# Introduction to Machine Learning

Fall 2022
University of Science and Technology of China

Lecturer: Jie Wang — Homework 6
Name: Yunqin Zhu — ID: PB20061372

**Notice,** to get the full credits, please present your solutions step by step.

### Exercise 1: SVM for Linearly Separable Cases

Given the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let

$$\mathcal{D}^+ = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = 1\}, \quad \mathcal{D}^- = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = -1\}.$$

Assume that $\mathcal{D}^+$ and $\mathcal{D}^-$ are nonempty and the training set $\mathcal{D}$ is linearly separable. We have shown in Lecture 13 that SVM can be written as

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \ & \min_i y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{1}$$

Moreover, we further transform Problem (1) to

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \ & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n. \end{aligned} \tag{2}$$

We denote the feasible set of Problem (2) by

$$\mathcal{F} = \{(\mathbf{w}, b) : y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, i = 1, \ldots, n\}.$$

1. The Euclidean distance between a linear classifier $f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ and a point $\mathbf{z}$ is

   $$d(\mathbf{z}, f) = \min_{\mathbf{x}} \{\|\mathbf{z} - \mathbf{x}\| : f(\mathbf{x}; \mathbf{w}, b) = 0\}.$$

   Please find the closed form of $d(\mathbf{z}, f)$.

   **Solution:**
   Denote $C = \{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{w}, \mathbf{x} \rangle = 0\}$, which is a subspace. As $\mathbf{w}$ is a normal vector of $C$, projecting an arbitraty $\mathbf{x}$ onto the orthogonal complement of $C$ yields

   $$\Pi_{C^\perp}(\mathbf{x}) = \mathbf{w}(\mathbf{w}^\top \mathbf{w})^{-1}\mathbf{w}^\top \mathbf{x} = \frac{\langle \mathbf{w}, \mathbf{x} \rangle}{\|\mathbf{w}\|^2}\mathbf{w}.$$

   Let $\mathbf{x}_0 \in \mathbb{R}^d$ satisfy $\langle \mathbf{w}, \mathbf{x}_0 \rangle = b$. Then, by the definition of projection, we have

   $$d(\mathbf{z}, f) = \|\mathbf{z} - \Pi_{C-\mathbf{x}_0}(\mathbf{z})\| = \|\Pi_{C^\perp}(\mathbf{z} + \mathbf{x}_0)\| = \frac{|\langle \mathbf{w}, \mathbf{z} \rangle + b|}{\|\mathbf{w}\|}. \qquad \blacksquare$$

2. Show that $\mathcal{F}$ is nonempty.

**Solution:**
Since the training set is linearly separable, there exists $(\hat{\mathbf{w}}, \hat{b})$ such that $y_i = \operatorname{sgn}(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b})$ for all $i$, i.e. $y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) > 0$. Letting $c = \min_i y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b})$, we have $y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) > c$ and hence $y_i(\langle \frac{\hat{\mathbf{w}}}{c}, \mathbf{x}_i \rangle + \frac{\hat{b}}{c}) \geq 1$ for all $i$. Therefore, $(\frac{\hat{\mathbf{w}}}{c}, \frac{\hat{b}}{c}) \in \mathcal{F}$. ■

3. Show that Problem (2) admits an optimal solution.

**Solution:**
Let $\mathcal{G} = \{\mathbf{w} : (\mathbf{w}, b) \in \mathcal{F}\}$, which is the image of $\mathcal{F}$ under an affine transformation. We note that $\mathcal{F}$ is nonempty, closed and convex, so is $\mathcal{G}$. Consider the following problem

$$\min_{\mathbf{w}} \ \frac{1}{2}\|\mathbf{w}\|^2, \ \text{s.t.} \ \mathbf{w} \in \mathcal{G}. \tag{3}$$

It is easy to see that Problem (2) attains its optimal value if and only if Problem (3) does. Since the objective function of Problem (3) is strongly convex, it is solvable and has a unique global minimum. Hence, Problem (2) is also solvable. ■

4. Let $(\mathbf{w}^*, b^*)$ be the optimal solution to Problem (2). Show that $\mathbf{w}^* \neq \mathbf{0}$.

**Solution:**
If $\mathbf{w} = \mathbf{0}$, then the constraint becomes $y_i b \geq 1$ for all $i = 1, \ldots, n$. Because both $\mathcal{D}^+$ and $\mathcal{D}^-$ are nonempty, we have $b \geq 1$ and $-b \geq 1$, which is impossible. That is, $(\mathbf{0}, b)$ cannot be a feasible solution, let alone an optimal solution. ■

5. Show that Problems (1) and (2) are equivalent; that is, they share the same set of optimal solutions.

**Solution:**
Denote the set of optimal solutions to Problems (1) and (2) by $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. According to Homework 5, $(\mathbf{w}^*, b^*) \in \mathcal{S}_2 \cap \textbf{int}\ \mathcal{F}$ only if the gradient of the objective equals $\mathbf{0}$, i.e. $\mathbf{w}^* = \mathbf{0}$, which contradicts the result in Exercise 1.4. So $\mathcal{S}_2 \subset \textbf{bd}\ \mathcal{F} \subset \mathcal{F}$, where $\textbf{bd}\ \mathcal{F}$ is the feasible set of Problem (1). Thus, we can conclude that $\mathcal{S}_2 \subset \mathcal{S}_1$. Since all solutions in $\mathcal{S}_1$ and $\mathcal{S}_2$ share the same optimal value, we have $\mathcal{S}_1 = \mathcal{S}_2$. ■

6. Let $(\mathbf{w}^*, b^*)$ be the optimal solution to Problem (2). Show that there exist at least one positive sample and one negative sample, respectively, such that the corresponding equality holds. In other words, there exist $i, j \in \{1, 2, \ldots, n\}$ such that

$$\begin{aligned} 1 = y_i =& \langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*, \\ -1 = y_j =& \langle \mathbf{w}^*, \mathbf{x}_j \rangle + b^*. \end{aligned}$$

**Solution:**
Denote $\mathcal{F}^+ = \{(\mathbf{w}, b) : \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1, \forall (\mathbf{x}_i, y_i) \in \mathcal{D}^+\}$ and $\mathcal{F}^- = \{(\mathbf{w}, b) : \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1, \forall (\mathbf{x}_i, y_i) \in \mathcal{D}^-\}$. Assume that $(\mathbf{w}^*, b^*) \in \textbf{bd}\ \mathcal{F}^+ \cap \textbf{int}\ \mathcal{F}^-$. Then there exists $\delta > 0$ such that the neighborhood $B = \{(\mathbf{w}, b) : \|(\mathbf{w}, b) - (\mathbf{w}^*, b^*)\| < \delta\} \subset \mathcal{F}^-$. Since $(\mathbf{w}^*, b^*) \in \mathcal{F} \cap B = \mathcal{F}^+ \cap \mathcal{F}^- \cap B = \mathcal{F}^+ \cap B$, it is also the local minimum, and thus the global minimum of the following convex optimization problem

$$\min_{\mathbf{w}, b} \ \frac{1}{2}\|\mathbf{w}\|^2, \ \text{s.t.} \ (\mathbf{w}, b) \in \mathcal{F}^+.$$

However, the above problem attains its optimal value at $\mathbf{w}^* = \mathbf{0}$, which contradicts the result in Exercise 1.4. Therefore, $(\mathbf{w}^*, b^*) \notin \mathbf{bd}\ \mathcal{F}^+ \cap \mathbf{int}\ \mathcal{F}^-$. Similarly, we can show that $(\mathbf{w}^*, b^*) \notin \mathbf{int}\ \mathcal{F}^+ \cap \mathbf{bd}\ \mathcal{F}^-$. So $(\mathbf{w}^*, b^*) \in \mathbf{bd}\ \mathcal{F}^+ \cap \mathbf{bd}\ \mathcal{F}^-$, which implies that there exists at least one positive sample and one negative sample such that the corresponding equality holds. ■

7. Show that the optimal solution to Problem (2) is unique.

**Solution:**
Again, consider Problem (3) in Exercise 1.3, which is strongly convex and admits a unique solution. It follows that the optimal solutions $(\mathbf{w}^*, b^*)$ to Problem (2) share the same $\mathbf{w}^*$. To see that $b^*$ is also unique, we let $\mathbf{x}_i$ be a support vector (we have shown its existence). Then $y_i(\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*) = 1$, which leads to $b^* = y_i - \langle \mathbf{w}^*, \mathbf{x}_i\rangle$ and completes the proof. ■

8. Can we remove the inequalities that hold strictly at the optimum to Problem (2) without affecting the solution? Please justify your claim rigorously.

