Abbreviations:

- BPTT: Back propagation through time
- CNN: Convolution neural network
- FiLM: Feature-wise linear modulation
- FFNN: Feedforward neural network
- LSTM: Long short-term memory
- MSE: Mean squared error
- NN: Neural network
- RNN: Recurrent neural network
- SGD: Stochastic gradient descent

FeedForward Networks

1) In binary classification, one wants to predict a probability $p(y = 1 \mid x)$ that an input x belongs to class 1. If $p(y = 1 \mid x) > 1/2$, one predicts that x is of class 1, otherwise one predicts it to be of class 0. To get a prediction in the interval (0,1), one usually uses the sigmoid output $\sigma(x) = \dfrac{1}{1 + \exp(-x)}$. If no nonlinear layer is added, Show that the NN is equivalent to the logistic regression model.
2) Following the backpropagation algorithm, compute expressions for the gradients of the weights and biased for the network in the XOR example in Sec 6.1, with the MSE loss function.
3) Show that the softmax output reduces to the sigmoid output when the number of classes equals to 2.
4) Suppose a FFNN is trained to infer the parameters of a 1-D conditional mixture model $p(y \mid x) = \sum_{i=1}^{m} p^{(i)}(x)\lambda^{(i)}(x) \ \exp[-\lambda^{(i)}(x)\ y]$, where the number of components m is known, $\lambda^{(i)}(x) > 0$ is the rate parameter of the i-th exponential component at x and $p^{(i)}(x)$ is the mixture probability of component i at x. Design a suitable output layer of the network. What is the corresponding loss function?
5) Argue that the number of parameters (weights and bias) in a NN is not a good measure of model complexity. Specifically, give an example to support or illustrate your reasoning. Any suggestion of a better measure of model complexity?
6) In class, we discussed the intuition why a "Barrel" form of NN shape is used for classification (i.e., expand dimensions to disentangle the classes and then compress dimensions irrelevant to classification). What is the intuition to use a "Barrel" form for regression?
7) List 2 or more advantages of SGD in compared with the ordinary GD, justify your answers.

Regularizations

8) Consider modifying the cross entropy loss by introducing a prior $p(\theta)$ on the weights θ, so that the loss is given by $-\log p(y \mid x, \theta)p(\theta)$, thereby maximizing a Bayesian posterior distribution instead of the likelihood. Show that a Gaussian prior gives rise to L_2 regularization and a Laplace prior to L_1 regularization.
9) Consider a binary classification problem using NN with highly imbalanced data - sample size of one class is much larger than that of the other class. It is expected that the NNs formulated/trained in the standard settings tend to be biased toward the majority class. Assume that you cannot collect more data, propose 2 ways to mitigate the potential problems raised by imbalanced data and justify your proposals. Hint: You can propose strategies that apply to the data, the loss function, the regularization and validation schemes, etc.
10) Consider injecting noise $\epsilon \sim N(0,\sigma^2)$ at the input $x \in \mathbb{R}$ of a fully connected FF NN. Assume that the output layer is scalar and the network is trained by minimizing the MSE. That is, in the limit of infinite sample $(x, y) \sim p(x, y)$, we minimize $J(\theta; x, y, \epsilon) = E_{p(x,y,\epsilon)}[\hat{y}(x + \epsilon, \theta) - y)^2]$. **(i)** Show that $\hat{y}(x + \epsilon) = \hat{y}(x) + \epsilon\,\hat{y}'(x) + \epsilon^2\hat{y}''(x)/2 + O(\epsilon^3)$. **(ii)** Show that $J(\theta; x, y, \epsilon) \approx E_{p(x,y)}[(\hat{y}(x) - y)^2 + \sigma^2\hat{y}'(x)^2 + \sigma^2(\hat{y}(x) - y)\hat{y}''(x)]$ for small σ^2. **(iii)** A fundamental result of statistical learning is that the minimum of the MSE with respect to $\hat{y}$ is of

the form $\hat{y}_{opt}(x) = E[y\,|\,x]$. Show that at this minimum, we have $E_{p(x,y)}[\sigma^2(E[y\,|\,x] - y)\hat{y}''(x)] = 0$. Hence, close to the minimum of the MSE and for small σ^2, we have $J(\theta; x, y, \epsilon) \approx E_{p(x,y)}[(\hat{y}(x) - y)^2 + \sigma^2\hat{y}'(x)^2] = E_{\text{MSE}} + \sigma^2 E_{\text{Reg}}$.

