

1. Dropout on Linear Regression Recall that linear regression optimizes:

$$\mathcal{L}(\mathbf{w}) = \|\mathbf{y} - X\mathbf{w}\|_2^2 \quad (1)$$

One way of using *dropout* on the d -dimensional input features \mathbf{x}_i involves keeping each feature at random with probability p (and zeroing it out if not kept). This makes our learning objective effectively become

$$\mathcal{L}(\tilde{\mathbf{w}}) = \mathbb{E}_{R \sim \text{Bernoulli}(p)} \left[\|\mathbf{y} - (R \odot X)\tilde{\mathbf{w}}\|_2^2 \right] \quad (2)$$

where \odot is the element-wise product, and the random binary matrix $R \in \{0, 1\}^{n \times d}$ is such that $R_{i,j} \sim_{i.i.d} \text{Bernoulli}(p)$. We use $\tilde{\mathbf{w}}$ to remind you that this is learned by dropout.

Show that we can manipulate (2) to eliminate the expectations and get:

$$\mathcal{L}(\tilde{\mathbf{w}}) = \|\mathbf{y} - pX\tilde{\mathbf{w}}\|_2^2 + p(1-p)\|\tilde{\Gamma}\tilde{\mathbf{w}}\|_2^2 \quad (3)$$

with $\tilde{\Gamma}$ being a diagonal matrix whose j -th diagonal entry is the norm of the j -th column of the training matrix X .

2. Feature Dimensions in CNN

We are going to describe a convolutional neural net using the following pieces:

- CONV3-F denotes a convolutional layer with F different filters, each of size $3 \times 3 \times C$, where C is the depth (i.e. number of channels) of the activations from the previous layer. Padding is 1, and stride is 1.
- POOL2 denotes a 2×2 max-pooling layer with stride 2 (pad 0)
- FLATTEN just turns whatever shape input tensor into a one-dimensional array with the same values in it.
- FC-K denotes a fully-connected layer with K output neurons.

Note: All CONV3-F and FC-K layers have biases as well as weights. **Do not forget the biases when counting parameters.**

We are going to use this network to do inference on a single input. Fill in the missing entries in this table of the size of the activations at each layer, and the number of parameters at each layer. You can/should write your answer as a computation (e.g. $128 \times 128 \times 3$) in the style of the already filled-in entries of the table.

Layer	Number of Parameters	Dimension of Activations
Input	0	$28 \times 28 \times 1$
CONV3-10		$28 \times 28 \times 10$
POOL2	0	$14 \times 14 \times 10$
CONV3-10	$3 \times 3 \times 10 \times 10 + 10$	
POOL2		
FLATTEN	0	490
FC-3		3