**Solution:**
Yes, we can. Define $\mathcal{I}_{\text{ac}} = \{i : |\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*| = 1\}$ and $\mathcal{I}_{\text{ia}} = \{i : |\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*| > 1\}$ as the index sets of the support and non-support vectors, respectively, where $(\mathbf{w}^*, b^*)$ is the optimal solution to Problem (2). Moreover, we define $\mathcal{F}_{\text{ac}} = \{(\mathbf{w}, b) : \langle \mathbf{w}, \mathbf{x}_i\rangle + b \geq y_i, \forall i \in \mathcal{I}_{\text{ac}}\}$ and $\mathcal{F}_{\text{ia}} = \{(\mathbf{w}, b) : \langle \mathbf{w}, \mathbf{x}_i\rangle + b \geq y_i, \forall i \in \mathcal{I}_{\text{ia}}\}$. Then $(\mathbf{w}^*, b^*) \in \mathbf{bd}\ \mathcal{F}_{\text{ac}} \cap \mathbf{int}\ \mathcal{F}_{\text{ia}}$. Similar to Exercise 1.6, there exist $\delta > 0$ such that the neighborhood $B = \{(\mathbf{w}, b) : \|(\mathbf{w}, b) - (\mathbf{w}^*, b^*)\| < \delta\} \subset \mathcal{F}_{\text{ia}}$. Since $(\mathbf{w}^*, b^*) \in \mathcal{F} \cap B = \mathcal{F}_{\text{ac}} \cap \mathcal{F}_{\text{ia}} \cap B = \mathcal{F}_{\text{ac}} \cap B$, it is the local minimum, and thus the global minimum of the following convex optimization problem

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2, \ \text{s.t.} \ (\mathbf{w}, b) \in \mathcal{F}_{\text{ac}},$$

which is the desired result. ■

9. Find the dual problem of (2) and the corresponding optimality conditions.

**Solution:**
To find the dual problem, we first construct the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i \left(y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1\right),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is the dual variable. We next find the dual function

$$\begin{aligned} q(\boldsymbol{\alpha}) &= \inf_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \\ &= \inf_{\mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \langle \mathbf{w}, \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i\rangle \right\} + \inf_b \left\{ -b \sum_{i=1}^n \alpha_i y_i \right\} + \sum_{i=1}^n \alpha_i \\ &= -\frac{1}{2}\left\| \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right\|^2 + \sum_{i=1}^n \alpha_i. \end{aligned}$$

To attain the above infimum, we must have

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \mathbf{0} \quad \text{and} \quad \nabla_b L(\mathbf{w}, b, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \alpha_i y_i = 0.$$

That is, **dom** $q = \{\boldsymbol{\alpha} : \sum_{i=1}^{n} \alpha_i y_i = 0\}$. The dual problem is

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \ & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{n} \alpha_i, \\ \text{s.t.} \ & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \ i = 1, \ldots, n. \end{aligned}$$

Since the primal problem is convex quadratic and solvable, the dual problem is also solvable and the duality gap is zero. To sum up, the KKT conditions are

$$\begin{aligned} & \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i, \quad \sum_{i=1}^{n} \alpha_i^* y_i = 0, && \text{(Lagrangian optimality)} \\ & y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) \geq 1, \quad i = 1, \ldots, n, && \text{(primal feasibility)} \\ & \alpha_i^* \geq 0, \quad i = 1, \ldots, n, && \text{(dual feasibility)} \\ & \alpha_i^* \left( y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) - 1 \right) = 0, \quad i = 1, \ldots, n, && \text{(complementary slackness)} \end{aligned}$$

where $\mathbf{w}^*$ and $b^*$ are the primal optimal solution, and $\boldsymbol{\alpha}^*$ is the dual optimal solution. ■

### Exercise 2: Discussions on Geometric Multiplier and Duality Gap

Consider the primal problem

$$\begin{aligned}\min_{\mathbf{x}}\ & f(\mathbf{x}) \\ \text{s.t. } & g_i(\mathbf{x}) \leq 0,\ i = 1, \cdots, m, \\ & h_i(\mathbf{x}) = 0, i = 1, \cdots, p, \\ & \mathbf{x} \in X.\end{aligned} \tag{4}$$

Let

$$S = \{(\mathbf{g}(\mathbf{x}), \mathbf{h}(\mathbf{x}), f(\mathbf{x})) : \mathbf{x} \in X\} \subset \mathbb{R}^{m+p+1}. \tag{5}$$

Are the following claims on the geometric multiplier and the duality gap for the primal problem correct? Justify the claims rigorously if they are correct. Otherwise please give a counterexample for each.

1. The geometric multiplier for the primal problem (4) always exists.

   **Solution:**
   False. Consider the following primal problem

   $$\begin{aligned}\min_{x}\ & f(x) = x \\ \text{s.t. } & g(x) = x^2 \leq 0.\end{aligned}$$

   The dual function is

   $$q(\lambda) = \inf_x \left\{x + \lambda x^2\right\} = -\frac{1}{4\lambda},$$

   whose supremum $q^* = 0$ cannot be attained over $\lambda \geq 0$. Therefore, the geometric multiplier does not exist. ■

2. If the geometric multiplier exists, then it is unique.

   **Solution:**
   False. Consider the following primal problem

   $$\begin{aligned}\min_{x}\ & f(x) = x^2 \\ \text{s.t. } & g(x) = x^2 \leq 0.\end{aligned}$$

   The dual function is

   $$q(\lambda) = \inf_x \left\{x^2 + \lambda x^2\right\} = 0 = q^*.$$

   That is, any $\lambda \geq 0$ is a geometric multiplier. ■

3. If the geometric multiplier exists, then the duality gap is zero.

**Solution:**
True. The geometric multiplier is defined as some $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ such that $\boldsymbol{\lambda}^* \geq \mathbf{0}$ and

$$f^* = \inf_{x \in X} L(x, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = q(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$$

By the weak duality theorem, we have $q^* \leq f^*$. On the other hand, $q^* \geq q(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = f^*$. So $q^* = f^*$, i.e. the duality gap is zero. ■

4. If the duality gap is zero, there exists at least one geometric multiplier.

   **Solution:**
   False. Consider the primal and dual problems in Exercise 2.1. We have $f^* = q^* = 0$ but the geometric multiplier does not exist. ■

5. Let $(\lambda^*, \mu^*)$ be a geometric multiplier. Then, the problem $\mathbf{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \mu^*)$ always admits at least one solution, where $L(\mathbf{x}, \lambda, \mu)$ is the Lagrangian for (4).

   **Solution:**
   False. Consider the following primal problem

$$\begin{aligned} \min_x \ & f(x) = e^x \\ \text{s.t. } & g(x) = e^x - 1 \leq 0. \end{aligned}$$

   The Lagrangian is $L(x, \lambda) = e^x + \lambda e^x - \lambda$, whose infimum with respect to $x$ is $-\lambda$ but cannot be attained. However, the dual problem

$$\begin{aligned} \max_\lambda \ & -\lambda = 0 \\ \text{s.t. } & \lambda \geq 0 \end{aligned}$$

   admits a unique optimal solution $\lambda^* = 0$, which is a geometric multiplier. ■

6. If $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a geometric multiplier and the problem $\mathbf{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ admits a solution $\mathbf{x}^*$, then $\mathbf{x}^*$ is feasible.