11) Eqs. 7.50-51 in the book show that the variance can be big when the models are correlated with covariance $c > 0$. Reason why this does not affect the application of bagging to FF NN. Then explain what are the main advantage(s) of dropout over bagging in deep learning.
12) This question refers to the "Supplementary reading for Dropout" in the course page. NOTE: In this note, p is the probability to set the neuron variable to zero. **A)** Explain in terms of the concept of SGD how the training procedure in the section "2.1 Detailed workflow of dropout" works to minimize the dropout loss $\sum_{\mu} p(\mu)J(Y, X, \theta)$, where $\mu \sim p(\mu)$ is the mask vector, $p(\mu)$ is the Bernoulli distribution, and $J(Y, X, \theta)$ is the usual cross entropy loss. **B)** It says in the note that "…it's imperative to rescale the vector $y_1, y_2, \cdots, y_{1000}$ by multiplying it with $1/(1-p)$ …", explain why this rescaling is needed. **C)** In Fig. 5b of the note, it says that during the test phase, it is the weights, instead of the neurons during the training phase (see Fig. 5a), that are multiplied by the probability p. Explain why this is done so.

Optimizations

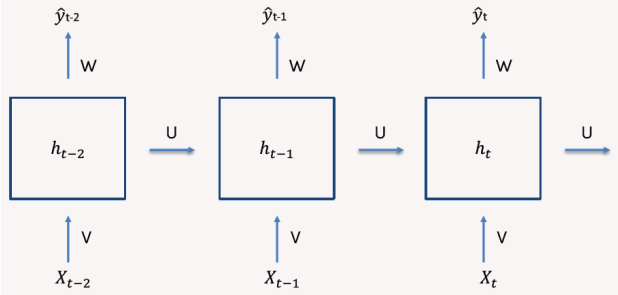
13) In the Adam algorithm (Algorithm, 8.7), the accumulated 1st and 2nd moments are normalized ($\hat{s} \leftarrow s/(1-\rho_1^t)$ and $\hat{r} \leftarrow r/(1-\rho_2^t)$). Discuss what is the purpose of this normalization.
14) In the Adam algorithm (Algorithm, 8.7), explain how the 1st, 2nd moments ($\hat{s}, \hat{r}$) and the adaptive learning rate $\epsilon/\sqrt{\hat{r}}$ behave when the training **A)** enters the bottom of a deep local minimum; **B)** enters a relatively flat plateau of the loss function. **C)** Give two possible scenarios that the parameter update $\Delta\theta = -\epsilon\hat{s}/\sqrt{\hat{r}}$ can become small, i.e., either reaching convergence or failing to learn. Justify your answers.

