TTIC 31230 Fundamentals of Deep Learning
Midterm Exam
Thursday February 9

Problem 1. (15 points) 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 we want

$$\hat{g}^t \approx \frac{1}{k}\sum_{i=1}^{k} \nabla_\Theta\, \ell^{t-i}(\Theta)$$

Give a value of β_1 which corresponds to this backward-window averaging.

$$\beta_1 = 1 - \frac{1}{k}$$

Note that the backward looking window and the Adam smoothing rule then both assign weight $\frac{1}{k}$ to each new data point.

Problem 2. (15 points) Suppose that at training time we construct a random matrix ϵ withere $\epsilon_{i,j}$ is drawn form a zero-mean unit-variance Gaussian. Then at train time we do

$$\begin{aligned}
y_i &= \text{Relu}\left(\sum_j (W_{i,j} + \epsilon_{i,j}) x_j\right) \\
\Theta &\texttt{ -= } \nabla_\Theta\, \ell(\Theta, \epsilon)
\end{aligned}$$

Give a corresponding weight scaling rule for computing y_i at test time where the input to the Relu activation has the same expectation as the input to the Relu at train time.

No weight scaling is required in this case. The expection of $w_i x_i$ is the same at test time as at training time.

Problem 3. (30 points) We consider a three dimensional convolutional neural network applied to video data where we have

$$L2[b, x, y, t, c] = \sum_{u,v,z,c'} F[u, v, c'] * L1[b, x + u, y + v, t + z, c', c]$$

Here we have that $L1$ has shape (B, H, W, T, C'), $L2$ has shape (B, H, W, T, C) and F has shape (U, V, Z, C', C).

a. Convert this to a `+=` expression of the form

$$\forall b, x, y, t, c, u, v, z, c' \quad L2.\text{value}[\cdots] \texttt{ += } F.\text{value}[\cdots] L1.\text{value}[\cdots]$$

$\forall b, x, y, t, c, u, v, z, c' \quad L2.\text{value}[b, x, y, t, c] \texttt{ += } F.\text{value}[u, v, z, c', c] L1.\text{value}[x+u, y+v, z+t, c']$

b. Write a Numpy vector operation implementation of this `+=` expression for the forward method using a Python loop over x, y and t and a call to `matmul` on appropriate and reshapings.

```
for x in range(H)
    for y in range(W)
        for t in Range(T)

            L1Slice = Li.value[:,x:x+U,y:y+V,t:t+Z,:]
            L1Reshaped = LiSlice.reshape((B,-1))
            FReshaped = F.value.reshape((-1,C))
            L2[:,x,y,t,:] += np.matmul(L1Reshaped,FReshaped)
```

b. Write the corresponding `+=` expressions for the backpropagation to F and $L1$.

$\forall b, x, y, t, c, u, v, z, c'$

$$\begin{aligned} F.\text{grad}[u, v, w, c', c] &\texttt{ += } L2.\text{grad}[b, x, y, t, c]\ L1.\text{value}[x + u, y + v, z + t, c'] \\ L1.\text{grad}[b, x, y, t, c'] &\texttt{ += } L2.\text{grad}[b, x, y, t, c]\ F.\text{value}[u, v, z, c', c] \end{aligned}$$

c. Write a Numpy vector operation implementation your backpropagation `+=` expressions. Again, you can use a Python loop over x and y but use vector operations in the body of the loop.

```
for x in range(H)
    for y in range(W)
        for t in Range(T)

            L2gradSlice = L2.grad[:,x,y,t,:].reshape(B,1,1,1,1,C)
            L1slice = L1.value[:,x:x_U,y:y+V,t:t+Z,:]
            L1gradslice = L1.grad[:,x:x_U,y:y+V,t:t+Z,:]
            F.grad += (L2gradSlice * L1Slice.reshape((B,U,V,Z,C',1))).sum(axis=0)
            L1gradslice += (L2.gradSlice*F.value).sum(axis=5)
```

