**TTIC 31230 Fundamentals of Deep Learning, Winter 2019**

**Trainability Problems**

**Problem 1.** This problem is on initialization. Consider a single unit defined by

$$u = f\left(\left(\sum_{i=1}^{N} W[i]x[i]\right) - B\right).$$

where $B$ is initialized to zero and $f$ is an activation function such as a sigmoid or ReLU. The vector $x$ is a random variable determined by a random draw of a training example. Assume that the components of $x$ are independent and that each component has zero mean and unit variance. Suppose that we initialize each weight in $W$ from a distribution with zero mean and variance $\sigma$ and that is symmetric about zero — (the probability that $w[i] = z$ equals the probability that $w[i] = -z$). Consider $y = \sum_i W[i]x[i]$ as a random variable defined by the distribution on $x$ and the independent random distribution on $W$. Recall that the variance $\sigma^2$ of a sum of independent random variables is the sum of the variances and the variance of a product of zero mean independent random variables is the product of the variances.

(a) What value of the initialization variance $\sigma$ for $W[i]$ gives zero mean and unit variance for $y$? Show your derivation.

(b) For a sigmoid activation function what is the mean of $u$.

(c) For a sigmoid activation function is the variance of $u$ larger than, equal to, or smaller than the variance of $y$?

(d) What is the largest possible variance of the output of a sigmoid?

(e) If we assume recursively that each component of $x$ has a positive nonzero mean, should we change the initialization of the threshold $B$?

**Problem 2.** This problem is on the initialization of ResNet filters. Consider the following residual skip connection where $R_{\ell+1}$ is computed with an $N \times N$ filter.

$$\begin{array}{l}
\text{for } b, x, y, j\,\Delta x, \Delta y, j' \\
\qquad R_{\ell+1}[b,x,y,j] \quad \texttt{+=} \quad W_{\ell+1}[\Delta x, \Delta y, j', j]\; L_\ell[b, x+\Delta x, y+\Delta y, j'] \\
\\
\text{for } b, x, y, j \\
\qquad R_{\ell+1}[b,x,y,j] \quad \texttt{-=} \quad B_{\ell+1}[j] \\
\\
\text{for } b, x, y, j \\
\qquad L_{\ell+1}[b,x,y,j] \quad = \quad L_\ell[b,x,y,j] + R_{\ell+1}[b,x,y,j]
\end{array}$$

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 with activations.
Assume that $L_0[b, x, y, j]$ is computed from the input in some unspecified way such that $L_0[b, x, y, j]$ has unit variance. Assume that the values $L_\ell[b, x, y, j]$ and $R_{\ell+1}[b, x, y, j]$ are all independent. Suppose that each weight $W_\ell[\Delta x, \Delta y, j, j']$ is drawn independently at random from a distribution with zero mean and variance $\sigma_W$. Recall that the variance $\sigma^2$ 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 $\sigma_\ell^2$ of $L_{\ell+1}[b, x, y, j]$ as a function of $\ell$, the filter dimension $D = \Delta X = \Delta Y$, shthe feature dimension $J$, and the weight variance $\sigma_W^2$.

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

(c) Assuming $L_L.\text{grad}[b, x, y, j]$ has unit variance, and that all components of $L_\ell.\text{grad}[b, x, y, j]$ and $R_\ell.\text{grad}[b, x, y, j]$ are independent, give an expression for the variance $\sigma^2_{\ell,\text{grad}}$ of the components of $L_\ell.\text{grad}[b, x, y, j]$ as a function of $\ell$, $D$, $J$ and $\sigma_W$.