Due Date : February 22nd, 24:00
Student : Ziqiang Wang

1. **Selection of Activation Function (10 pts)** We will compare two different activation functions in the following question. Recall the definition of $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
   (a) (2 pts) Find the derivative of the sigmoid function $\sigma'(x)$ and express it in terms of the sigmoid function $\sigma(x)$.
   (b) (2 pts) Find the derivative of the $\tanh'(x)$ function and express it in terms of the $\tanh(x)$ function.
   (c) (2 pts) Upper bound the value of $\sigma'(x)$ with a constant (you can use AM–GM inequality).
   (d) (2 pts) Upper bound the value of $\tanh'(x)$ with a constant (you can use GM-HM inequality or the property that the square of real number is always non-negative).
   (e) (2 pts) Compare the two upper bounds and explain what impact would this difference have on optimization.

Useful inequalities:

Inequality of Arithmetic and Geometric Means (AM-GM)

$$\frac{x_1 + x_2 + \dots x_n}{n} \geq \sqrt[n]{x_1 x_2 \dots x_n} \tag{1}$$

Inequality of Geometric and Harmonic Means (GM-HM)

$$\sqrt[n]{x_1 x_2 \dots x_n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots \frac{1}{x_n}} \tag{2}$$

The above inequalities hold for any real positive numbers $x_1, x_2, \dots x_n$ with equality if and only if $x_1 = x_2 = \dots = x_n$.

My Answer:

(a)

$$\begin{aligned}
\sigma'(x) &= \frac{d}{dx}\frac{1}{1+e^{-x}} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \sigma(x) \cdot (1 - \sigma(x))
\end{aligned}$$

(b)

$$\begin{aligned}
\tanh'(x) &= \frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} \\
&= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\
&= 1 - \tanh^2(x)
\end{aligned}$$

(c)

$$\begin{aligned}\sigma(x)\cdot(1-\sigma(x)) &= \sigma(x)\cdot\left(1-\frac{1}{1+e^{-x}}\right)\\ &= \sigma(x)\cdot\frac{e^{-x}}{1+e^{-x}}\\ &\leq \frac{\sigma(x)+\frac{e^{-x}}{1+e^{-x}}}{2^2}=\frac{1}{4}\end{aligned}$$

where the inequality uses AM-GM.

(d)

$$\begin{aligned}\tanh'(x) &= \frac{(e^x+e^{-x})(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2}\\ &= \frac{2^2}{(e^x+e^{-x})^2}\\ &\leq \frac{1}{e^x}\cdot\frac{1}{e^{-x}}=1\end{aligned}$$

where the inequality uses GM-HM.

(e) The upper bound for $\sigma'(x)$ is smaller than that for $\tanh'(x)$. This means that the gradients of the tanh function can be larger than those of the sigmoid function. In practice, this can result in faster convergence during optimization. However, the choice of activation function should ultimately depend on the specific problem at hand.

2. **Cross Entropy Properties (12 pts)**

Cross-entropy loss function (a popular loss function) is given by:

$$\mathrm{CE}(p,x) = -x\log(p)-(1-x)\log(1-p)$$

.

Please refer to this loss for (a) and (b) parts.

(a) (2 pts) **Cross Entropy and Maximum Likelihood** For this derivation, we assume that x is binary, i.e. $x\in\{0,1\}$. Derive the cross-entropy loss function using the maximum likelihood principle for $x\in\{0,1\}$.

(b) (2 pts) **Cross Entropy and KL divergence** Suggest a probabilistic interpretation of the cross-entropy loss function when $x\in(0,1)$. (Hint: KL divergence between two distributions)

(c) (4 pts) **Discrete distribution - Maximum Entropy** Let X be a random variable which takes n values with probabilities $p_1,p_2,\ldots,p_n$ with $p_i>0,\forall i$. What is the distribution that maximizes entropy $H(X)=-\sum_{i=1}^{n}p_i\log p_i$? Derive the upper bound for the entropy $H(X)$ expressed as a function of n. (Hint : use Jensen Inequality)

(d) (4 pts) **Continuous distribution (known mean μ and variance σ^2) - Maximum Entropy** Given mean μ and variance σ^2, what is the continuous distribution that maximizes differential entropy $h(X)=-\int_x f(x)\log f(x)dx$? Prove it.

