Chapter 7 Solutions

Problem 7.1

The first and the second derivatives of function 'f(x)' are given by

$$f'(x) = 3x^2 + 12x - 3,$$
$$f''(x) = 6x$$

To find the positions of maxima and minima we solve equation for the points where the first derivative takes zero values:

$$3x^2 + 12x - 3 = 0 \Rightarrow x_{1,2} = -2 \pm \sqrt{5}$$

Substituting each of the roots into the second derivative we obtain that

$$f''(x_1) = f''(-2 + \sqrt{5}) > 0,$$
$$f''(x_2) = f''(-2 - \sqrt{5}) < 0,$$

i.e. $x_1 = -2 + \sqrt{5}$ is the minimum, whereas $x_2 = -2 - \sqrt{5}$ is the maximum of function $f(x)$.

Problem 7.2

The full batch update rule for gradient descent is

$$\theta_{i+1} = \theta_i - \gamma_i \sum_{n=1}^{N} (\nabla L_n(\theta_i))^T$$

For a mini-batch of size 1 at every step we would choose only one of the data points, i.e. we would choose randomly n among the values $1...N$ and calculate the update as

$$\theta_{i+1} = \theta_i - \gamma_i (\nabla L_n(\theta_i))^T$$

Problem 7.3

a)

True. Indeed, any two points in the intersection of the two convex sets can be connected by a path, which belongs to each of the sets (since they are convext), and therefore is also a part of the intersection.

b)

False. Indeed, if the two sets are completely disjoint, their union may not contain the points joining a point of one set to a point of the other.

c)

c)

False. Subtracting one set from the other may remove some of the points that belong to teh paths connecting the points of the remaining set.

Problem 7.4

a)

True. If $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex functions, it means that for any $\theta \in [0,1]$:

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}),$$
$$g(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta g(\mathbf{x}) + (1-\theta)g(\mathbf{y}).$$

By adding these inequalities we immediately obtain the condition of convexity for their sum, $h(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$:

$$h(\theta\mathbf{x} + (1-\theta)\mathbf{y}) = f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) + g(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) + \theta g(\mathbf{x}) + (1-\theta)g(\mathbf{y}) =$$

b)

False. For example; the difference of $f(x) = ax^2$ and $g(x) = bx^2$ is convex if $a \geq b$, but concave otherwise

c)

False. For example the product of two convex functions $f(x) = (x-a)^2$ and $g(x) = (x+a)^2$, $h(x) = (x-a)^2(x+a)^2$ has two minima and a maximum at $x = 0$. For examlple, if we take $x = a$, $y = -a$ and $\theta = 0.5$, then

$$h(\theta x + (1-\theta)y) = h(0) = a^4 > 0 = 0.5h(a) + 0.5h(-a) = \theta h(x) + (1-\theta)h(y),$$

i.e. the condition fo convexity is not satisfied.

d)

The maximum of a function is not a function, so it does not have property of convexity. If however we talk about the maximum as a convext set, consisting of only one point, then it is trivially convex.

Problem 7.5

We first introduce vectors $\mathbf{y} = (x_0, x_1, \xi)^T$ and $\mathbf{c} = (p_0, p_1, 1)^T$, so that $\mathbf{p}^T\mathbf{x} + \xi = \mathbf{c}^T\mathbf{y}$.

We then introduce vector $\mathbf{b} = (0, 3, 0)$ and matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

which allow us to write the three constraints as $\mathbf{A}\mathbf{y} \leq \mathbf{b}$.