   **Solution:**
   False. Consider the following primal problem

$$\begin{aligned} \min_x \ & f(x) = \begin{cases} -x, & x < 0, \\ 0, & x \geq 0 \end{cases} \\ \text{s.t. } & g(x) = x \leq 0. \end{aligned}$$

   The Lagrangian is

$$L(x, \lambda) = \begin{cases} (\lambda - 1)x, & x < 0, \\ \lambda x, & x \geq 0. \end{cases}$$

   $\inf_x L(x, \lambda)$ is finite if and only if $0 \leq \lambda \leq 1$. We see that $q^* = f^* = 0$ and any $\lambda^* \in [0, 1]$ is a geometric multiplier. Take $\lambda^* = 0$ for example. In this case, we have $x^* \in [0, \infty)$. However, only $x^* = 0$ is feasible for the primal problem. ■

7. Let $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be a geometric multiplier. Then, $\mathbf{x}^*$ is a global minimum of the primal problem if and only if $\mathbf{x}^*$ is feasible and $\mathbf{x}^* \in \mathop{\mathbf{argmin}}_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.

   **Solution:**
   False. Consider the primal and dual problem in Exercise 2.6. Letting $\lambda^* = 1$, we have $x^* \in (-\infty, 0]$, which is feasible. However, only $x^* = 0$ is optimal for the primal problem. ■

8. $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a geometric multiplier if and only if $\boldsymbol{\lambda}^* \geq \mathbf{0}$ and among all hyperplanes with normal $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, 1)$ that contain the set $S$ in their positive halfspace, the highest attained level of interception of the vertical axis is $f^*$, where

$$f^* = \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = 0, \mathbf{x} \in X\}.$$

   **Solution:**
   True. We only need to show that among all hyperplanes with normal $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, 1)$ that contain the set $S$ in their positive halfspace, the highest attained level of interception of the vertical axis is $\inf_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$. In fact, by definition, any hyperplane with normal $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, 1)$ can be written as $L = f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, \mathbf{g}(\mathbf{x})\rangle + \langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x})\rangle$. The hyperplane contains $S$ in its positive halfspace if and only if $L \leq \inf_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$. Setting $\mathbf{g}(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ as zero, we obtain the interception $f(\mathbf{x}) = L$, whose maximum with respect to $\mathbf{x}$ is $\inf_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$. ■

**Exercise 3: The Dual Problem of SVM**

Suppose that the training set is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let

$$\mathcal{D}^+ = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = 1\}, \quad \mathcal{D}^- = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = -1\}.$$

Assume that $\mathcal{D}^+$ and $\mathcal{D}^-$ are nonempty. The soft margin SVM takes the form of

$$\begin{aligned} &\min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ &\qquad \xi_i \geq 0, \; i = 1, \ldots, n, \end{aligned} \tag{6}$$

The corresponding dual problem is

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle - \sum_{i=1}^n \alpha_i \\ &\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0, \\ &\qquad \alpha_i \in [0, C], i = 1, \ldots, n. \end{aligned} \tag{7}$$

1. Show that the problems (6) and (7) always admit optimal solutions.

   **Solution:**
   The primal problem has convex quadratic objective with linear constraints. Moreover, it is bounded from below. By Proposition 5 in Lecture 13, the primal and dual problems have optimal solutions and the duality gap is zero. ■

2. Let $(\mathbf{w}^*, b^*)$ be the solution to (6) and $\alpha^*$ be the corresponding solution to (7).

   (a) When is $\alpha_i^*$ equal to $C$, $i = 1, \ldots, n$? Please give an example and find the corresponding solutions.

   (b) When is $\mathbf{w}^*$ equal to zero? Please give an example and find the corresponding solutions.

   Notice that, you need to find all the primal and dual optimal solutions if they are not unique.

   **Solution:**

   (a) In addition to the feasibility of the primal and dual problems, we have the complementary slackness

   $$\alpha_i^*(y_i(\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*) - 1 + \xi_i^*) = (C - \alpha_i^*)\xi_i^* = 0, \; i = 1, \ldots, n,$$

   and the Lagrangian optimality

   $$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i, \quad \sum_{i=1}^n \alpha_i^* y_i = 0.$$

When $\alpha_i^*$ equals $C$ for all $i = 1, \dots, n$, the complementary slackness implies that $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) = 1 - \xi_i^*$, i.e. each $\mathbf{x}_i$ is either a support vector or wrongly classified. Moreover, the Lagrangian optimality implies that $\sum_{i=1}^n y_i = 0$, i.e. $\mathcal{D}^+$ and $\mathcal{D}^-$ have the same number of samples, and

$$\mathbf{w}^* = C \left( \sum_{y_i=1} \mathbf{x}_i - \sum_{y_j=-1} \mathbf{x}_j \right). \tag{8}$$

Conversely, if $\mathcal{D}^+$ and $\mathcal{D}^-$ have the same number of samples, and there exists $b^*$ such that

$$y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) \le 1, \ \forall\, i = 1, \dots, n, \tag{9}$$

where $\mathbf{w}^*$ is defined by (8), then $\alpha_i^* = C$ for all $i = 1, \dots, n$.

For example, let $C = \frac{1}{2}$ and $\mathcal{D} = \{(1,1), (-1,-1)\}$. By (8), we have $w^* = 1$. Plugging it into (9) yields $b^* = 0$ and $\xi^* = \mathbf{0}$. In this case, the dual problem is

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}(\alpha_1 + \alpha_2)^2 - \alpha_1 - \alpha_2 \\ \text{s.t.} \ & \alpha_1 - \alpha_2 = 0, \\ & 0 \le \alpha_1, \alpha_2 \le 1. \end{aligned}$$

It is easy to see that the dual optimal solution is $\alpha_1^* = \alpha_2^* = \frac{1}{2}$, as desired.

(b) If $\mathbf{w}^* = \mathbf{0}$ is optimal, then the primal feasibility reduces to

$$\xi_i^* = \max\{0, 1 - y_i b^*\}, \ i = 1, \dots, n.$$

Then, the primal problem becomes

$$\min_b \ \left\{ |\mathcal{D}^+| \max\{0, 1-b\} + |\mathcal{D}^-| \max\{0, 1+b\} \right\}.$$

We can see that the optimal value is $2 \min \{|\mathcal{D}^+|, |\mathcal{D}^-|\}$, and the optimal $b^*$ satisfies

- If $|\mathcal{D}^+| > |\mathcal{D}^-|$, then $b^* = 1$.
- If $|\mathcal{D}^+| < |\mathcal{D}^-|$, then $b^* = -1$.
- If $|\mathcal{D}^+| = |\mathcal{D}^-|$, then $b^* \in [-1, 1]$.

The complementary slackness implies that

$$\alpha_i^*(y_i b^* - 1 + \xi_i^*) = (C - \alpha_i^*)\xi_i^* = 0, \ i = 1, \dots, n,$$

and hence

- If $|\mathcal{D}^+| > |\mathcal{D}^-|$, then $\alpha_i^* = C$ if $y_i = -1$.
- If $|\mathcal{D}^+| < |\mathcal{D}^-|$, then $\alpha_i^* = C$ if $y_i = 1$.
- If $|\mathcal{D}^+| = |\mathcal{D}^-|$, then $\alpha_i^* = C$ for all $i = 1, \dots, n$, where we use $\sum_{i=1}^n \alpha_i^* y_i = 0$

We then discuss the KKT conditions.

