Homework

PB22010344 黄境

2024 年 12 月 19 日

EXERCISE 1. **SVM for Linearly Separable Cases**

SOLUTION. 1. Denote $\{\mathbf{x}: f(\mathbf{x};\mathbf{w},b)=0\}$ by H. Set $\mathbf{x}_0 \in H$, we have

$$\begin{aligned} d(\mathbf{z},f) &= \|\mathbf{z}-\mathbf{x}_0\|\cos\langle \mathbf{z}-\mathbf{x}_0,\mathbf{w}\rangle \\ &= \|\mathbf{z}-\mathbf{x}_0\|\frac{|\langle \mathbf{z}-\mathbf{x}_0,\mathbf{w}\rangle|}{\|\mathbf{z}-\mathbf{x}_0\|\|\mathbf{w}\|} \\ &= \frac{|\langle \mathbf{z},\mathbf{w}\rangle - \langle \mathbf{x}_0,\mathbf{w}\rangle|}{\|\mathbf{w}\|} \\ &= \frac{|\langle \mathbf{z},\mathbf{w}\rangle + b|}{\|\mathbf{w}\|} \end{aligned}$$

2. By the definition of linear separable, there exists $(\hat{\mathbf{w}},\hat{b})$, s.t.

$$y_i f(\mathbf{x}_i;\hat{\mathbf{w}},\hat{b}) > 0,\ \forall\, i\in[n].$$

Since $n<\infty$, we have

$$m \triangleq \min_{i\in[n]} y_i f(\mathbf{x}_i;\hat{\mathbf{w}},\hat{b}) > 0.$$

Multiply $(\hat{\mathbf{w}},\hat{b})$ by $\frac{1}{m}>0$, we obtain

$$y_i f(\mathbf{x}_i;\frac{\hat{\mathbf{w}}}{m},\frac{\hat{b}}{m}) = \frac{1}{m} y_i f(\mathbf{x}_i;\hat{\mathbf{w}},\hat{b}) > \frac{1}{m}\cdot m = 1,\ \forall\, i\in[n].$$

Therefore, $\frac{1}{m}(\hat{\mathbf{w}}, \hat{b})$ is a feasible point of Problem (2), which implies that $\mathcal{F}$ is nonempty.

3. We prove that $\mathcal{F}$ is a nonempty, closed convex set first. From the last question we know that $\mathcal{F}$ is nonempty. Assume that

$$(\mathbf{w}_1, b_1),\ (\mathbf{w}_2, b_2) \in \mathcal{F},\ \theta \in [0, 1],$$

we have

$$\begin{aligned}
&\min_{i\in[n]} y_i(\langle \theta\mathbf{w}_1 + (1-\theta)\mathbf{w}_2, \mathbf{x}_i\rangle - (\theta b_1 + (1-\theta)b_2)) \\
&\geq \theta \min_{i\in[n]} y_i f(\mathbf{x}_i; \mathbf{w}_1, b_1) + (1-\theta)\min_{i\in[n]} y_i f(\mathbf{x}_i; \mathbf{w}_2, b_2) \\
&= \theta + (1-\theta) = 1
\end{aligned}$$

that is, $\theta(\mathbf{w}_1, b_1) + (1-\theta)(\mathbf{w}_2, b_2) \in \mathcal{F}$. Since the limit point of $\{(\mathbf{w}_i, b_i)\}$ has the same properties, similarly we can prove that $\mathcal{F}$ is closed. Therefore, $\mathcal{F}$ is a nonempty, closed convex set.

Since the objective function is strongly convex and continuous over its domain, by the Proposition 3 in Lecture 8, Problem (2) admits a unique global optimum, corresponding to a unique optimum point.

4. We only need to prove that at least one of the constraints holds as an equality at the optimum of Problem (2).

Assume that $(\mathbf{w}, b)$ is an optimum point of Problem (2), which satisfies

$$y_i f(\mathbf{x}_i; \mathbf{w}, b) > 1,\ \forall\, i \in [n].$$

Since $n < \infty$, we have

$$m \triangleq \min_{i\in[n]} y_i f(\mathbf{x}_i; \mathbf{w}, b) > 1.$$

Thus, set $(\mathbf{w}', b') = \frac{1}{m}(\mathbf{w}, b)$, we have

$$\min_{i\in[n]} y_i f(\mathbf{x}_i; \mathbf{w}', b') = 1,$$

which implies that

$$y_i f(\mathbf{x}_i; \mathbf{w}', b') \geq 1, \forall\, i \in [n].$$

Therefore, $(\mathbf{w}', b') \in \mathcal{F}$, while $\frac{1}{2}\|\mathbf{w}'\|^2 = \frac{1}{2m^2}\|\mathbf{w}\|^2 < \frac{1}{2}\|\mathbf{w}\|^2$, contraditing to the assumption that $(\mathbf{w}, b)$ is an optimum point. Thus complete the proof.

