CUID: _________ Student Last Name:______________ First Name:_____________

Problem 1. (20 points)

Part 1: (8 points)
Vector $\underline{\mathbf{x}}$ is the least-squares solution for linear system

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$$

when it minimizes

$$||\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}||_2$$

If **A** has linearly independent columns and if $\mathbf{A}^T\mathbf{A}$ is invertible then the system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ has a unique least squares solution

$$\underline{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\underline{\mathbf{b}}$$

If **A** does not have full rank, the least squares problem still has a solution, but it is no longer unique. There are many vectors $\underline{\mathbf{x}}$ that minimize $||\mathbf{A}\underline{\mathbf{x}} - \underline{\mathbf{b}}||_2$.
The optimal solution of the system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}$ is the vector $\underline{\mathbf{x}}$ that has minimum length and let it be denoted by $\underline{\mathbf{x}}^+$.
Then,

$$\underline{\mathbf{x}}^+ = \mathbf{A}^+\underline{\mathbf{b}} \text{ ,} \qquad \text{where } \mathbf{A}^+ \text{ is the pseudo-inverse of } \mathbf{A}.$$

Determine the least squares solution for the following data points which are of the form
([input vector],output)
([4,3], 1)
([8,6], 2)

Part 2: (8 points)
Suppose **A** has orthogonal columns $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots\ldots\ldots\ldots, \mathbf{w}_n$ of lengths (L2 norm) $\sigma_1, \sigma_2, \sigma_3, \ldots\ldots\ldots, \sigma_n$. Calculate $\mathbf{A}^T\mathbf{A}$. What are **U, Σ, V** in the singular value decomposition (SVD) of **A**?

Part 3: (4 points)
Let matrix **A** have n eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots.., \mathbf{e}_n$ which are known and eigenvalues $\lambda_1, \lambda_2, \ldots\ldots\ldots,$ λ_n which are unknown and need to be calculated.
For any **1<=t<=n,** write an expression for λ_t using the known quantities given.

CUID: _________ Student Last Name:______________ First Name:_____________

Problem 2. (20 points)

Regularization is a key concept which facilitates the successful operation of artificial neural networks (ANNs). Consider the “classical” L^2 parameter regularization, where parameter norm penalty $\Omega(\boldsymbol{\theta})$ is used, which only considers the ANN weights **w**.

1. Start with objective function $J(\boldsymbol{\theta};\mathbf{X},\mathbf{y})$ and use hyperparameter α to write a starting expression for the regularized objective function $\tilde{J}(\boldsymbol{\theta};\mathbf{X},\mathbf{y})$ which utilizes $\Omega(\boldsymbol{\theta}) = ½ ||\mathbf{w}||^2_2$. (3 points)
2. What is the set of numbers that α belongs to? (2 point)
3. What is the justification for not taking offsets into account, when studying ANN regularization? (2 point)
4. Write the gradient of the total (regularized) objective function $\tilde{J}(\boldsymbol{\theta};\mathbf{X},\mathbf{y})$, and for a single step for weight update (3 points)

Consider the expression for the objective function with the following quadratic approximation $\hat{J}(\mathbf{w}) = J(\mathbf{w}^*) + ½*(\mathbf{w}-\mathbf{w}^*)^T\mathbf{H}(\mathbf{w}-\mathbf{w}^*)$, in the vicinity of empirically optimal weights **w*.**

5. Write the expression for the location of the minimum of the regularized objective function and derive $\tilde{\mathbf{w}}$ **(**estimated **w**) as a function of **w*** (4 points)
6. Derive and discuss the behavior of $\tilde{\mathbf{w}}$ as α changes. Use the eigen representation and properties of the Hessian. (3 points)
7. What are the effects of L_2 parameter regularization? (3 points)

Note: You need to show derivations. Make sure that vectors and matrices show in your handwriting.

CUID: _________ Student Last Name:______________ First Name:_____________

Problem 3. Short answers (20 points) (no more than 2 sentences per question!)