My Answer:

(a) The probability function of one sample $x \sim \text{Bernoulli}(p)$ is :

$$P(x) = p^x(1-p)^{1-x}$$

The likelihood of observing a set of n samples $x^1, x^2, \ldots, x^m$ is :

$$L(p) = \prod_{i=1}^{n} P(x^i) = \prod_{i=1}^{n} p^{x^i}(1-p)^{1-x^i}$$

Taking the negative logarithm of the likelihood, we get :

$$\begin{aligned} -\log L(p) &= -\log \prod_{i=1}^{n} P(x^{(i)}) \\ &= -\sum_{i=1}^{n} \log P(x^i) \\ &= -\sum_{i=1}^{n} \big(x^i \log p + (1-x^i)\log(1-p)\big) \end{aligned}$$

Therefore, for one sample, the cross-entropy loss function is :

$$\text{CE}(p, x) = -x\log(p) - (1-x)\log(1-p)$$

(b) Based on the condition in (a), we can consider the probability distribution $Q(x)$ where $Q(1) = x$ and $Q(0) = 1 - x$. The cross-entropy loss between the true distribution $P(x)$ and $Q(x)$ can be written as :

$$\text{CE}(P, Q) = -\sum_{x \in 0,1} P(x)\log Q(x) = -x\log p - (1-x)\log(1-p)$$

This cross-entropy can be transformed to :

$$\begin{aligned} \text{CE}(P, Q) &= -\sum_{x \in 0,1} P(x)\log Q(x) \\ &= -\sum_{x \in 0,1} P(x)\log P(x) + P(x)\log\frac{P(x)}{Q(x)} \\ &= H(P) + D_{\text{KL}}(P||Q) \end{aligned}$$

where $H(P)$ is the entropy of the distribution $P(x)$ and $D_{\text{KL}}(P||Q)$ is the KL divergence between $P(x)$ and $Q(x)$. Therefore, minimizing the cross-entropy $\text{CE}(P, Q)$ is equivalent to minimizing the KL $D_{\text{KL}}(P||Q)$.

(c) In condition of $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for all $i = 1, \ldots, n$, we can use Jensen's inequality :

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\left(\sum_{i=1}^{n} p_i\right)\log\left(\frac{\sum_{i=1}^{n} p_i}{n}\right) = -\log\frac{1}{n} = \log n.$$

The equality holds if and only if $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$. Therefore, the uniform distribution over the n values of X maximizes the entropy. And the upper bound for entropy $H(X)$ is $\log n$.

(d) The objective is to maximize $h(X) = -\int_x f(x)\log f(x)dx$ subject to :

$$\int_x f(x)dx = 1 \text{ and } \int_x xf(x)dx = \mu \text{ and } \int_x (x-\mu)^2 f(x)dx = \sigma^2$$

Introduce Lagrange multipliers λ_1, λ_2, and λ_3 for each of the constraints and form the Lagrangian :

$$L(f, \lambda_1, \lambda_2, \lambda_3) = -\int_x f(x)\log f(x)dx + \lambda_1\left(\int_x f(x)dx - 1\right) + \lambda_2\left(\int_x xf(x)dx - \mu\right) + \lambda_3\left(\int_x (x-\mu)^2 f(x)dx - \sigma^2\right)$$

Taking the derivative of the Lagrangian with respect to $f(x)$ and setting it equal to zero yields :

$$\frac{\partial L(f, \lambda_1, \lambda_2, \lambda_3)}{\partial f(x)} = -1 - \log f(x) - \lambda_1 - \lambda_2 x - \lambda_3(x-\mu)^2 = 0$$

Solving for $f(x)$, we get :

$$\begin{aligned} f(x) &= \exp\left(-1 - \lambda_1 - \lambda_2 x - \lambda_3(x-\mu)^2\right) \\ &= \exp(-1-\lambda_1)\cdot\exp\left(\lambda_2 x - \lambda_3(x-\mu)^2\right) \\ &= C\exp\left(-\lambda_3(x - \mu + \frac{\lambda_2}{2\lambda_3})^2\right), \end{aligned}$$

where $C > 0$.
$f(x)$ is symmetric about $\mu - \frac{\lambda_2}{2\lambda_3}$, thus $\lambda_2 = 0$ and $f(x) = C\exp\left(-\lambda_3(x-\mu)^2\right)$. Then according to $\int_x f(x)dx = 1$, we get :

$$\begin{aligned} \int_x C\exp\left(-\lambda_3(x-\mu)^2\right)dx &= 1 \\ \int_x C\exp\left(-\lambda_3 z^2\right)dz &= 1 \\ \int_x \exp\left(-\lambda_3 z^2\right)dz &= \frac{1}{C} \\ \sqrt{\frac{\pi}{-\lambda_3}} &= \frac{1}{C} \end{aligned}$$