5.(a) We only need to prove that $(\mathbf{0}, b) \notin \mathcal{F}$. If not, we have

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = y_i \cdot b \geq 1, \forall\, i \in [n].$$

Since $\mathcal{D}^+$, $\mathcal{D}^-$ are nonempty, take $y_i = 1$, $y_j = -1$, we have $b \geq 1$, $-b \geq 1$, leading to the contradiction. Therefore, $\mathbf{w}^* \neq \mathbf{0}$.

(b) WLOG, let $y_i = 1$, $\forall\, i \in [n]$. Follow the same steps in (a), we obtain $(\mathbf{0}, 1) \in \mathcal{F}$. Since $\frac{1}{2}\|\mathbf{w}\|^2 \geq 0$, $(\mathbf{0}, 1)$ is an optimum point.

6. Proof by contradiction. WLOG, we assume that $(\mathbf{w}, b)$ is an optimum point and

$$\langle \mathbf{w}, \mathbf{x}_i\rangle + b > 1, \forall\, y_i = 1,$$

$$\langle \mathbf{w}, \mathbf{x}_i\rangle + b \leq -1, \forall\, y_i = -1,$$

$$\langle \mathbf{w}, \mathbf{x}_0\rangle + b = -1.$$

Similarly, we have

$$m \triangleq \min_{y_i=1} y_i f(\mathbf{x}_i; \mathbf{w}, b) > 1.$$

Since $\delta = \frac{m-1}{2} > 0$, set $(\mathbf{w}', b') = (\mathbf{w}, b - \delta)$. Therefore,

$$\langle \mathbf{w}', \mathbf{x}_i\rangle + b' > m - \delta > 1, \forall\, y_i = 1,$$

$$\langle \mathbf{w}', \mathbf{x}_i \rangle + b' \leq -1 - \delta < -1, \, \forall \, y_i = -1,$$

$$\langle \mathbf{w}', \mathbf{x}_0 \rangle + b' = -1 - \delta < -1.$$

Noticing that $\frac{1}{2}\|\mathbf{w}'\|^2 = \frac{1}{2}\|\mathbf{w}\|^2$, which implies that $(\mathbf{w}', b')$ is also an optimum point, in which none of the constraints holds as an equality, contradicting to the conclusion we've got in question 4.

7. We have already proved these properties in question 3 and 4 respectively.

8. Yes we can. Problem(2) is a convex optimization problem, and the training set is linearly separable, which implies that LICQ is satisfied. Therefore, it satisfies Slater's condition.

First, consider the Lagrange function:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1), \; \boldsymbol{\alpha} \geq \mathbf{0}.$$

Calculate $\min\limits_{\mathbf{w}, b} L(\mathbf{w}, b, \boldsymbol{\alpha})$, we get

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \hat{\mathbf{w}} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Therefore, the dual problem is

$$\max_{\alpha} L(\hat{\mathbf{w}}, \hat{b}, \boldsymbol{\alpha}), \;\; \text{s.t. } \boldsymbol{\alpha} \geq \mathbf{0}, \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Next, suppose that $\hat{\boldsymbol{\alpha}}$ is an optimum point of dual problem, we have

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i, \; \hat{b} = y_0 - \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_0^T \mathbf{x}_i$$

is an optimum point of primal problem, where $y_0(\langle \hat{\mathbf{w}}, \mathbf{x}_0 \rangle + b) = 1$.

