward method in `+=` notation.

$$L'[\ldots] \mathrel{\texttt{+=}} A[\ldots](L[\ldots] + B[\ldots])$$

(fill in the indices in the above equation).

b. Write the `+=` notation for the backward method (to L, A and B).

c. Write the Python/Numpy code for the backward method using appropriate Numpy vector operations.

Problem 2. This problem is on momentum. Momentum is traditionally defined in terms of "velocity"

$$\hat{v}^{t+1} = \mu v^t + \eta \nabla_\Theta \ell^t(\Theta)$$

$$\Theta \mathrel{\texttt{-=}} v^{t+1}$$

In this class we have defined momentum by a smoothing operation on the gradient vector.

$$\begin{aligned}
\hat{g}^{t+1} &= \mu \hat{g}^{t} + (1-\mu)\nabla_{\Psi}\ \ell^t(\Psi) \\
\Psi^t &\mathrel{-}= \eta' \hat{g}^{t+1}
\end{aligned}$$

Assume that $\Psi^0 = \Theta^0$ and that $v^0 = \hat{g}^0 = 0$ and define

$$\eta' = \frac{\eta}{1-\mu}$$

Show by induction on t that we have

$$\begin{aligned}
v^t &= \eta'\ \hat{g}^t \\
\Psi^t &= \Theta^t
\end{aligned}$$

The semantics of μ and η' seem clearer in the gradient smoothing formulation. In particular, we expect the optimum value of η' to be near the optimal value of η when momentum is not used. We also expect that μ and η' are more nearly conjugate in meta-parameter search (we can optimize μ and η' independently).

Problem 3. Adam uses the equations.

$$\begin{aligned}
\hat{g}_i^{t+1} &= \beta_1 \hat{g}_i^t + (1-\beta_1)\left(\nabla_\Theta\ \ell^t(\Theta)\right)_i \\
s_i^{t+1} &= \beta_2 s_i^t + (1-\beta_2)\left(\nabla_\Theta\ \ell^t(\Theta)\right)_i^2 \\
\Theta_i^{t+1} &= \Theta_i^t - \frac{\eta}{\sqrt{s_i^{t+1}} + \epsilon}\ \hat{g}_i^{t+1}
\end{aligned}$$

Suppose that for each scalar parameter Θ_i we have that $\left(\nabla_\Theta\ \ell^t(\Theta)\right)_i$ is either 1 or 0 and the fraction of time that it is 1 is ϵ. If we want s_i^t to be a reasonable estimate of $\mathrm{E}\left[\left(\nabla_\Theta\ \ell^t(\Theta)\right)_i^2\right]$ what is a reasonable value of β_2 as a function of ϵ? Explain your answer.

Problem 4. An intuitive prior on numbers is the log-uniform prior where $\log_2 |\Theta_i|$ is taken to be uniformly distributed between, say, -10 and 10. A log-uniform prior corresponds to a density

$$p(\Theta_i) \propto 1/|\Theta_i|$$

This correspinds to the following regularizer.

$$\Theta^* = \operatorname*{argmin}_{\Theta} \ell(\Theta) + \lambda \sum_i \ln |\Theta_i|$$

Assuming that we only consider positive value of Θ_i (to make things simpler) give the conditions on $\partial\ell/\partial\Theta_i$ for a local optimum (or rather a stationary point — a point where the derivative of the objective function is zero).

Problem 5. This problem is on complex-step differentiation. Suppose that Θ is a parameter tensor and that we have computed Θ.grad using complex arithmetic at the parameter setting $\Theta + i\epsilon\Delta\Theta$ where $\epsilon = 2^{-50}$. Write the expression for $H\Delta\Theta$ where H is the Hessian in terms of the complex value for Θ.grad.

Problem 6. This problem concerns initialization. Consider a unit defined by a simple inner product of a weight vector and previous units followed by a Karpathy normalization and then a nonlinearity.

$$y = \sigma(W \cdot x)$$

or in component notation

$$\begin{aligned} y &= \sum_i w_i x_i \\ z &= \sigma(a(y+b)) \end{aligned}$$

a. Assume that the x_i are independent and have mean μ_x and variance σ_x^2. Assume that the the w_i are initialized independently to have mean μ_w and variance σ_w^2. If we want $a(y+b)$ to have zero mean and unit variance, how should we initialize a and b?

b. If a and b are initialized so that $a(y+b)$ has zero mean and unit variance, is there any advantage to using Xavier initialization on w? Explain your answer.

Problem 7. Consider a highway path update of the form

$$L_{i+1} = F_i * L_i + (1 - F_i) * D_i(L_i)$$

where F_i is a parameter (is independent of the problem instance) rather than being computed from L_i. In a Resnet-like CNN the diversion $D_i(F_i)$ uses different parameters for each i. Assume that F_i can also be a different parameter for each i. Do you think this is a reasonable CNN architecture? Explain your answer (your explanation is more important than your answer).

Problem 8. Suppose that at training time we construct a mask μ on the parameters so that

$$\begin{cases} \mu_i = \frac{1}{2} & \text{with probability } \alpha \\ \mu_i = 1 & \text{with probability } 1 - \alpha \end{cases}$$

Then at train time we do

$$\begin{aligned} y_i &= \mathrm{Relu}\left(\sum_j W_{i,j}\ \mu_j x_j\right) \\ \Theta &\mathrel{-}= \nabla_\Theta\ \ell(\Theta, \mu) \end{aligned}$$

Give a corresponding weight scaling rule for computing y_i at test time for this "half drop out" training algorithm.