According to $\int_x (x-\mu)^2 f(x)dx = \sigma^2$, we get :

$$\begin{aligned} \int_x C\exp\left(-\lambda_3(x-\mu)^2\right)(x-\mu)^2 dx &= \sigma^2 \\ \int_x C\exp\left(-\lambda_3 z^2\right)z^2 dz &= \sigma^2 \\ \int_x \exp\left(-\lambda_3 z^2\right)z^2 dz &= \frac{\sigma^2}{C} \\ \frac{1}{2}\sqrt{\frac{\pi}{-\lambda_3^3}} &= \frac{\sigma^2}{C} \\ \sqrt{\frac{\pi}{-\lambda_3}} &= \frac{-2\lambda_3\sigma^2}{C} \end{aligned}$$

Putting together :

$$\sqrt{\frac{\pi}{-\lambda_3}} = \frac{1}{C} = \frac{-2\lambda_3\sigma^2}{C}$$
$$1 = -2\lambda_3\sigma^2$$
$$\lambda_3 = -\frac{1}{2\sigma^2}$$

Then,

$$\sqrt{\frac{\pi}{\lambda_3}} = \frac{1}{C}$$
$$\sqrt{\frac{\pi}{\frac{1}{2\sigma^2}}} = \frac{1}{C}$$
$$C = \frac{1}{\sqrt{2\pi\sigma^2}}$$

Finally, plugging in,

$$f(x) = C\exp\left(\lambda_3(x-\mu)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Therefore, the normal distribution with mean μ and variance σ^2 maximizes the differential entropy $h(X)$.

3. **Output size and Parameters of Convolution Layers (5 pts)**
Consider a 3 hidden-layer convolutional neural network. Assume the input is a color image of size 128×128 in the RGB representation. The first layer convolves 64 8×8 kernels with the input, using a stride of 2 and zero-padding of 4. The second layer downsamples the output of the first layer with a 2×2 non-overlapping max pooling. The third layer convolves 128 4×4 kernels with a stride of 2 and zero-padding of 2.
   (a) (3 pts) What is the dimensionality of the output of the third layer?
   (b) (2 pts) Not including the biases, how many parameters are needed for the last layer?

   My Answer:

   (a) The output of the third layer will have a dimensionality of $128 \times 17 \times 17$.
   The dimensionality of input image is $3\times128\times128$. After the first layer, the dimensionality is $64 \times (\frac{128}{2}+1) \times (\frac{128}{2}+1) = 64 \times 65 \times 65$. Then through the downsampling layer, the dimensionality becomes $64\times32\times32$. Finally, the dimensionality of output from the third layer is $128 \times (\frac{32}{2}+1) \times (\frac{32}{2}+1) = 128 \times 17 \times 17$.
   (b) $64 \times 128 \times 4 \times 4 = 131,072$