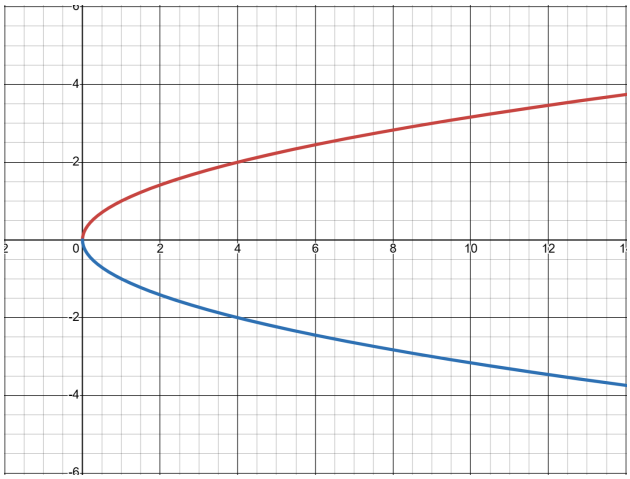
By the complementary slackness, for the inequalities that hold strictly, its corresponding $\alpha_i = 0$. Therefore, we can modify $\hat{\mathbf{w}}$ and $\hat{b}$ to:

$$\hat{\mathbf{w}} = \sum_{i \in I} \hat{\alpha}_i y_i \mathbf{x}_i, \; \hat{b} = y_0 - \sum_{i \in I} \hat{\alpha}_i y_i \mathbf{x}_0^T \mathbf{x}_i,$$

where $I = \{i : \; y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1\}$. Thus, we can remove the inequalities that hold strictly at the optimum without affecting the solution.

EXERCISE 2. **Discussions on Geometric Multiplier and Duality Gap**

SOLUTION. 1.(a) The gragh:



Since the set is supported by a hyperplane from below intercepts the vertical axis at the level f^*, the duality gap is zero.

On the other hand, the hyperplane's normal vector$(\lambda^*, 1)$ doesn't exist for $\lambda^* \to \infty$, therefore the geometric multiplier does not exist.

The Lagrangian dual function:

$$L(x, \lambda) = x + \lambda x^2, \; \lambda \geq 0$$

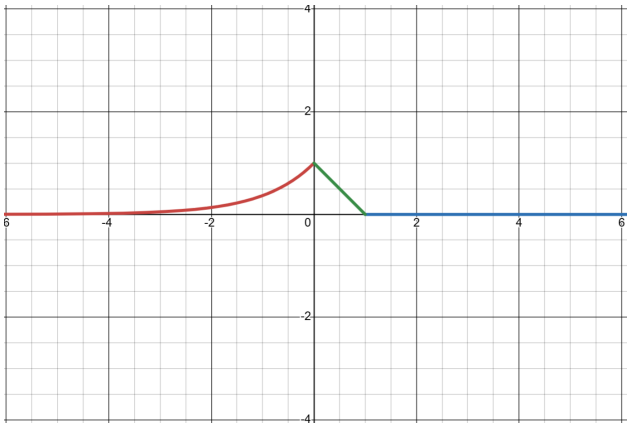
The dual problem:

$$\sup_{\lambda \geq 0} \inf_{x} x + \lambda x^2$$

Solve the dual problem and we get

$$x^* = 0.$$

(b) The gragh:



The supporting hyperplane is $\{(x, y) : y = 0\}$, therefore the geometric multiplier is $\lambda^* = 0$. Since the geometric multiplier exists, the dual gap is zero.

The Lagrangian dual function:

$$L(x, \lambda) = f(x) + \lambda x,\ \lambda \geq 0$$

The dual problem:

$$\sup_{\lambda \geq 0} \inf_{x} f(x) + \lambda x$$

Since

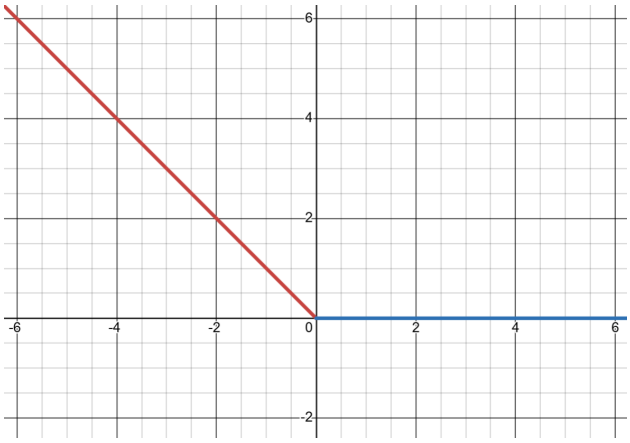
$$f(x) + \lambda x = \begin{cases} e^x + \lambda x,\ x < 0 \\ 1 - x + \lambda x,\ x \in [0, 1] \\ \lambda x,\ x > 0 \end{cases},$$

we have

$$\inf_x f(x) + \lambda x = \begin{cases} -\infty, \ \lambda > 0 \\ f^*, \ \lambda = 0 \end{cases}$$

Therefore, $\lambda^* = 0$, same as the λ^* we've found above.