1. Consider a convolutional layer that takes a 32x32x3 image as input. All 128 filters in the convolutional layer have size 9x9 and there is a 3x3 max-pooling operation after the convolution. What is the size of the output of the max-pooling operation?
2. State two things that are characteristic for stochastic gradient descent methods as opposed to classical (batch) gradient descent.
3. Describe deep learning in two sentences.
4. What is adversarial training?
5. List two examples of data augmentation.
6. Compare the frequency of occurrence of functional minima in low-dimensional and high-dimensional spaces. Provide intuitive explanation.
7. Draw a vectorial figure which illustrates the method of Nesterov momentum.
8. Sketch the behavior of the verification error as the model becomes overfit.
9. Write down the expressions for the softmax activation function and its derivative.
10. What are the advantages of ReLU as compared to the sigmoid activation function?

Problem 4. True/False Questions (20 points)

1. While highway networks are capable of attaining a higher accuracy than Fitnets on CIFAR-10 and MNIST datasets, highway networks need slightly more parameters to do so.

2. Highway networks successfully utilize a gating mechanism to pass information almost unchanged through many layers.

3. While capable of giving impressive results, a significant drawback of generative adversarial networks is that they cannot be trained via backpropagation.

4. Convolutional neural networks require max-pooling to give good results.

5. A big advantage of CNN over MLP is that it allows for fewer model parameters.

6. Weight normalization is inspired by batch normalization, thus it is a probabilistic method that shares batch normalization’s property of adding noise to the gradients.

7. In weight normalization, no additional memory is required and additional computation is quite negligible.

8. The first couple of hidden layers of a CNN captures the high-level content of the image, while lower level content is captured by the latter subsequent layers.

9. Increasing the number of layers while keeping the number of parameters the same always increases the maximum testing accuracy of the network.

10. There cannot be a convolutional layer after a fully connected layer.

CUID: _________ Student Last Name:______________ First Name:_____________

Problem 5. Backpropagation (20 points)

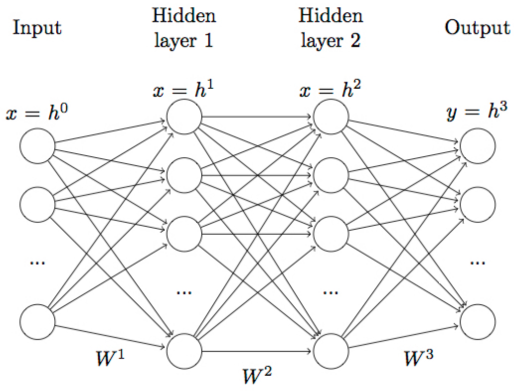
Consider a deep neural network with one input layer, two hidden layers, and one output layer. The input and output layers both have N neurons, while each hidden layer has M neurons. We use sigmoid activation function throughout the entire network, leading to the following feed-forward connection:

$$\mathbf{h}^{l+1} = \sigma(\mathbf{h}^{l}\mathbf{W}^{l+1})$$

where l is the layer index, **h** denotes the layer input, and **W** denotes the weight. We are omitting the bias term in each layer. The entire model is expressed as **y** = f(**x**).

For a given set of inputs $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k\}$ with their associated output $\{\mathbf{t}^1, \mathbf{t}^2, \ldots, \mathbf{t}^k\}$, we use L2 loss as the objective function, which we denote as L_2 norm without *sqrt*:

$$L_{obj} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{N}(\mathbf{t}_j^i - \mathbf{f}(\mathbf{x}_j^i))^2$$



Part 1: (8 points)

Derive $\frac{\partial L}{\partial h^1}$.

Part 2: (8 points)

Residual connections are now widely used to train very deep neural networks. An example of residual connection is like follows: given a (sub)network with input x and output **y** = f(**x**), we add an extra connection directly from the input to the output, i.e., the new output is expressed as **y** = f(**x**) + g(**x**), where g(**x**) could be an identity mapping (g(**x**) = **x**) if the input and output has same shape, or it could be a linear transformation (g(**x**) = **xW**) for changing the shape.

Suppose we change the number of hidden units in the first hidden layer into N, and then add a residual connection between the first hidden layer and the output layer. Derive $\frac{\partial L}{\partial h^1}$.

Part 3: (4 points)

Deep neural networks may have the *vanishing gradient problem*, which means that the gradient for some parameters may vanish near zero. Use your results above, describe why this problem may happen when using sigmoid activation function, what’s the consequence of this problem, and how residual connection may alleviate it. (Hint: look into the property of sigmoid function, and your results for the derivatives.)