We now can write the problem as a standard linear program:

$\max_{\mathbf{y}\in\mathbb{R}^3} \mathbf{c}^T\mathbf{y}$
subject to $\mathbf{A}\mathbf{y} \leq \mathbf{b}$

Problem 7.6

Let us define $\mathbf{c} = (-5, -3)^T$, $\mathbf{b} = (33, 8, 5, -1, 8)^T$ and

$$\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 2 & -4 \\ -2 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}$$

The linear program is then written as

$\min_{\mathbf{x}\in\mathbb{R}^2} \mathbf{c}^T\mathbf{x}$
subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$

The Lagrangian of this problem is

$$\mathcal{L}(\mathbf{x}, \lambda) = \mathbf{c}^T\mathbf{x} + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = (\mathbf{c}^T\mathbf{x} + \lambda^T\mathbf{A})\mathbf{x} - \lambda^T\mathbf{b} = (\mathbf{c}\mathbf{x} + \mathbf{A}^T\lambda)^T\mathbf{x} - \lambda^T\mathbf{b}$$

Taking gradient in respect to $\mathbf{x}$ and setting it to zero we obtain the extremum condition

$$\mathbf{c}\mathbf{x} + \mathbf{A}^T\lambda = 0,$$

that is

$$\mathcal{D}(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^2} \mathcal{L}(\mathbf{x}, \lambda) = -\lambda^T\mathbf{b}$$

that is the dual problem is given by

$\max_{\lambda\in\mathbb{R}^5} -\mathbf{b}^T\lambda$
subject to $\mathbf{c}\mathbf{x} + \mathbf{A}^T\lambda = 0$ and $\lambda \geq 0$

In terms of the original values of the parameters it can be thus written as

$$\max_{\lambda \in \mathbb{R}^5} -\begin{bmatrix} 33 \\ 8 \\ 5 \\ -1 \\ 8 \end{bmatrix}^T \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}$$

$$\text{subject to } -\begin{bmatrix} 5 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 & 2 & -2 & 0 & 0 \\ 2 & -4 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} = 0$$

$$\text{and } \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} \geq 0$$

Problem 7.7

We introduce $\mathbf{Q} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$, $\mathbf{c} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Then the quadratic problem takes form

$$\min_{\mathbf{x} \in \mathbb{R}^2} \tfrac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{c}^T \mathbf{x}$$
$$\text{subject to } \mathbf{A}\mathbf{x} \leq \mathbf{b}$$

The Lagrangian corresponding to this problem is

$$\mathcal{L}(\mathbf{x}, \lambda) = \tfrac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{c}^T \mathbf{x} + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \tfrac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + (\mathbf{c}^T + \mathbf{A}^T\lambda)^T \mathbf{x} - \lambda^T \mathbf{b}$$

where $\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix}$ We minimize the Lagrangian by setting its gradient to zero, which results in

$$\mathbf{Q}\mathbf{x} + \mathbf{c} + \mathbf{A}^T\lambda = 0 \Rightarrow \mathbf{x} = -\mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\lambda)$$

Substituting this back into the Lagrangian we obtain

$$\mathcal{D}(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^2} \mathcal{L}(\mathbf{x}, \lambda) = -(\mathbf{c} + \mathbf{A}^T\lambda)^T \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\lambda) - \lambda^T \mathbf{b}$$

The dual problem is now

$$\max_{\lambda \in \mathbb{R}^4} -(\mathbf{c} + \mathbf{A}^T\lambda)^T \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T\lambda) - \lambda^T \mathbf{b}$$
$$\text{subject to } \lambda \geq 0,$$

where the parameter vectors and matrices are defined above.

Problem 7.8

The primal problem can be written in teh standard form as

$\min_{\mathbf{w}\in\mathbb{R}^D} \frac{1}{2}\mathbf{w}^T\mathbf{w}$
subject to $1 - \mathbf{x}^T\mathbf{w} \le 0$

The Lagrangian is then

$\mathcal{L}(\mathbf{w}, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \lambda(1 - \mathbf{x}^T\mathbf{w})$

Taking gradient in respect to $\mathbf{w}$ we obtain the position of the minimum: $\mathbf{w} = \lambda\mathbf{x}$

Thus, the dual Lagrangian is

$\mathcal{D}(\lambda) = \min_{\mathbf{w}\in\mathbb{R}^D} \mathcal{L}(\mathbf{w}, \lambda) = -\frac{\lambda^2}{2}\mathbf{x}^T\mathbf{x} + \lambda$