(c) The gragh:



Similar to (a), the dual gap is zero, and $\lambda^* \in [0, 1]$.

The Lagrangian dual function:

$$L(x, \lambda) = f(x) + \lambda x, \ \lambda \geq 0$$

The dual problem:

$$\sup_{\lambda \geq 0} \inf_x f(x) + \lambda x$$

Therefore we get

$$f(x) + \lambda x = \begin{cases} x(1 - \lambda), \ x < 0 \\ \lambda x, \ x \geq 0 \end{cases},$$

which implies that

$$\inf_x f(x) + \lambda x = \begin{cases} -\infty, \ \lambda \in (1, \infty) \\ 0, \ \lambda \in [0, 1] \end{cases}$$

Therefore, $\lambda^* \in [0, 1]$.

2.(a) Incorrect. A counterexample is 1.(a).

(b) Incorrect. A counterexample is 1.(c).

(c) Correct. It is Proposition 1 in Lecture 13.

(d) Incorrect. 1.(a) can also be a counterexample.

(e) Incorrect. $\mathbf{x}^*$ for $f(\mathbf{x}^*) = f^* = \inf_{\mathbf{x}\in X}(\mathbf{x}, \lambda^*, \mu^*)$ may not exist.
Set $f(x) = f(x) + sgn(x)$ in 1.(b) and we get a counterexample, in which $\lambda^* = 0$ exists, while x^* does not exist, since $f^* = -1 = \lim_{x\to-\infty} f(x)$.

(f) Incorrect. A counterexample is 1.(b), in which $\mathbf{argmin}_{x\in X} L(x, \lambda^*, \mu^*) = [1, \infty)$, and obviously the feasible set $\mathcal{F} \cap [1, \infty) = \emptyset$.

EXERCISE 3. **Exercises of Dual Problems**

SOLUTION. 1. Denote the feasible set, optimal value and optimal solution by $\mathcal{F}$, m and $\mathbf{x}^*$ respectively.
(1)

$$\mathcal{F} = \{x : x \leq -2 \text{ or } 2 \leq x \leq 4,\ x \geq 0\} = [2, 4]$$

$$m = \min_{x\in\mathcal{F}}(x + 1)^2 + 1 = 10$$

$$x^* = 2$$

(2)

$$\mathcal{F} = \{(x_1, x_2) : x_1 + x_2 \leq 1\}$$

$$m = \min_{x \in \mathcal{F}} \frac{1}{2} e^{1-x_2} + \frac{1}{2} e^{1-x_2} + e^{2x_2} = 3(\frac{e}{2})^{\frac{2}{3}}$$

By the condition for taking the equality from the inequality of arithmetic and geometric means,

$$x^* = (\frac{2 + \ln 2}{3}, \frac{1 - \ln 2}{3})$$

2.(1) The Lagrangian dual function

$$q(x, \lambda_1, \lambda_2) = \inf_{x \in \mathbb{R}} x^2 + 2x + 3 + \lambda_1 (x+2)(x-2)(x-4) + \lambda_2 [1 - (x+1)^3]$$

The dual problem

$$\sup_{\lambda_1, \lambda_2 \geq 0} q(x, \lambda_1, \lambda_2)$$

The KKT condition

Primal Feasibility:

$x^* \in \mathcal{F}$

Dual Feasibility:

$\lambda_1,\ \lambda_2 \geq 0$

Lagrangian Optimality:

$x^* \in \mathbf{argmin}_{x \in \mathbb{R}}\ x^2 + 2x + 3 + \lambda_1 (x+2)(x-2)(x-4) + \lambda_2 [1 - (x+1)^3]$

Complementary Slackness:

$\lambda_1 (x+2)(x-2)(x-4) = 0,\ \lambda_2 [1 - (x+1)^3] = 0$

(2) The Lagrangian dual function

$$q(\mathbf{x}, \mu) = \inf_{\mathbf{x} \in \mathbb{R}^2} e^{x_1} + e^{2x_2} + \mu(x_1 + x_2 - 1)$$

The dual problem

$$\sup_{\mu} q(\mathbf{x}, \mu)$$

The KKT condition

Primal Feasibility:

$\mathbf{x}^* \in \mathcal{F}$

Lagrangian Optimality:

$\mathbf{x}^* \in \underset{(x_1,x_2)\in\mathbb{R}^2}{\textbf{argmin}}\ e^{x_1} + e^{2x_2} + \mu(x_1 + x_2 - 1)$

3. (1) Since objective function is convex, $\mathcal{F}$ is a convex set, the optimal value is finite, by Proposition 4 in Lecture 13, the strong duality holds.