- If $|\mathcal{D}^+| > |\mathcal{D}^-|$, $\mathbf{w}^* = \mathbf{0}$ is optimal if and only if there exists $\alpha_i^* \in [0, C]$ for each $y_i = 1$ such that $\sum_{y_i=1} \alpha_i^* = C|\mathcal{D}^-|$ and $\sum_{y_i=1} \alpha_i^* \mathbf{x}_i = C \sum_{y_j=-1} \mathbf{x}_j$.
- If $|\mathcal{D}^+| < |\mathcal{D}^-|$, $\mathbf{w}^* = \mathbf{0}$ is optimal if and only if there exists $\alpha_i^* \in [0, C]$ for each $y_i = -1$ such that $\sum_{y_i=-1} \alpha_i^* = C|\mathcal{D}^+|$ and $\sum_{y_i=-1} \alpha_i^* \mathbf{x}_i = C \sum_{y_j=1} \mathbf{x}_j$.
- If $|\mathcal{D}^+| = |\mathcal{D}^-|$, $\mathbf{w}^* = \mathbf{0}$ is optimal if and only if $\sum_{y_i=1} \mathbf{x}_i = \sum_{y_j=-1} \mathbf{x}_j$.

A trivial case is that $\mathcal{D}^+ = \emptyset$ or $\mathcal{D}^- = \emptyset$. If so, we also have $\boldsymbol{\alpha}^* = \mathbf{0}$, $\boldsymbol{\xi}^* = \mathbf{0}$. However, this case contradicts the assumption in the problem statement.

For another example, let $C = 1$, $\mathcal{D}^- = \{(-2, -1), (2, -1)\}$ and $\mathcal{D}^+ = \{(-1, 1), (1, 1)\}$. Clearly, the aforementioned conditions can be satisfied, and hence $\mathbf{w}^* = \mathbf{0}$ is optimal. To verify this, consider the dual problem

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\boldsymbol{\alpha}^\top \begin{pmatrix} 4 & -4 & -2 & 2 \\ -4 & 4 & 2 & -2 \\ -2 & 2 & 1 & -1 \\ 2 & -2 & -1 & 1 \end{pmatrix} \boldsymbol{\alpha} + \mathbf{1}^\top \boldsymbol{\alpha} \\ \text{s.t.} \quad & -\alpha_1 - \alpha_2 + \alpha_3 + \alpha_4 = 0, \\ & 0 \le \alpha_1, \alpha_2, \alpha_3, \alpha_4 \le C. \end{aligned}$$

Solving this problem, we obtain $\alpha_1^* = \alpha_2^* = \alpha_3^* = \alpha_4^* = 1$ and $p^* = 4$, which is consistent with the established result. Futhermore, the optimal solution to the primal problem is $\mathbf{w}^* = \mathbf{0}$, $b^* \in [-1, 1]$ and $\xi_i^* = 1 - y_i b^*$ for all $i = 1, \ldots, n$. ■

**Exercise 4: An Example of the Soft Margin SVM**

Recall that the soft margin SVM takes the form of

$$\begin{aligned}&\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ &\quad \xi_i \geq 0,\ i = 1, \ldots, n,\end{aligned} \tag{10}$$

where $C > 0$.

1. The function of the slack variables used in the optimization problem for soft margin hyperplanes takes the form $\sum_{i=1}^n \xi_i$. We could also use $\sum_{i=1}^n \xi_i^p$, where $p > 1$. The soft margin SVM becomes

$$\begin{aligned}&\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i^p, \\ &\text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, i = 1, \ldots, n, \\ &\quad \xi_i \geq 0,\ i = 1, \ldots, n.\end{aligned} \tag{11}$$

Please find the dual problem of (11) and the corresponding optimal conditions.

**Solution:**
To find the dual problem, we first construct the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i^p - \sum_{i=1}^n \alpha_i\left(y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i\right) - \sum_{i=1}^n \mu_i\xi_i,$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ are the dual variables. We next find the dual function

$$\begin{aligned}q(\boldsymbol{\alpha}, \boldsymbol{\mu}) &= \inf_{\mathbf{w},b,\boldsymbol{\xi}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \\ &= \inf_{\mathbf{w}}\left\{\frac{1}{2}\|\mathbf{w}\|^2 - \langle \mathbf{w}, \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i\rangle\right\} + \inf_b\left\{-b\sum_{i=1}^n \alpha_i y_i\right\} \\ &\quad + \sum_{i=1}^n \inf_{\xi_i}\{C\xi_i^p - (\alpha_i + \mu_i)\xi_i\} + \sum_{i=1}^n \alpha_i\end{aligned}$$

To attain the above infimum, we must have

$$\begin{aligned}&\nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = \mathbf{0}, \\ &\nabla_b L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = -\sum_{i=1}^n \alpha_i y_i = 0, \\ &\nabla_{\xi_i} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = Cp\xi_i^{p-1} - (\alpha_i + \mu_i) = 0,\ i = 1, \ldots, n.\end{aligned}$$

Then

$$q(\boldsymbol{\alpha}, \boldsymbol{\mu}) = -\frac{1}{2}\left\|\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i\right\|^2 + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n \frac{(1-p)(\alpha_i+\mu_i)}{p}\left(\frac{\alpha_i+\mu_i}{Cp}\right)^{\frac{1}{p-1}},$$

where $\mathbf{dom}\, q = \{\boldsymbol{\alpha} : \sum_{i=1}^n \alpha_i y_i = 0\}$. Note that we can always set $\mu_i = 0$ to achieve a smaller $\xi_i$ and thus a smaller objective. So $\boldsymbol{\mu}^* = \mathbf{0}$. The dual problem reduces to

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \ & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle - \frac{p-1}{p}\sum_{i=1}^n \alpha_i \left(\frac{\alpha_i}{Cp}\right)^{\frac{1}{p-1}}, \\ \text{s.t.} \ & \sum_{i=1}^n \alpha_i y_i = 0, \quad \alpha_i \geq 0,\ i = 1, \ldots, n. \end{aligned}$$

By Proposition 4 in Lecture 13, the duality gap is zero. The KKT conditions are

$$\begin{aligned} & \mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i, \quad \sum_{i=1}^n \alpha_i^* y_i = 0, \quad \xi_i^* = \left(\frac{\alpha_i}{Cp}\right)^{\frac{1}{p-1}} && \text{(Lagrangian optimality)} \\ & y_i(\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*) \geq 1 - \xi_i^*, \quad i = 1, \ldots, n, && \text{(primal feasibility)} \\ & \alpha_i^* \geq 0, \quad i = 1, \ldots, n, && \text{(dual feasibility)} \\ & \alpha_i^* \left(y_i(\langle \mathbf{w}^*, \mathbf{x}_i\rangle + b^*) - 1 + \xi_i\right) = 0, \quad i = 1, \ldots, n, && \text{(complementary slackness)} \end{aligned}$$

where $\mathbf{w}^*, b^*, \boldsymbol{\xi}^*$ are the primal optimal solution, and $\boldsymbol{\alpha}^*$ are the dual optimal solution. ■

As shown in Figure 1, the training set is $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{11}$, where $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i \in \{+1, -1\}$. Suppose that we use the soft margin SVM to classify the data points and get the optimal parameters $\mathbf{w}^*$, $b^*$, and $\boldsymbol{\xi}^*$ by solving the problem Eq. (11).

2. Please write down the equations of the separating hyperplane ($H_0$) and the marginal hyperplanes ($H_1$ and $H_2$) in terms of $\mathbf{w}^*$ and $b^*$.

