**Problem 1**
Represent the derivative of the following scalar functions with respect to $\mathbf{X} \in \mathbb{R}^{D \times D}$.

(a) $f(\mathbf{X}) = tr(\mathbf{X}^2)$. Here, $tr(A)$ is the trace of a square matrix $A$.

(b) $g(\mathbf{X}) = tr(\mathbf{X}^3)$.

(c) $h(\mathbf{X}) = tr(\mathbf{X}^k)$ for $k \in \mathbb{N}$.

**Problem 2**

In order to alleviate overfitting in logistic regression, regularization technique can be used with the $L_2$-norm of the weight parameter, $||\mathbf{w}||^2 = \sum_{i=1}^{D} w_i^2$. When we are given $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^D$ and $y_1, \dots, y_N \in \{0, 1\}$ for training set, we want to derive the update rule for $\mathbf{w} \in \mathbb{R}^D$ in order to minimize the following loss function $L(\mathbf{w})$ having $L_2$-norm,

$$
\begin{aligned}
L(\mathbf{w}) &= \sum_{i=1}^{N} \Big( -y_i \ln f(\mathbf{x}_i; \mathbf{w}) - (1 - y_i) \ln(1 - f(\mathbf{x}_i; \mathbf{w})) \Big) + ||\mathbf{w}||^2. && (1) \\
f(\mathbf{x}; \mathbf{w}) &= \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})} && (2)
\end{aligned}
$$

(1) Derive the update rule for $\mathbf{w}$ when we use gradient descent.

(2) Discuss the effect of $L_2$-norm regularization.

**Problem 3**

Consider the following function,

$$f(\mathbf{x};\mathbf{w}) = \frac{1-\exp(-\mathbf{w}^T\mathbf{x})}{1+\exp(-\mathbf{w}^T\mathbf{x})}, \tag{3}$$

having the shape in Fig. 1.

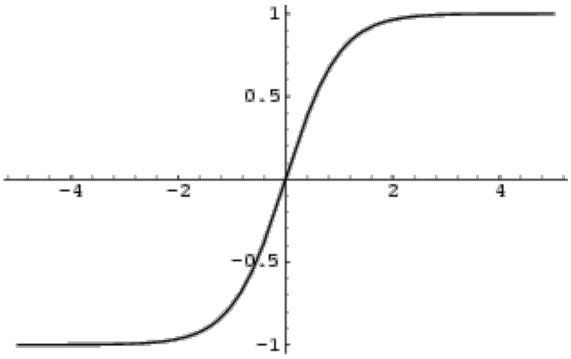


Figure 1: $f(\mathbf{x};\mathbf{w})$

Find the gradient descent update rule for $\mathbf{w}$ to minimize the loss

$$L = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i;\mathbf{w}) - y_i)^2. \tag{4}$$

Here, we use $N$ number of $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{-1, 1\}$ for $i \in \{1, \dots, N\}$ which are given in advance.

**Problem 4**

For a given data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, derive the closed-form solution for $\mathbf{w}$ that maximizes the following probability $\boldsymbol{P}$.

$$\boldsymbol{P} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - f(\mathbf{x}_i; \mathbf{w}))^2\right), \tag{5}$$

with $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^\top \mathbf{x}$ for $\mathbf{x}, \mathbf{w} \in \mathbb{R}^D$.

**Problem 5**

The Frobenius dot product $\langle \mathbf{A}, \mathbf{B} \rangle$ is defined as

$$\langle \mathbf{A}, \mathbf{B} \rangle = tr(\mathbf{A}^\top \mathbf{B}), \tag{6}$$

for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ using the scalar function $tr(\cdot)$ for the trace of a matrix,

$$tr(\mathbf{M}) = \sum_{i}^{N} M_{ii}, \quad \text{for } \mathbf{M} \in \mathbb{R}^{N \times N}. \tag{7}$$

Find the derivative of $\langle \mathbf{A}, \mathbf{B} \rangle$ with respect to $\mathbf{A}$ using the definition of the matrix derivative:

$$\left[ \frac{d}{d\mathbf{A}} \langle \mathbf{A}, \mathbf{B} \rangle \right]_{ij} = \frac{\partial}{\partial A_{ij}} \langle \mathbf{A}, \mathbf{B} \rangle. \tag{8}$$

**Problem 6**

We are given a scalar function $f(\mathbf{X}) = \mathbf{w}^\top \mathbf{X} \mathbf{w}$ for $\mathbf{w} \in \mathbb{R}^D, \mathbf{X} \in \mathbb{R}^{D \times D}$. Find the derivative of $f(\mathbf{X})$ with respect to the vector $\mathbf{w}$. Use the definition of the vector derivative $\left[\frac{df}{d\mathbf{w}}\right]_k = \frac{\partial f}{\partial w_k}$, where $w_k$ is the $k$-th element of $\mathbf{w}$.

## Problem 7

For a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$, Suppose that we have a two-layer neural network with residual connections shown in Figure 2. Each component is given as follows:

$$L(W) = \frac{1}{2}\sum_{i=1}^{N}(g(\mathbf{x}_i) - y_i)^2 \tag{9}$$

$$g(\mathbf{x}) = \sigma\left(\sum_{d=1}^{D} w_{2,d} \cdot h_d(\mathbf{x})\right) \tag{10}$$

$$h_i(\mathbf{x}) = x_i + \sigma\left(\sum_{m=1}^{M} w_{1,i,m} \cdot z_m(\mathbf{x})\right) \tag{11}$$

$$z_i(\mathbf{x}) = \sigma\left(\sum_{d=1}^{D} w_{0,i,d} \cdot x_d\right) \tag{12}$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{13}$$

Here, the residual connection from each $x_i$ to the $h_i$ node in the second layer is represented in Eq. (11). The weight $w_{i,j,k}$ connects the $j$-th node of $i$-th layer to the $k$-th node of $(i+1)$-th layer.

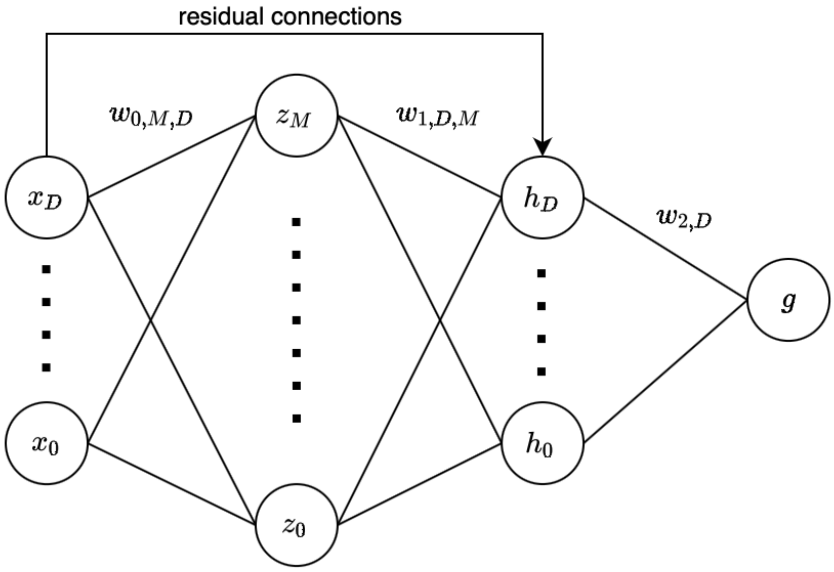


Figure 2: A two-layer network

(a) Calculate derivative of $L$ with respect to $h_j$.

(b) Calculate derivative of $L$ with respect to $w_{1,d,m}$.