(2) Consider $\frac{\partial L(\mathbf{x},\mu)}{\partial \mathbf{x}}$, we have

$$\frac{\partial L(\mathbf{x},\mu)}{\partial x_1} = e^{x_1} + \mu,$$

$$\frac{\partial L(\mathbf{x},\mu)}{\partial x_2} = 2e^{2x_2} + \mu,$$

thus for fixed μ,

$$q(\mathbf{x},\mu) = \begin{cases} -\infty,\ \mu \geq 0 \\ q(\ln(-\mu), \frac{1}{2}\ln(-\frac{\mu}{2}), \mu),\ \mu < 0 \end{cases}$$

Let $g(\mu) = q(\ln(-\mu), \frac{1}{2}\ln(-\frac{\mu}{2}), \mu) = -(\frac{5+\ln 2}{2})\mu + \frac{3}{2}\mu\ln(-\mu)$, we have

$$g'(\mu) = \frac{3}{2}(\ln(-\mu) - \frac{2+\ln 2}{3}) \Rightarrow \mu^* = -exp(\frac{2+\ln 2}{3}) < 0$$

and

$$g* = g(\mu^*) = \frac{3}{2}\mu^* = 3(\frac{e}{2})^{\frac{2}{3}} = m$$

Therefore, the strong duality holds.

EXERCISE 4. **The Dual Problem of SVM**

SOLUTION. 1.(a) Since objective function is convex, The feasible set is convex, and the optimal value is finite, by Proposition 4 in Lecture 13, the strong duality holds.

(b) Since the the form of dual problem is the same as which we've got in exercise 1.8, we claim that, by solving the problems in (2) and (5), we will arrive at the same separating hyperplane respectively.

2. The Lagrangian dual function

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i^p - \sum_{i=1}^{n}\lambda_i(y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) + \xi_i - 1) - \sum_{i=1}^{n}\lambda_{n+i}\xi_i$$

Take derivative, we obtain

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \hat{\mathbf{w}} = \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i$$

$$\nabla_b L = -\sum_{i=1}^{n}\lambda_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\lambda_i y_i = 0$$

$$\nabla_{\xi_i} L = Cp\xi_i^{p-1} - \lambda_i - \lambda_{n+i} = 0 \quad \Rightarrow \quad \xi_i = (\frac{\lambda_i + \lambda_{n+i}}{Cp})^{\frac{1}{p-1}}$$

Therefore, the dual problem has

$$\lambda_i \le pC\xi_i^{p-1},\ i = 1, 2, \ldots, n.$$
$$\lambda_{n+i} = C\xi_i^{p-1} - \lambda_i \ge 0,\ i = 1, 2, \ldots, n.$$

and we can remove λ_{n+i}.

Combining all together, the dual problem is

$$\max_{\boldsymbol{\lambda}\ge 0} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle + \sum_{i=1}^{n}\lambda_i + \frac{1-p}{C^{\frac{1}{p-1}} p^{\frac{p}{p-1}}}\sum_{i=1}^{n}(\lambda_i + \lambda_{n+i})^{\frac{1}{p-1}},$$

$$\text{s.t.} \sum_{i=1}^{n} \lambda_i y_i = 0,$$
$$\lambda_i \in [0, Cp\xi_i^{p-1}],\ i = 1, 2, \ldots, n.$$

3. Combining with the conclusion we've got in exercise 1.8, we have
(a)

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$$

for support vector $(\mathbf{x}_i, y_i)$.

(b)

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq -1$$

for misclassified point $(\mathbf{x}_i, y_i)$. (c)

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \in (-1, 1)$$

for point $(\mathbf{x}_i, y_i)$ in the region between the marginal hyperplanes.

4.(a) Similarly, we have

$$H_0 : \langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = 0$$

$$H_1 : \langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = 1$$

$$H_2 : \langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = -1$$

(b) Supporting vectors:x_2, x_5, x_7, x_8, x_{10}, x_{12}, x_{13}.
Non-support vectors: others.