   **Solution:**
   $H_0 = \{\mathbf{x} : \langle \mathbf{w}^*, \mathbf{x}\rangle + b^* = 0\}$. $H_1 = \{\mathbf{x} : \langle \mathbf{w}^*, \mathbf{x}\rangle + b^* = 1\}$. $H_2 = \{\mathbf{x} : \langle \mathbf{w}^*, \mathbf{x}\rangle + b^* = -1\}$. ■

3. Please find the support vectors and the non-support vectors.

   **Solution:**
   The support vectors are $\mathbf{x}_5$, $\mathbf{x}_7$, $\mathbf{x}_{10}$, which are on the marginal hyperplanes. All the other data points are non-support vectors. ■

4. Please find the values (or ranges) of the optimal slack variables $\xi_i^*$ for $i = 1, 2, \ldots, 11$. (*Hint: The possible answers are* $\xi_i^* = 0$, $0 < \xi_i^* < 1$, $\xi_i^* = 1$, *and* $\xi_i^* > 1$). How do the slack variables change when the parameter $C$ increases and decreases?

   **Solution:**
   $\xi_1, \xi_2, \xi_3, \xi_5, \xi_7, \xi_8, \xi_9, \xi_{10}, \xi_{11} = 0$. $0 < \xi_6 < 1$. $\xi_4, \xi_8 > 1$. The slack variables decrease when $C$ increases and increase when $C$ decreases. ■

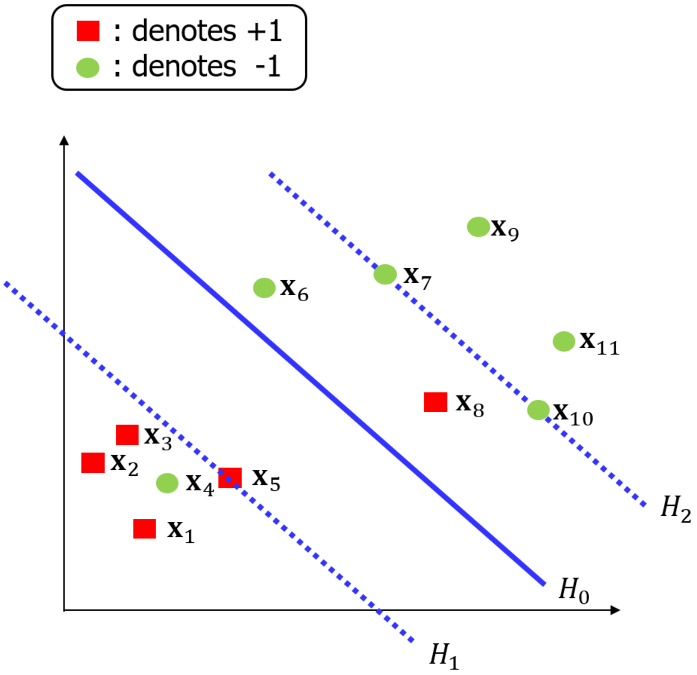


Figure 1: Classifying the data points using the soft margin SVM. $H_0$ is the separating hyperplane. $H_1$ and $H_2$ are the marginal hyperplanes.

**Exercise 5: Neural Networks**

1. The softmax function $\mathbf{f}:\mathbb{R}^n\to\mathbb{R}^n$ is defined by:

$$f_i(\mathbf{x})=\frac{\exp(x_i)}{\sum_{k=1}^n\exp(x_k)}, i=1,\dots,n,$$

where $x_i$ is the $i^{th}$ component of $\mathbf{x}\in\mathbb{R}^n$. The function $\mathbf{f}(\mathbf{x})=(f_1(\mathbf{x}),f_2(\mathbf{x}),\dots,f_n(\mathbf{x}))^\top$ converts each input $\mathbf{x}$ into a probability (stochastic) vector in which all entries are nonnegative and add up to one.

(a) Please find the gradient and Jacobian matrix of $\mathbf{f}(\mathbf{x})$, i.e., $\nabla\mathbf{f}(\mathbf{x})$ and $\mathbf{J}_\mathbf{f}(\mathbf{x})$.

(b) Show that $\mathbf{f}(\mathbf{x})=\mathbf{f}(\mathbf{x}-c\mathbf{1})$, where $c=\max\{x_1,x_2,...,x_n\}$ and $\mathbf{1}$ is a vector all of whose components are one. When do we need this transformation?

**Solution:**

(a) Note that the partial derivatives of $\mathbf{f}(\mathbf{x})$ are:

$$\frac{\partial f_i}{\partial x_j}=f_i(\delta_{ij}-f_j)=\begin{cases}f_i(1-f_i), & i=j\\ -f_if_j, & i\neq j\end{cases}.$$

So the gradient of $\mathbf{f}(\mathbf{x})$ is:

$$\nabla\mathbf{f}(\mathbf{x})=\begin{pmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n}\end{pmatrix}=\begin{pmatrix}f_1(1-f_1) & -f_1f_2 & \cdots & -f_1f_n\\ -f_2f_1 & f_2(1-f_2) & \cdots & -f_2f_n\\ \vdots & \vdots & \ddots & \vdots\\ -f_nf_1 & -f_nf_2 & \cdots & f_n(1-f_n)\end{pmatrix}.$$

The Jacobian, as the transpose of the gradient, is the same

$$\mathbf{J}_\mathbf{f}(\mathbf{x})=\nabla\mathbf{f}(\mathbf{x})^\top=\begin{pmatrix}f_1(1-f_1) & -f_2f_1 & \cdots & -f_nf_1\\ -f_1f_2 & f_2(1-f_2) & \cdots & -f_nf_2\\ \vdots & \vdots & \ddots & \vdots\\ -f_1f_n & -f_2f_n & \cdots & f_n(1-f_n)\end{pmatrix}.$$

(b) For any $c\in\mathbb{R}$, we have

$$f_i(\mathbf{x})=\frac{\exp(x_i)}{\sum_{k=1}^n\exp(x_k)}=\frac{\exp(x_i-c)}{\sum_{k=1}^n\exp(x_k-c)}=f_i(\mathbf{x}-c\mathbf{1}).$$

Thus, $\mathbf{f}(\mathbf{x})=\mathbf{f}(\mathbf{x}-c\mathbf{1})$. When $\max\{x_1,x_2,...,x_n\}$ is large, we need this transformation to avoid overflow in the numerical computation. ■

2. Consider the neural network with a single hidden layer in Figure 2. Let $\mathbf{x}\in\mathbb{R}^3$ be an input vector, and $\mathbf{y}$ be its corresponding output of the network. $f$ implies that there exist four units in the hidden layer, each of which is followed by a sigmoid activation function $\sigma$, converting its input $\mathbf{z}$ to output $\mathbf{a}$. Suppose that the ground truth label vector of $\mathbf{x}$ is $[0,0,1]^\top$ and we use the cross entropy introduced in Lecture 15 as the loss function.

(a) Please find the update formula for the $j^{\text{th}}$ weight of the $i^{\text{th}}$ hidden unit , i.e., $w^1_{ij}$ where $i \in \{1, 2, 3, 4\}$ and $j \in \{1, 2, 3\}$.

(b) Can we initialize all the parameters, i.e., weights and bias, of the neural network to zero? Please state you conclusion.