4. **MLP Mixer (16 pts)**

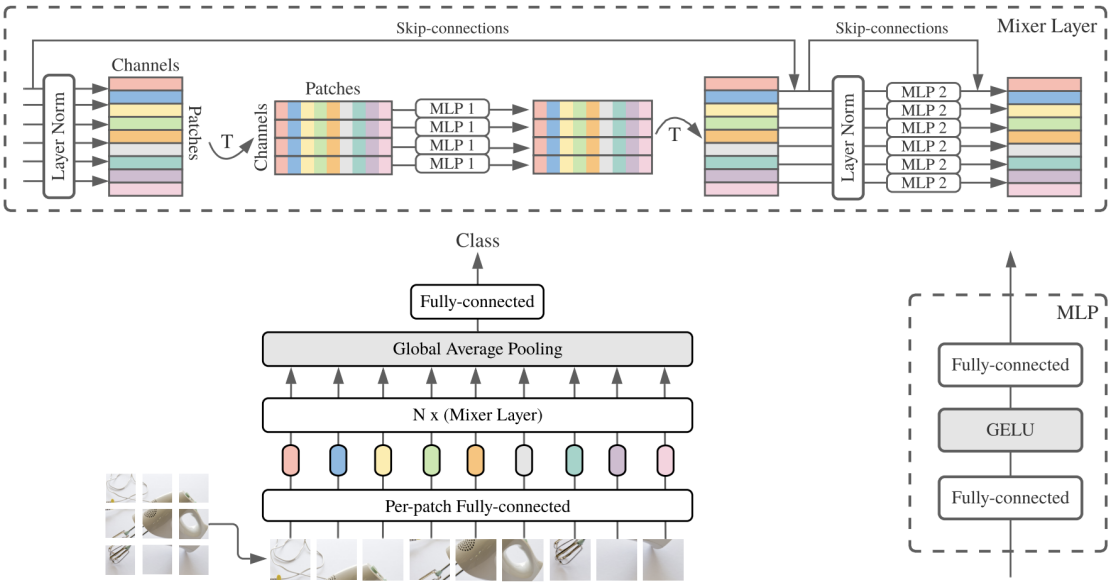

FIGURE 1 – (Borrowed from the MLPMixer paper.) MLP-Mixer consists of per-patch linear embeddings, Mixer layers, and a classifier head. Mixer layers contain one token-mixing MLP and one channel-mixing MLP, each consisting of two fully-connected layers and a GELU nonlinearity. Other components include : skip-connections, dropout, and layer norm on the channels.

(a) (2 pts) **MLP Mixer Dimensions** Let's assume that Mixer architecture is being applied to an input image of size 64×64. The Mixer's output is of size 16×128. Determine the patch resolution P, number of patches S, as hidden dimension C(channels).

(b) (2 pts) **MLP Mixer Complexity** Show that the computational complexity of the MLP Mixer is linear in terms of number of input patches.

(c) (6 pts) **Input Transformation - Channel Mixing MLP** Consider the following scenario : The original input image A is of size 9×9. We convert the input image into non-overlapping patches of size 3×3, and then linearly project all patches with the same projection matrix. The result of these operations is a matrix X of size 9×6. Then we apply the *channel-mixing MLP* that acts on rows of X, and is shared across all rows. The result of this operation is matrix U size 9×6. Now consider a modified image A such that $A_{\text{modified}} = PA$, where we define matrix P in the following manner:

$$P = \begin{bmatrix} e_{\pi(1)} \\ e_{\pi(2)} \\ \vdots \\ e_{\pi(9)} \end{bmatrix} \tag{3}$$

Here e_k is k-th basis vector and π represents the permutation of indices from $1 \ldots 9$. Find all possible P such that by permuting rows of U_{modified} we can get back matrix U.

(d) (6 pts) Select one of your solutions for P and find $P_{reverse}$ such that $P_{\text{reverse}} U_{\text{modified}} = U$.

My Answer:

(a) Patch resolution $P = 16 \times 16$, number of patches $S = 16$, and hidden dimension $C = 128$.

(b) For the linear projecting and channel-mixing parts, the patches are processed separately and in parallel, thus the calculation amount in the two parts is linearly in terms of number of input patches N.

For the token-mixing part, the complexity is $O(2NC_hC)$, where C_h is the hidden width in the token-mixing block, and C is the hidden dimension. Therefore, the complexity is also linear in terms of N in this part.

Overall, the complexity of the MLPMixer is linear in terms of number of input patches.

(c) The linear projecting and channel-mixing parts separately process the patches, which means that they do not change the order of patches. And rows of matrix U just correspond to patches respectively, so if we only change the order of the patches without shuffling the pixels within a patch, the order of the rows in U is permuted correspondingly, then we can get back U only by permutation.

PA permutes rows of image A, thus the valid solutions for P are those that permute the order of the patch rows, then we can get $3 \times 2 \times 1 = 6$ solutions for P :

$$
\begin{aligned}
&[e_1^T, e_2^T, e_3^T, e_4^T, e_5^T, e_6^T, e_7^T, e_8^T, e_9^T]^T,\\
&[e_1^T, e_2^T, e_3^T, e_7^T, e_8^T, e_9^T, e_4^T, e_5^T, e_6^T]^T,\\
&[e_4^T, e_5^T, e_6^T, e_1^T, e_2^T, e_3^T, e_7^T, e_8^T, e_9^T]^T,\\
&[e_4^T, e_5^T, e_6^T, e_7^T, e_8^T, e_9^T, e_1^T, e_2^T, e_3^T]^T,\\
&[e_7^T, e_8^T, e_9^T, e_1^T, e_2^T, e_3^T, e_4^T, e_5^T, e_6^T]^T,\\
&[e_7^T, e_8^T, e_9^T, e_4^T, e_5^T, e_6^T, e_1^T, e_2^T, e_3^T]^T.
\end{aligned}
$$

(d) Based on the fact that the linear projecting and channel-mixing parts do not change the order of patches, for $P = [e_1^T, e_2^T, e_3^T, e_7^T, e_8^T, e_9^T, e_4^T, e_5^T, e_6^T]^T$, PA achieves $A_{modified} = A.permute(0, 1, 2, 6, 7, 8, 3, 4, 5)$, so $U_{modified} = U.permute(0, 1, 2, 6, 7, 8, 3, 4, 5)$. Therefore, $U = U_{modified}.permute(0, 1, 2, 6, 7, 8, 3, 4, 5) = P_{reverse}U_{modified}$, then we get $P_{reverse} = [e_1^T, e_2^T, e_3^T, e_7^T, e_8^T, e_9^T, e_4^T, e_5^T, e_6^T]$.