Problem 3. (20 points) Suppose we use the following regularizer

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}}\, \ell(\Theta) + \lambda||\Theta||$$

Here $||\Theta|| = \sqrt{\sum_i \Theta_i^2}$ is the standard geometric length of the vector Θ. This regularization is different from standard L_2 regularization in that we are using $||\Theta||$ rather than $||\Theta||^2$.

a. Give a condition on $\nabla_\Theta\, \ell(\Theta)$ for a local optimum (or rather a stationary point — a point where the derivative of the objective function is zero) with $\Theta \neq 0$. Hint: write $||\Theta||$ as $\sqrt{||\Theta||^2}$ and use the fact that $\nabla_\Theta\, ||\Theta||^2 = 2\Theta$.

$$\begin{aligned}
\nabla_\Theta ||\Theta|| &= \nabla_\Theta \sqrt{||\Theta||^2} \\
&= \frac{1}{2}\frac{1}{\sqrt{||\Theta||^2}} 2\Theta \\
&= \frac{\Theta}{||\Theta||}
\end{aligned}$$

So we get

$$\nabla_\Theta \ell(\Theta) = -\lambda \frac{\Theta}{||\Theta||}$$

b. Show that any stationary point of this objective is also a stationary point of standard L_2 regularization for some other regularization constant λ'.

The corresponding stationarity condition for L_2 regulariazation i

$$\nabla_\Theta \, \ell(\Theta) = -\lambda' \Theta$$

Taking $\lambda' = \lambda / ||\Theta||$ we get that Θ is a stationary point of both regularizations.

c. Consider a value of Θ small enough that the following first order expansion of $\ell(\Theta)$ around $\Theta = 0$ is accurate

$$\ell(\Theta) \approx \ell(0) + (\nabla_\Theta \, \ell(\Theta) \; @ \; \Theta = 0) \cdot \Theta$$

If we have

$$(\nabla_\Theta \, \ell(\Theta) \; @ \; \Theta = 0) \cdot \Theta + \lambda ||\Theta|| > 0 \tag{1}$$

for all Θ then any sufficiently small deviation from $\Theta = 0$ increases the regularized objective function and $\Theta = 0$ is a local minuma of the regularized objective.

Give the condition on $\nabla_\Theta \, \ell(\Theta)$ at $\Theta = 0$ such that (1) holds for all directions Θ.

If we hold $||\Theta||$ constant and minimize the first inner product we see that Θ should be in the opposite direction of $\nabla_\Theta \, \ell(\Theta)$ and the expression becomes

$$\lambda ||\Theta|| - ||\nabla_\Theta \, \ell(\Theta)|| \, ||\Theta||$$

This expression is strictly positive if and only if

$$||\nabla_\Theta \, \ell(\Theta)|| < \lambda$$

Problem 5. (20 points) This problem is on Hessian-vector products. Assume that the Hessian of the function $\ell(\Theta)$ is constant. In other words $\ell(\Theta)$ is a quadratic function of the vector Θ. Optimizing a quadratic objective in very high dimension is still nontrivial. Suppose we are at parameter vector Θ and have measured the gradient vector g (for the total objective, not just one sample problem) and we use complex-step differentiation to measure the Hessian-vector product $\dot{g} = Hg$. Give the value of λ at which $\ell(\Theta - \lambda g)$ is minimized (holding Θ and g fixed while varying λ). Hint: write the second order Taylor expansion

of $\ell(\Theta)$ around Θ. For a quadratic function the second order Taylor expansion is exact.

$$\ell(\Theta + \Delta\Theta) = \ell(\Theta) + g \cdot \Delta\Theta + \Delta\Theta^\top H \Delta\Theta$$

$$\ell(\Theta - \lambda g) = \ell(\Theta) - \lambda ||g||^2 + \frac{1}{2}\lambda^2 g \cdot \dot{g}$$

Setting the derivative with respec to λ to zero gives

$$\lambda(g \cdot \dot{g}) = ||g||^2$$

or

$$\lambda = \frac{||g||^2}{g \cdot \dot{g}}$$