**Solution:**

(a) For simplicity, we do not consider the bias. Then, we have

$$\begin{cases} z_i = \sum_{j=1}^{3} w^1_{ij} x_j, \\ a_i = \sigma(z_i), \\ y_i = \sum_{j=1}^{4} w^2_{ij} a_j, \\ \mathbf{p} = \text{softmax}\,(\mathbf{y}), \\ \text{Loss} = -\log p_3. \end{cases}$$

The partial derivative of the loss with respect to $y_i, w^2_{ij}, z_i, w^1_{ij}$ are

$$\begin{aligned} \frac{\partial \text{Loss}}{\partial y_i} &= \frac{\partial \text{Loss}}{\partial p_3}\frac{\partial p_3}{\partial y_i} = -\frac{1}{p_3} p_3(\delta_{3i} - p_3) = p_3 - \delta_{3i}, \\ \frac{\partial \text{Loss}}{\partial w^2_{ij}} &= \frac{\partial \text{Loss}}{\partial y_i}\frac{\partial y_i}{\partial w^2_{ij}} = (p_3 - \delta_{3i}) a_j, \\ \frac{\partial \text{Loss}}{\partial z_i} &= \frac{\partial a_i}{\partial z_i}\sum_{j=1}^{3}\frac{\partial \text{Loss}}{\partial y_j}\frac{\partial y_j}{\partial a_i} = a_i(1 - a_i)\sum_{j=1}^{3}(p_3 - \delta_{3j}) w^2_{ij}, \\ \frac{\partial \text{Loss}}{\partial w^1_{ij}} &= \frac{\partial \text{Loss}}{\partial z_i}\frac{\partial z_i}{\partial w^1_{ij}} = \left\{ a_i(1 - a_i)\sum_{k=1}^{3}(p_3 - \delta_{3k}) w^2_{ik} \right\} x_j. \end{aligned}$$

Suppose that the learning rate is $\eta$. Then the update formula for the $j^{\text{th}}$ weight of the $i^{\text{th}}$ hidden unit is

$$w^1_{ij} \leftarrow w^1_{ij} - \eta \left\{ a_i(1 - a_i)\sum_{k=1}^{3}(p_3 - \delta_{3k}) w^2_{ik} \right\} x_j.$$

(b) Yes, we can, but it is not recommended. When all the parameters are initialized to zero, the inital forward propagation ends with $\mathbf{z} = \mathbf{y} = \mathbf{0}$ and $\mathbf{a} = \mathbf{p} = (1/3, 1/3, 1/3)$ and the backward propagation ends with $w^1_{1j} = w^1_{2j} = w^1_{3j} = w^1_{4j}$ for all $j = 1, 2, 3$ and $w^2_{i1} = w^2_{i2} = w^2_{i3} = w^2_{i4}$ for all $i = 1, 2, 3$, implying that all the hidden nodes are symmetric. As a result, in the following iterations, the parameters of four hidden nodes will be updated in the same way, which is not what we want. ■

3. Consider a convolutional neural network as shown in Table 1.

(a) The convolutional layer parameters are denoted as "conv⟨filter size⟩-⟨number of filters⟩".

(b) The fully connected layer parameters are denoted as "FC⟨number of neurons⟩".

(c) The window size of pooling layers is 2.

(d) The stride of convolutinal layers is 1.

(e) The stride of pooling layers is 2.

(f) You may want to use padding in both convolutional and pooling layers if necessary.

(g) For convenience, we assume that there is no activation function and bias.

Suppose that the input is a **210** $\times$ **160 RGB** image. Please derive the size of all feature maps and the number of parameters.

**Solution:**
Denote the layers shown in Table 1 as C1, C2, S3, C4, C5, S6, F7, F8, respectively.

The size of the input is $3 \times 210 \times 160$.
The size of the output of C1 (conv3-32) is $32 \times (210 - 2) \times (160 - 2) = 32 \times 208 \times 158$.
The size of the output of C2 (conv5-32) is $32 \times (208 - 4) \times (158 - 4) = 32 \times 204 \times 154$.
The size of the output of S3 (max pool) is $32 \times 204/2 \times 154/2 = 32 \times 102 \times 77$.
The size of the output of C4 (conv3-64) is $64 \times (102 - 2) \times (77 - 2) = 64 \times 100 \times 75$.
The size of the output of C5 (conv5-64) is $64 \times (100 - 4) \times (75 - 4) = 64 \times 96 \times 71$.
The size of the output of S6 (max pool) is $64 \times 96/2 \times \lceil 71/2 \rceil = 64 \times 48 \times 36$.
The size of the output of F7 (FC128) and F8 (FC10) are 128 and 10, respectively.

The number of parameters in C1 is $32 \times 3 \times 3 \times 3 = 864$.
The number of parameters in C2 is $32 \times 32 \times 5 \times 5 = 25600$.
The number of parameters in C4 is $64 \times 32 \times 3 \times 3 = 18432$.
The number of parameters in C5 is $64 \times 64 \times 5 \times 5 = 102400$.
The number of parameters in F7 is $64 \times 48 \times 36 \times 128 = 14155776$.
The number of parameters in F8 is $128 \times 10 = 1280$.
The total number of parameters is 14304352. ■

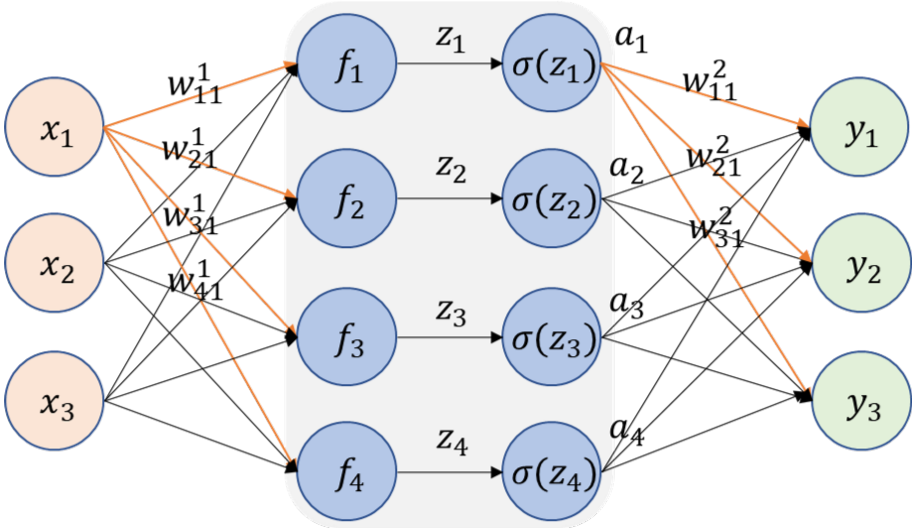


Figure 2: A neural network with a single hidden layer.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| conv3-32 | conv5-32 | max pool | conv3-64 | conv5-64 | max pool | FC-128 | FC-10 |

Table 1: The architecture of convolutional neural network

**Exercise 6: Exercises of Dual Problems (Optional)**

1. Consider the optimization problem

$$\begin{aligned} \min_x \, & x^2+1 \\ \text{s.t.}\, & (x-2)(x-4) \le 0, \\ & x \in \mathbb{R}. \end{aligned}$$

   (a) Give the feasible set, the optimal value, and the optimal solution.

   (b) Plot the objective $x^2+1$ versus $x$. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x,\lambda)$ versus $x$ for a few positive values of $\lambda$. Verify the lower bound property, i.e., $p^* \ge \inf_x L(x,\lambda)$ for $\lambda \ge 0$, where $p^*$ is the optimal value. Derive and sketch the Lagrange dual function.

   (c) State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^*$. Does strong duality hold?

   (d) (**Sensitivity analysis**) Let $p^*(u)$ denote the optimal value of the problem

$$\begin{aligned} \min_x \, & x^2+1 \\ \text{s.t.}\, & (x-2)(x-4) \le u, \\ & x \in \mathbb{R}, \end{aligned}$$

   as a function of the parameter $u$. Please plot $p^*(u)$ and verify that $\mathrm{d}p^*(0)/\mathrm{d}u = -\lambda^*$.

**Solution:**

(a) The feasible set is $[2,4]$. The optimal value is 5. The optimal solution is 2.