Convolutional NN

15) Show that convolution is translational equivariant. Is convolution reflective equivariant, i.e., equivariant under the transformation $i \rightarrow -i$ and/or $j \rightarrow -j$ for pixels in a 2D image?
16) Consider the following setup: An input with dimension 112 (pixels) x 112 (pixels) x 128 (channels), a 3 x 3 convolutional kernel, stride = 2, no zero padding, and number of channels in the feature map = 256. How many pixels does the feature map have? What is the total number of trainable parameters (weights + biases) of the kernel?
17) Explain why in the Batch Normalization $BN(x) = \gamma \odot \dfrac{x - \hat{\mu}_B}{\hat{\sigma}_B} + \beta$ only the mini-batch, instead of the full training set, is used to evaluate the sample mean $\hat{\mu}_B$ and standard deviation $\hat{\sigma}_B$? If x has dimension 112 (pixels) x 112 (pixels) x 128 (channels), what is the total number of trainable parameters in the above Batch Normalization?
18) Consider a 2D input feature map X where each pixel intensity is an independent and identically distributed (i.i.d.) random variable sampled from the standard normal distribution $N(0,1)$. Now a 2×2 Max Pooling with stride 2 (i.e. non-overlapping windows) is applied. **A)** Calculate the expected value $E(Y)$ of a single pixel in the output feature map Y. (Hint: To find the expectation, you need to use the probability density function of the maximum of n i.i.d. Gaussian variables). **B)** Based on your results, explain why a ReLU activation placed after a Max Pooling layer might be less statistically efficient than one placed before it, assuming no Batch Normalization is used.
19) Provide a simple example to illustrate why the up convolution operation used in the U-Net is also called "transposed" convolution, i.e., illustrate what is transposed. Hint: No need to include channels.

Recurrent NN & BPTT

20) Consider the following RNN, write down with clear steps the BPTT derivative $\partial L_t/\partial V$, similar to that of $\partial L_t/\partial W$ discussed in the class. Does the vanishing/exploding gradient problem also exist in $\partial L_t/\partial V$?



21) Explain using the BPTT derivatives $\partial L_t/\partial W$ and $\partial L_t/\partial V$ above why it is difficult for the original RNN to learn about long term dependence.
22) In class, we have discussed that the problem of vanishing/exploding gradient occurs in vanilla RNNs due to the recurrent multiplication of the derivatives $\partial h_t/\partial h_{t-1}$. To investigate the same issue in LSTM, we can look at the recurrent multiplication of the derivatives $\partial C_t/\partial C_{t-1}$, with the cell state $C_t = f_t C_{t-1} + i_t \tilde{C}_t$. Use the notation of LSTM discussed in the class and the following chain rule: $\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial f_t}\frac{\partial f_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial i_t}\frac{\partial i_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial \tilde{C}_t}\frac{\partial \tilde{C}_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial C_{t-1}}$, derive an expression for $\partial C_t/\partial C_{t-1}$ and discuss if the problem of vanishing/exploding gradient still occur in LSTM. For simplicity, you can assume the number of the hidden states $h = 1$.

Attention & Beyond

23) Let $X \in \mathbb{R}^{N\times D}$ be the input to a multi-head self attention with h heads and with the scaled dot product attention score function, where N is the number of tokens in the sequence and D is the embedding dimension. Derive the explicit equation for the output $Y \in \mathbb{R}^{N\times D}$ of the multi-head self attention in terms of the input and the appropriate trainable weight matrices.
24) Let the queries $q \in R^d$ and keys $k \in R^d$ are independent random vectors whose components are independent and identically distributed random variables with zero mean and unit variance. **A)** Derive the means and variance of the dot product $Z = q \cdot k$. **B)** What is the asymptotic distribution of the dot product as $d \to \infty$? c) Let $S(Z_i)$ be the softmax output for a specific key i among a set of N keys, use the above results to prove that without the usual scaling factor $1/\sqrt{d}$, the expected value of the maximum attention weight approach 1 as d becomes large, collapsing the distribution to a point mass.
25) In class, we discussed conditioning using FiLM and cross-attention: $FiLM(x\,|\,c) = \gamma(c)x + \beta(c)$; and cross-attention uses c as keys and values, and x as queries. Here c is the conditioning vector. **A)** Let $x \in R^{N\times D}$ (a sequence of N tokens) and conditioning vector $c \in R^D$. What are the number of trainable parameters applied to x for a single FiLM layer and a single cross-attention layer, respectively? **B)** Is it possible to impose constraints on the projection matrices (W_Q, W_K, W_V) to make a cross-attention layer conditioning identical to a FiLM conditioning applied independently to each token? Justify your answer.