EXERCISE 5. **Neural Networks**

SOLUTION. 1.(a) We have

$$J\mathbf{f}(\mathbf{x}) = (\frac{\partial f_i}{\partial x_j})_{ij}.$$

For $i = j$,

$$\begin{aligned}\frac{\partial f_i}{\partial x_i} &= \frac{e^{x_i}\sum_{k=1}^n e^{x_k} - e^{x_i}\cdot e^{x_i}}{(\sum_{k=1}^n e^{x_k})^2} \\ &= \frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}} - (\frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}})^2 \\ &= f_i(\mathbf{x})(1 - f_i(\mathbf{x}))\end{aligned}$$

For $i \neq j$,

$$\begin{aligned}\frac{\partial f_i}{\partial x_j} &= \frac{-e^{x_j}\cdot e^{x_i}}{(\sum_{k=1}^n e^{x_k})^2} \\ &= -\frac{e^{x_i}}{\sum_{k=1}^n e^{x_k}} \cdot \frac{e^{x_j}}{\sum_{k=1}^n e^{x_k}} \\ &= -f_i(\mathbf{x})f_j(\mathbf{x})\end{aligned}$$

Therefore, we have

$$J\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x})(1-f_1(\mathbf{x})) & -f_1(\mathbf{x})f_2(\mathbf{x}) & \dots & -f_1(\mathbf{x})f_n(\mathbf{x}) \\ -f_2(\mathbf{x})f_1(\mathbf{x}) & f_2(\mathbf{x})(1-f_2(\mathbf{x})) & \dots & -f_2(\mathbf{x})f_n(\mathbf{x}) \\ \dots & & & \dots \\ -f_n(\mathbf{x})f_1(\mathbf{x}) & -f_n(\mathbf{x})f_2(\mathbf{x}) & \dots & f_n(\mathbf{x})(1-f_n(\mathbf{x})) \end{pmatrix},$$

and $\nabla\mathbf{f}(\mathbf{x}) = (J\mathbf{f}(\mathbf{x}))^T$.

(b) Since

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

we have $\mathbf{f}(\mathbf{x} - c\mathbf{1}) = \mathbf{f}(\mathbf{x})$.

Calculating e^x directly is prone to overflow when $x \gg 0$. To prevent this situation, we can apply this transformation.

2. Denote $\frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}}$ by T_i. Since

$$\frac{\partial f_i}{\partial x_j} = \frac{1}{T_i}\frac{\partial T_i}{\partial x_j},$$

similar to question 1(a), we obtain

$$J\mathbf{f}(\mathbf{x}) = \begin{pmatrix} 1 - T_1 & -T_2 & \dots & -T_n \\ -T_1 & 1 - T_2 & \dots & -T_n \\ \dots & & & \dots \\ -T_1 & -T_2 & \dots & 1 - T_n \end{pmatrix},$$

Since $-x$ is convex, we only need to prove that $h(\mathbf{x}) = \ln \sum_{k=1}^{n} e^{x_k}$ is convex. By the property of e^x, we have

$$\begin{aligned}(\sum_{k=1}^{n} e^{\theta x_{1,k}})(\sum_{k=1}^{n} e^{(1-\theta)x_{2,k}}) &= \sum_{k=1}^{n} e^{\theta x_{1,k} + (1-\theta)x_{2,k}} + \sum_{p \neq q}^{n} e^{\theta x_{1,p} + (1-\theta)x_{2,q}} \\ &\geq \sum_{k=1}^{n} e^{\theta x_{1,k} + (1-\theta)x_{2,k}},\ \forall \theta \in [0,1].\end{aligned}$$

Take logarithm on both sides, we obtain

$$\begin{aligned}\theta h(\mathbf{x}_1) + (1-\theta)h(\mathbf{x}_2) &= \theta \ln \sum_{k=1}^{n} e^{x_{1,k}} + (1-\theta) \ln \sum_{k=1}^{n} e^{x_{2,k}} \\ &= \ln \sum_{k=1}^{n} e^{\theta x_{1,k}} + \ln \sum_{k=1}^{n} e^{(1-\theta)x_{2,k}} \\ &\geq \ln \sum_{k=1}^{n} e^{\theta x_{1,k} + (1-\theta)x_{2,k}} \\ &= h(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2)\end{aligned}$$

Thus, $-f_i$ is convex, which implies that f_i is concave.