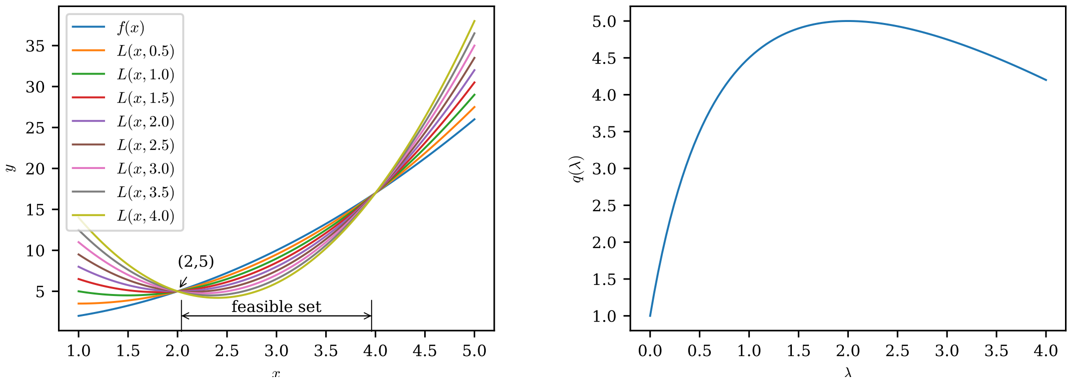
(b) The Lagrangian is

$$L(x,\lambda) = x^2+1+\lambda(x-2)(x-4).$$

The plot of the Lagrangian is shown in Figure 3a. It is clear that $p^* \ge \inf_x L(x,\lambda)$. The dual function is

$$\begin{aligned} q(\lambda) &= \inf_x L(x,\lambda) = \inf_x \{x^2+1+\lambda(x-2)(x-4)\} \\ &= \inf_x \left\{ (\lambda+1)\left(x-\frac{3\lambda}{\lambda+1}\right)^2 + \frac{-\lambda^2+9\lambda+1}{\lambda+1} \right\} \\ &= \frac{-\lambda^2+9\lambda+1}{\lambda+1}. \end{aligned}$$

Its plot is shown in Figure 3b.



(a) The Lagrangian (b) The dual function

Figure 3: The Lagrangian and the dual function

(c) The dual problem is

$$\max_{\lambda} \frac{-\lambda^2 + 9\lambda + 1}{\lambda + 1},$$
$$\text{s.t. } \lambda \geq 0.$$

It is a concave maximization problem. The dual optimal value is 5 and the dual optimal solution is $\lambda = 2$. The strong duality holds.

(d) The optimal solution is $x^* = 3 - \sqrt{u+1}$ and the optimal value is $p^*(u) = u + 11 - 6\sqrt{u+1}$. The plot of $p^*(u)$ is shown in Figure 4.

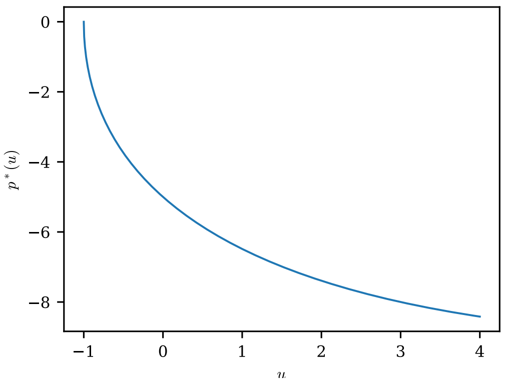


Figure 4: The sensitivity analysis

The derivative of $p^*(u)$ is

$$\frac{\mathrm{d}p^*(u)}{\mathrm{d}u} = 1 - \frac{3}{\sqrt{u+1}}.$$

Hence $\mathrm{d}p^*(0)/\mathrm{d}u = -2 = -\lambda^*$. ■

2. Please use the duality to show that in three-dimensional space, the (minimum) distance from the origin to a line is equal to the maximum over all (minimum) distances of the origin from planes that contain the line.

**Solution:**
Suppose that $\mathbf{x} \in \mathbb{R}^3$. Let the line be defined by the equations $\langle \mathbf{w}_1, \mathbf{x} \rangle + b_1 = 0$ and $\langle \mathbf{w}_2, \mathbf{x} \rangle + b_2 = 0$, where $\mathbf{w}_1 \perp \mathbf{w}_2$ and $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$. Then, the optimization problem is

$$\begin{aligned} &\min_{\mathbf{x}} \|\mathbf{x}\|^2, \\ &\text{s.t. } \langle \mathbf{w}_1, \mathbf{x} \rangle + b_1 = 0, \\ &\quad\ \ \langle \mathbf{w}_2, \mathbf{x} \rangle + b_2 = 0. \end{aligned}$$

The Lagrangian is

$$L(\mathbf{x}, \lambda_1, \lambda_2) = \|\mathbf{x}\|^2 + \lambda_1 \left( \langle \mathbf{w}_1, \mathbf{x} \rangle + b_1 \right) + \lambda_2 \left( \langle \mathbf{w}_2, \mathbf{x} \rangle + b_2 \right).$$

The dual function is

$$\begin{aligned} q(\lambda_1, \lambda_2) &= \inf_{\mathbf{x}} L(\mathbf{x}, \lambda_1, \lambda_2) = \inf_{\mathbf{x}} \left\{ \|\mathbf{x}\|^2 + \lambda_1 \left( \langle \mathbf{w}_1, \mathbf{x} \rangle + b_1 \right) + \lambda_2 \left( \langle \mathbf{w}_2, \mathbf{x} \rangle + b_2 \right) \right\} \\ &= - \left\| \frac{\lambda_1 \mathbf{w}_1 + \lambda_2 \mathbf{w}_2}{2} \right\|^2 + \lambda_1 b_1 + \lambda_2 b_2 \\ &= -\frac{1}{4} \left\{ (\lambda_1 - 2b_1)^2 + (\lambda_2 - 2b_2)^2 \right\} + b_1^2 + b_2^2 \le b_1^2 + b_2^2. \end{aligned}$$

The equality holds if and only if $\lambda_1 = 2b_1$ and $\lambda_2 = 2b_2$. Since the primal problem is convex quadratic and the constraints are linear, there is no duality gap, and thus the optimal solution to the primal problem is also $b_1^2 + b_2^2$. That is, the minimum distance from the origin to the line is $\sqrt{b_1^2 + b_2^2}$.

On the other hand, any plane that contains the line can be written as $\langle \mu_1 \mathbf{w}_1 + \mu_2 \mathbf{w}_2, \mathbf{x} \rangle + \mu_1 b_1 + \mu_2 b_2 = 0$. By Exercise 1.1, the distance from the origin to the plane is $\frac{|\mu_1 b_1 + \mu_2 b_2|}{\sqrt{\mu_1^2 + \mu_2^2}}$. The optimization problem can be formulated as

$$\begin{aligned} &\max_{\mu_1, \mu_2} \ (\mu_1 b_1 + \mu_2 b_2)^2, \\ &\text{s.t. } \mu_1^2 + \mu_2^2 \le 1. \end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$(\mu_1 b_1 + \mu_2 b_2)^2 \le \left( \mu_1^2 + \mu_2^2 \right) \left( b_1^2 + b_2^2 \right) \le b_1^2 + b_2^2.$$

where the equality holds if and only if $\mu_1 = \frac{b_1}{\sqrt{b_1^2 + b_2^2}}$ and $\mu_2 = \frac{b_2}{\sqrt{b_1^2 + b_2^2}}$. That is, the maximum over all distances of the origin from planes that contain the line is $\sqrt{b_1^2 + b_2^2}$, equal to the minimum distance of the origin from the line. ■

### Exercise 7: Some Network Layers, Linear Transformation and Gradient (Optional)

In this exercise, we explore several kinds of network layers in the view of linear transformation.

1. **1-dimensional convolutional layer.** Suppose we have an input $\mathbf{x} \in \mathbb{R}^n$ and filter $\mathbf{w} \in \mathbb{R}^k$ $(n > k)$. We can compute the convolution of $\mathbf{x} * \mathbf{w}$ as follows:

   - Take the convolutional filter $\mathbf{w}$ and align it with the beginning of $\mathbf{x}$. Take the dot product of $\mathbf{w}$ and the $\mathbf{x}[0 : k-1]$ and assign that as the first entry of the output.
   - Suppose we have stride $s$. Shift the filter down by $s$ indices, and now take the dot product of $\mathbf{w}$ and $\mathbf{x}[s : k-1+s]$ and assign to the next entry of your output.
   - Repeat until we run out of entries in $\mathbf{x}$.

