PID:_____

True/False

(10 pts: +1 for correct, −0.5 for incorrect, 0 for no answer.) If you would like to justify an answer, feel free. Please CLEARLY CIRCLE ONLY T or F.

1. T F Maximum Likelihood Estimation (MLE) can be applied directly to any kind of data without any assumptions about the underlying distribution of the data.

2. T F Ideally a softmax activation in the output layer of a multi-class problem should only be used if the classes are mutually exclusive.

3. T F In K-fold cross validation, the model with the lowest validation loss is guaranteed to perform the best on the test set, given an optimal set of hyperparameters.

4. T F Stride in a CNN determines the size of the jump from one receptive field to another.

5. T F In CNNs, the pooling layer is used to reduce the spatial dimensions (width and height) of the input volume for the next convolutional layer, without affecting the depth of the volume.

6. T F When using stochastic gradient descent, the order in which we give the training data to the network does not matter. More precisely, the weights learned by the network are the same irrespective of the order of the training data sequence.

7. T F A higher value of the momentum parameter signifies more weight given to past gradients.

8. T F It is very difficult for a neural network to perform well if we initialize all its weights and biases to 0 because the features learned (if any) will all be the same at each layer.

9. T F Pixels in one region of an image don't typically inform what pixels in another region might be, and this property is represented in convnets by small receptive fields.

Multiple Choice

1. (2 pts) Suppose you are training a multilayer perceptron model, and during training, your train loss converges at a higher value than expected. Which tweak is *least* likely to lower the training loss?

(a) Increase the number of neurons in each hidden layer of the model.

(b) Increase the number of layers in the model.

(c) Add L2 regularization to the weights update rule.

(d) Normalize your data using some method like z-scoring or min-max normalization.

2. (2 pts) Which of the following statements about normalization is true?

(a) Training data is normalized first before being split into train and validation sets.

(b) Once you normalize data, you cannot unnormalize the data unless you have the original normalizing parameters.

(c) Normalization means transforming the data into a range strictly between -1 and 1.

(d) It is not recommended to normalize data where each input dimension varies greatly in numerical scale.

3. (2 pts) Consider a neural network with 1 layer. The loss function is binary cross entropy loss:

$$E = -\big(t\log(y) + (1-t)\log(1-y)\big).$$

The activation function is the standard sigmoid, $g(x) = \dfrac{1}{1+e^{-x}}$. Which of the following is correct? Given: x is the input to the layer, w are the layer weights.

(a) $\dfrac{\partial E}{\partial w} = w^\top(y-t)$

(b) $\dfrac{\partial E}{\partial w} = w^\top(t-y)$

(c) $\dfrac{\partial E}{\partial w} = x^\top(y-t)$

(d) $\dfrac{\partial E}{\partial w} = x^\top(t-y)$

4. (2 pts) Recall Rumelhart's design for a regularization term:

$$C = \frac{\|W\|^2}{\|W\|^2+1}.$$

Which of the following intuitions best fits his idea?

(a) Penalize big weights less while penalizing small weights more.

(b) Penalize big weights more while penalizing small weights less.

(c) Penalize big weights and small weights directly proportional to their magnitude.

(d) Penalize big weights and small weights on the same scale.

5. (2 pts) Logistic regression is a _____ regression technique that is used to model data having a _____ outcome.

(a) Linear, binary

(b) Linear, numeric

(c) Nonlinear, binary

(d) Nonlinear, numeric

6. (2 pts) Which of the following is/are true about the momentum update rule?

(a) It has no hyperparameters.

(b) It better avoids local minima by keeping running gradient statistics.

(c) If the network has n parameters, momentum requires that we keep track of $O(n)$ extra parameters.

(d) b and c.

PID:_____

Short answer

Answer each question in no more than 2–3 sentences unless specified otherwise.

a) (3 pts) Describe (1) what L1/L2 regularization are used for and (2) how they differ. In addition, (3) if the complexity term for L2 regularization is $C = \|W\|_2^2$, then what is the closed form for $\frac{\partial C}{\partial w}$ where w is some arbitrary weight in a neural network?

b) (2 pts) List two features of the visual world that are reflected in the design of Convolutional Neural Networks (CNNs). Be clear about what feature in the visual world corresponds to what structure or function in CNNs.

c) (2 pts) What are two advantages of using Stochastic Gradient Descent over Batch learning?

d) (2 pts) What is the difference between PCA (normalized by the standard deviation) and Z-scoring?

e) (2 pts) A student is trying to get a network to classify images as cats vs. dogs. They designed a CNN with a single output neuron. The final output of their network, z, is given by $z = \sigma(\text{ReLU}(a))$, where σ is the logistic function, and he classifies all inputs with a final value $z \geq 0.5$ as cat images. What problem will they encounter?

PID:_____

Problem 4: Backprop

a. (3 pts) Briefly explain why the forward pass is non-linear and the backward pass is linear. Equations will help get full credit.

b. (12 pts) Consider the simple networks below. The initial weights and biases are as shown (biases are shown as a weight from units with a "1" inside). All units use the ReLU activation function $g(a) = \max(a, 0)$. **Note that the inputs and weights are different panel to panel.**

(a) (4 pts) Fill in the activations of the neurons in Figure 1 (after applying ReLU).