Actually, every solution for P in the last question is mutually inverse with itself.

5. **Gradient Descent Convergence (12 pts)**

(a) (6 pts) **Convex Function Convergence** Consider the following function:

$$
f(x) = \begin{cases} \frac{3}{4}(1-x)^2 - 2(1-x) & \text{if } x > 1 \\ \frac{3}{4}(1+x)^2 - 2(1+x) & \text{if } x < -1 \\ x^2 - 1 & \text{otherwise} \end{cases} \tag{4}
$$

Show that f is a convex function. Find its unique minimizer and its gradient. Consider the following algorithm : $x_t = x_{t-1} - \eta f'(x_{t-1})$ where $\eta = 1$. Will this algorithm converge to a stationary point if it starts at point x_0, where $x_0 > 1$? Why or why not?

(b) (6 pts) **Prove Convergence of Gradient Descent to Stationary Point in Non-Convex case** Suppose we are trying to minimize the function $F(w)$ that is L-smooth. Let F_* be the minimal function value (i.e. the value at the global minima). Using $\eta = \frac{1}{L}$, prove that gradient descent will "almost" converge to a stationary point in a bounded (and polynomial) number of steps. Precisely,

$$
\min_{k<K} \|\nabla F(w^{(k)})\|^2 \leq \frac{2L}{K}(F(w^{(0)}) - F_*) \tag{5}
$$

Hints:

i. L-smoothness implies that:

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \eta\|\nabla F(w^{(k)})\|^2 + \frac{1}{2}\eta^2 L\|\nabla F(w^{(k)})\|^2 \tag{6}$$

Combine this with $\eta = \frac{1}{L}$

ii. Use the fact that the minimum of a sequence of elements is less than the average of the sequence.

My Answer:

(a) To show that f is convex, we need to show that its second derivative is non-negative. First, we compute the first derivative of f :

$$f'(x) = \begin{cases} \frac{3}{2}x + \frac{1}{2} & \text{if } x > 1 \\ -\frac{3}{2}x - \frac{1}{2} & \text{if } x < -1 \\ 2x & \text{otherwise} \end{cases} \tag{7}$$

Then, the second derivative of f is :

$$f''(x) = \begin{cases} \frac{3}{2} & \text{if } x > 1 \\ \frac{3}{2} & \text{if } x < -1 \\ 2 & \text{otherwise} \end{cases} \tag{8}$$

Since $f''(x) \geq 0$ for all x, f is convex.
To find the minimizer of f, we need to solve $f'(x) = 0$. For $x > 1$, $\frac{3}{2}x + \frac{1}{2} > 0$. For $x < -1$, $-\frac{3}{2} - \frac{1}{2} > 0$. For $-1 \leq x \leq 1$, we have $2x = 0$, which implies $x = 0$. Thus, the unique minimizer of f is $x = 0$. And the gradient of f is $f'(x)$ above.
The algorithm $x_t = x_{t-1} - \eta f'(x_{t-1})$ with $\eta = 1$ is gradient descent with a fixed updating rate. Since f has a unique global minimum at $x = 0$, any trajectory of gradient descent will converge to the point. Therefore, if the algorithm starts at $x_0 > 1$, it will converge to the stationary point $x = 0$.

(b) From the Hint i, we can get :

$$\frac{1}{2L}|\nabla F(w^{(k)})|^2 \leq F(w^{(k)}) - F(w^{(k+1)}). \tag{9}$$

Then, using the Hint ii and the inequality 9, we can get :

$$\begin{aligned} \min_{k<K} |\nabla F(w^{(k)})|^2 &\leq \frac{1}{K}\sum_{k=0}^{K-1} |\nabla F(w^{(k)})|^2 \\ &\leq \frac{2L}{K}\sum_{k=0}^{K-1} \left(F(w^{(k)}) - F(w^{(k+1)})\right) \\ &= \frac{2L}{K}\left(F(w^{(0)}) - F(w^{(K)})\right) \end{aligned}$$

Finally, using the fact that the minimal function value $F_* \leq F(w^{(K)})$, we obtain the desired inequality :

$$\min_{k<K} \|\nabla F(w^{(k)})\|^2 \leq \frac{2L}{K}(F(w^{(0)}) - F_*)$$