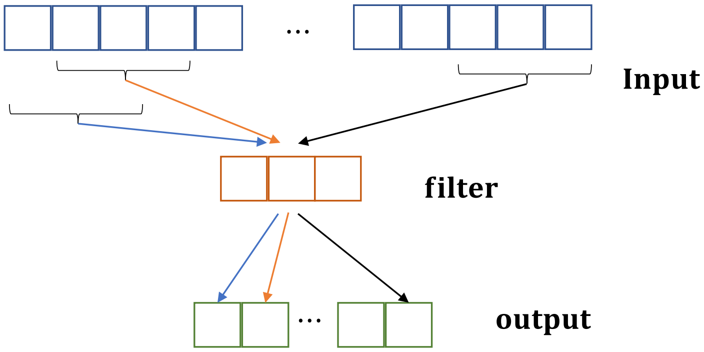


Figure 5: 1-dimensional convolutional layer.

Now we set the stride $s$ to be 1:

$$\mathbf{y} = \mathbf{x} * \mathbf{w} = \left( \sum_{i=1}^{k} w_i x_i, \sum_{i=1}^{k} w_i x_{i+1}, \cdots, \sum_{i=1}^{k} w_i x_{i+n-k} \right) \in \mathbb{R}^{n-k+1}.$$

Is the 1-dimensional convolutional operation a linear transformation? If so, please find the transformation matrix, then write down the gradient with respective to $\mathbf{x}$.

**Solution:**
Yes, it is. The transformation matrix $\mathbf{W} = (W_{ij}) \in \mathbb{R}^{(n-k+1)\times n}$ satisfies $\mathbf{y} = \mathbf{W}\mathbf{x}$, so

$$W_{ij} = \begin{cases} w_{j-i+1} & \text{if } i \le j \le i+k-1, \\ 0 & \text{otherwise.} \end{cases}$$

The gradient of $\mathbf{y}$ with respect to $\mathbf{x}$ is $\mathbf{W}^\top = (W_{ji}) \in \mathbb{R}^{n\times(n-k+1)}$. ■

2. $1 \times 1$ **convolutional layer.** Convolutional operations are linear transformations. We study a simple case, $1 \times 1$ convolutional operation, in this question. Suppose a convolutional layer takes as inputs the RGB $3 \times 28 \times 28$ images $\mathbf{X} = (x_{ijk}) \in \mathbb{R}^{3\times28\times28}$. Suppose that the convolutional layer has three $3 \times 1 \times 1$ filters where the $i^{\text{th}}$ filter is denoted by $\mathbf{w}_i \in \mathbb{R}^3$. We set stride $= 1$ and padding $= 0$.

Specifically, we denote the output by $\mathbf{Y} = (y_{ijk}) \in \mathbb{R}^{3\times 28\times 28}$, then

$$y_{ijk} = \sum_{t=1}^{3} w_{it}x_{tjk}, \quad i \in \{1,2,3\}, j,k \in \{1,\cdots,28\}.$$

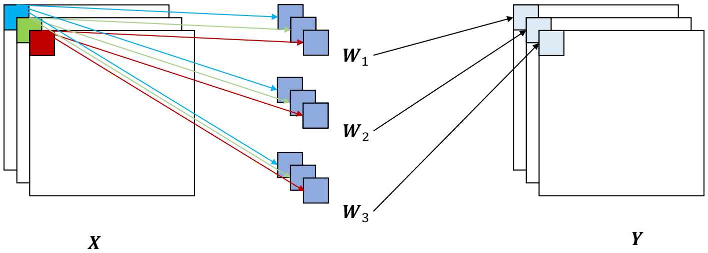


Figure 6: 1 × 1 convolutional layer.

Now we flatten the output $\mathbf{Y}$ to attain a $3 \times 28 \times 28$-dimensional vector,

$$\mathbf{y} = (y_{1,1,1}, y_{1,1,2}, \cdots, y_{1,1,28}, y_{1,2,1}, y_{1,2,2}, \cdots, y_{1,28,28}, y_{2,1,1}, y_{2,1,2}, \cdots, y_{3,28,28}).$$

We can also flatten $\mathbf{X}$ to attain a $3 \times 28 \times 28$-dimensional vector $\mathbf{x}$.

(a) Is the $1 \times 1$ convolutional operation a linear transformation? If so, Please find the transformation matrix.

(b) Please show that the $1 \times 1$ convolutional operation is invertible if and only if the matrix $(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3)$ is invertible.
**Hint**: let $A = (a_{ij})_{m\times m}$, $B \in \mathbb{R}^{n\times n}$, then the $mn \times mn$ matrix

$$\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \dots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{pmatrix}$$

is called the Kronecker product of $A$ and $B$, denoted by $A \otimes B$. Furthermore, $\det(A\otimes B) = (\det(A))^n(\det(B))^m$.

(c) Suppose $\mathbf{x}$ is sampled from a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, please find the density function of $\mathbf{y}$ if the $1 \times 1$ convolutional operation is invertible.

**Solution:**

(a) Yes.

(b)

(c) ■

3. **Pooling layer.** We know that average pooling and overlapping pooling are linear transformations, but not the max pooling.

(a) Suppose an average pooling layer has window size 2 × 2 and stride 2, as illustrated in the following picture. The pooling layer takes as inputs the 4 × 4 matrices. Please find the transformation matrix of the average pooling layer.

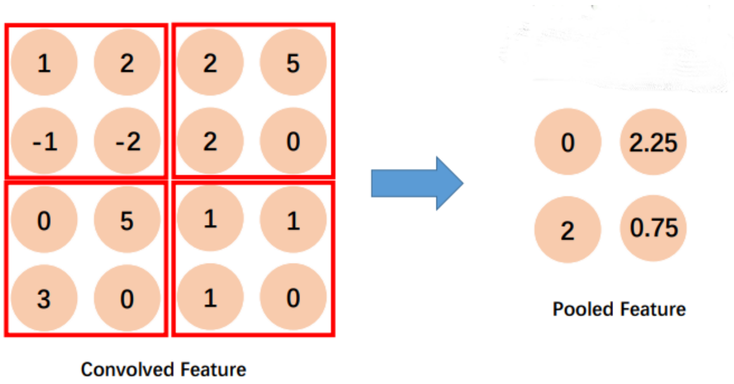


Figure 7: average pooling.

(b) Max pooling is generally not linear transformation. Consider the following example we studied in this course. The max pooling layer has window size 2 × 2 and stride 2, as illustrated in the following picture. The pooling layer takes as inputs the 4 × 4 matrices. Please find the subgradient of the max pooling operation. Then give an explanation of the "gradient" we studied in our course.

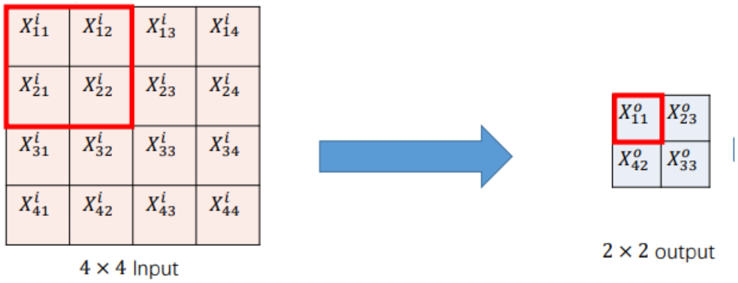


Figure 8: max pooling.

**Solution:**

(a)

(b) ■