U1: _____

U2: _____

U3: _____

U4: _____

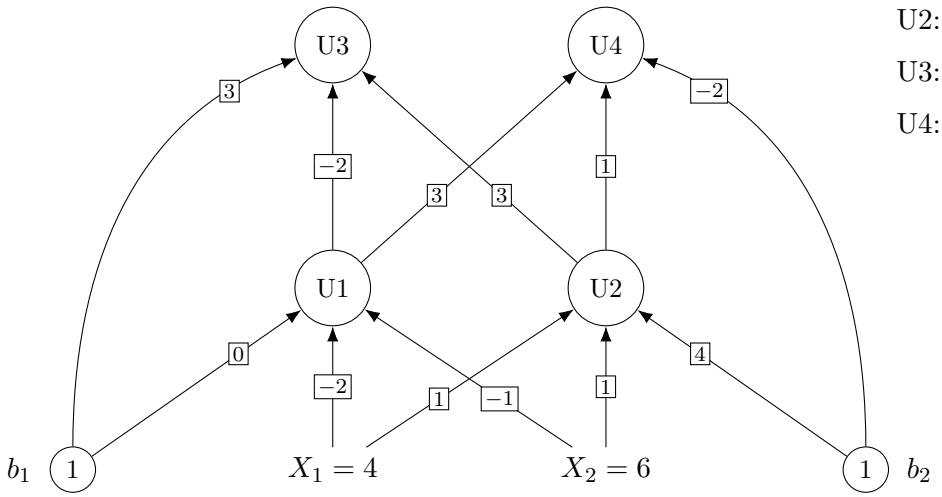

Figure 1: Compute the Activations

(b) (4 pts) Fill in the deltas of the 4 nodes computed via backprop in Figure 2 in the space provided. If a ReLU unit's activity is 0, assume the slope is 0.

δ U3: _____

δ U4: _____

δ U1: _____

δ U2: _____

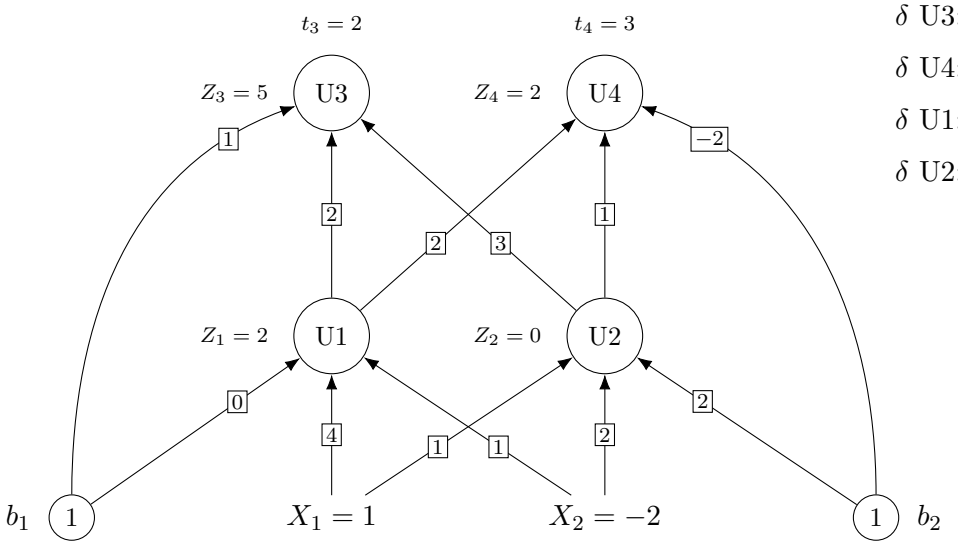

Figure 2: Compute the deltas

(c) (4 pts) Fill in the new weights and biases in Figure 3, assuming the learning rate is 1.

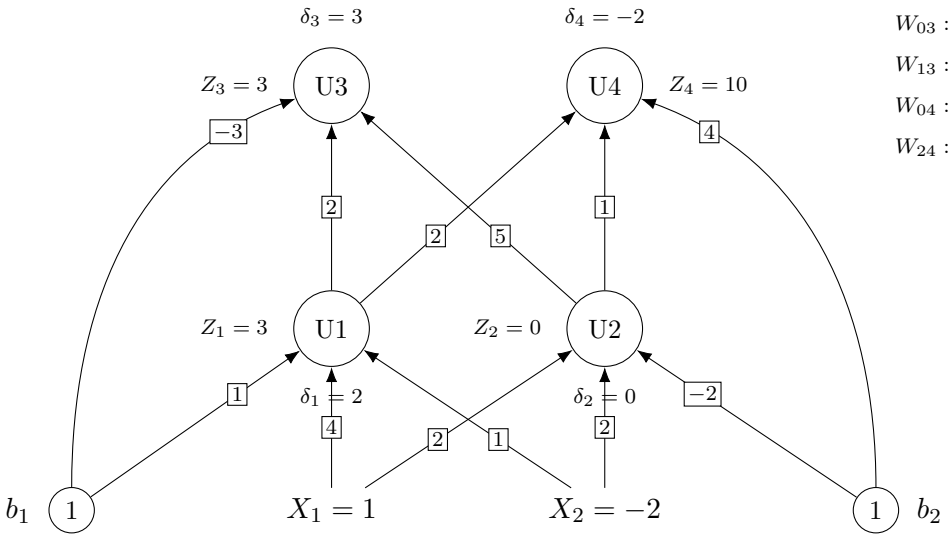

Figure 3: Change the weights

Acknowledgments

All materials in this practice exam were directly adapted from Prof. Gary Cottrell.