

TTIC 31230 Fundamentals of Deep Learning, Winter 2019

Trainability Problems

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

$$u = f \left(\left(\sum_i W[i]x[i] \right) - B \right).$$

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 zero mean and unit variance. Suppose that we initialize each weight in W from a distribution with zero mean and variance σ . 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 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.

- What value of the initialization variance σ gives zero mean and unit variance for y ? Show your derivation.
- For a sigmoid activation function, and for the components of x and W drawn from zero mean Gaussians, what is the mean of u .
- For a sigmoid activation function, is the variance of u larger than, equal to, or smaller than the variance of y ?
- A ReLU activation yields a non-zero mean for u than a sigmoid activation. Is the variance of the ReLU version smaller or larger than the variance of the σ version?
- 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 the diversion is a convolution with an $N \times N$ filter.

$$\begin{aligned} &\text{for } b, x, y, j \Delta x, \Delta y, j' \\ &R_{\ell+1}[b, x, y, j] \quad += \quad W_{\ell+1}[\Delta x, \Delta y, j', j] \quad L_{\ell}[b, x + \Delta x, y + \Delta y, j'] \end{aligned}$$

$$\begin{aligned} &\text{for } b, x, y, j \\ &R_{\ell+1}[b, x, y, j] \quad -= \quad B_{\ell+1}[j] \end{aligned}$$

$$\begin{aligned} &\text{for } b, x, y, j \\ &L_{\ell+1}[b, x, y, j] \quad = \quad L_{\ell}[b, x, y, j] + R_{\ell+1}[b, x, y, j] \end{aligned}$$

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 $R_\ell[b, x, y, j]$ 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 σ_W . 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 σ_ℓ^2 of $L_{\ell+1}[b, x, y, j]$ as a function of ℓ , the filter dimension $D = \Delta X = \Delta Y$, the feature dimension J , and the weight variance σ_W^2 .

(b) Using $(1 + \epsilon)^N \approx e^{\epsilon N}$ solve for the value of σ_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 $\sigma_{\ell, \text{grad}}^2$ of the components of $L_\ell.\text{grad}[b, x, y, j]$ as a function of ℓ , D , J and σ_W .