3. The cross entropy can directly use the output of log softmax function, instead of calculating softmax function first, then take the logarithm. Since

$$f_i(\mathbf{x}) = \ln \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} = x_i - \ln \sum_{k=1}^{n} e^{x_k},$$

the amount of computation is much less than what softmax function needs. Furthermore, the form of gradient of log softmax function is more simple.

4.(a) We have

$$f_i = \sum_{j=1}^{3} w_{ij}^1 x_j = 1.2, \forall\, i.$$

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}} = \frac{1}{1+e^{-1.2}} = 0.7685, \forall\, 1.$$

$$y_i = \sum_{j=1}^{4} w_{ij}^2 \sigma(z_j) = 0.7685, \forall\, i.$$

Therefore, output $\mathbf{y} = (0.7685, 0.7685, 0.7685)$. The loss L is

$$L = -\sum_{i=1}^{4} p_i \ln softmax(y_i) = -\ln \frac{1}{3} = 1.0986.$$

(b) Since $L = L(\mathbf{y})$, where $\mathbf{y} = \mathbf{W}^2\mathbf{a} = \mathbf{W}^2\sigma(\mathbf{z})$, we have

$$\frac{\partial L}{\partial \mathbf{W}^2} = \frac{\partial L}{\partial \mathbf{y}} \mathbf{a}^T.$$

From

$$\frac{\partial L}{\partial y_i} = \frac{q_i}{y_i},$$

we have

$$\frac{\partial L}{\partial \mathbf{W}^2} = (\frac{q_i}{y_i} a_j)_{ij} = (\mathbf{q} \odot \frac{1}{\mathbf{y}}) \cdot \mathbf{a}^T$$

Therefore, by the chain rule we have

$$\frac{\partial L}{\partial \mathbf{W}^1} = (\frac{\partial \mathbf{y}}{\partial \mathbf{z}})^T \frac{\partial L}{\partial \mathbf{y}} \frac{\partial \mathbf{z}}{\partial \mathbf{W}^1}.$$

Respectively,

$$\frac{\partial \mathbf{z}}{\partial \mathbf{W}^1} = \mathbf{x}^T,$$

$$\frac{\partial L}{\partial \mathbf{y}} = \mathbf{q} \odot \frac{1}{\mathbf{y}},$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{W}^2 diag(\sigma'(z_1),\, \sigma'(z_2),\, \sigma'(z_3)) = \mathbf{W}^2 diag(\sigma'(\mathbf{z})).$$

Thus we obtain

$$\frac{\partial L}{\partial \mathbf{W}^1} = diag(\sigma'(\mathbf{z}))(\mathbf{W}^2)^T(\mathbf{q} \odot \frac{1}{\mathbf{y}})\mathbf{x}^T.$$

In this question, $\mathbf{x} = (1,\, 1,\, 1)$, $\mathbf{q} = (0,\, 0,\, 1)$, therefore we have

$$\frac{\partial L}{\partial \mathbf{W}^2} = \begin{pmatrix} \frac{a_1}{y_1} & \frac{a_2}{y_1} & \frac{a_3}{y_1} & \frac{a_4}{y_1} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\frac{\partial L}{\partial \mathbf{W}^1} = diag(\sigma'(\mathbf{z}))(\mathbf{W}^2)^T \frac{\partial L}{\partial \mathbf{W}^2}$$

(c) We have already represented them in the form of matrix.

(d) Set all the parameters to $\mathbf{0}$ in the formula above, we have

$$\frac{\partial L}{\partial \mathbf{W}^1} = \frac{\partial L}{\partial \mathbf{W}^2} = \mathbf{0},$$

which will make the gradient decent unable to work.

(e) Since

$$\sigma'(z_i) = \begin{cases} 0,\, z_i \leq 0 \\ 1,\, z_i > 0 \end{cases}$$

Replace σ' in (b) and we can get the new results.

EXERCISE 6. **Convolutional Neural Networks and Some Advanced Networks Structure**

SOLUTION. 1.

Layer	conv3-32	conv5-32	max pool	conv3-64
Feature map size	32×208×158	32×204×154	32×102×77	64×100×75
Number of parameters	288	800	1	576
Layer	conv5-64	max pool	FC-128	FC-10
Feature map size	64×96×71	64×48×36	128×1×1	10×1×1
Number of parameters	1600	1	14155776	1280

表 1: The parameters of the convolutional nerual network

The table is shown above.