Problem 7.9

Assuming standard dot product, convex conjugate is defined as

$f^*(\mathbf{s}) = \sup_{\mathbf{x}\in\mathbb{R}^D} \mathbf{s}^T\mathbf{x} - f(\mathbf{x})$

Since function $f(\mathbf{x})$ is continuous, differentiable and convex, looking for the supremum is equivalent to looking for the maximum, and can be done by setting the following gradient to zero:

$\nabla_{\mathbf{x}}\left(\mathbf{s}^T\mathbf{x} - f(\mathbf{x})\right) = \mathbf{s}^T - \nabla_{\mathbf{x}} f(\mathbf{x}) = 0$

that is $s_d = \frac{\partial}{\partial x_d} f(\mathbf{x}) = \log x_d - 1 \Rightarrow x_d = e^{s_d+1}$

Then

$f^*(\mathbf{s}) = \sum_{d=1}^D s_d x_d - \sum_{d=1}^D x_d \log x_d = \sum_{d=1}^D s_d e^{s_d+1} - \sum_{d=1}^D e^{s_d+1}(s_d + 1) = -\sum_{d=1}^D e^{s_d+1}$

Problem 7.10

Without loss of generality we can assume that $\mathbf{A} = \mathbf{A}^T$ is a symmetric matrix, as well as its inverse. By setting to zero gradient

$\nabla_{\mathbf{x}}\left(\mathbf{s}^T\mathbf{x} - f(\mathbf{x})\right) = \mathbf{s}^T - \mathbf{x}^T\mathbf{A} - \mathbf{b} = 0$

that is

$\mathbf{s} = \mathbf{A}\mathbf{x} + \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{A}^{-1}(\mathbf{s} - \mathbf{b})$

Therefore

$$f^*(\mathbf{s}) = \frac{1}{2}(\mathbf{s}-\mathbf{b})\mathbf{A}^{-1}(\mathbf{s}-\mathbf{b}) - c$$

Problem 7.11

By definition of convex conjugate:

$$L^*(\beta) = \sup_{\alpha\in\mathbb{R}} \beta\alpha - \max\{0, 1-\alpha\} = \sup_{\alpha\in\mathbb{R}} \begin{cases} \beta\alpha, \text{ if } \alpha > 1 \\ \beta\alpha - 1 + \alpha, \text{ if } \alpha < 1 \end{cases} = \sup_{\alpha\in\mathbb{R}} \begin{cases} \beta\alpha, \text{ if } \alpha > 1 \\ (\beta+1)\alpha - 1, \text{ if } \alpha <\end{cases}$$

We must here distinguish three cases:

a) if $\beta > 0$, then $\beta\alpha$ and $(\beta+1)\alpha - 1$ are both increasing with α, and take their maximum values at the maximum posible value of α, that is

$$L^*(\beta) = \sup_{\alpha\in\mathbb{R}} \begin{cases} +\infty, \text{ if } \alpha > 1 \\ \beta, \text{ if } \alpha < 1 \end{cases} = +\infty$$

b) if $\beta < -1$ then then $\beta\alpha$ and $(\beta+1)\alpha - 1$ are both decreasing with α, and take their maximum values at the minimum possible value of α, that is

$$L^*(\beta) = \sup_{\alpha\in\mathbb{R}} \begin{cases} \beta, \text{ if } \alpha > 1 \\ +\infty, \text{ if } \alpha < 1 \end{cases} = +\infty$$

c) Finally, if $-1 \le \beta \le 0$, then $\beta\alpha$ is decreasing with α, whereas $(\beta+1)\alpha - 1$ is growing, i.e.

$$L^*(\beta) = \sup_{\alpha\in\mathbb{R}} \begin{cases} \beta, \text{ if } \alpha > 1 \\ \beta, \text{ if } \alpha < 1 \end{cases} = \beta$$

We thus obtain

$$L^*(\beta) = \begin{cases} \beta \text{ if } -1 \le \beta \le 0 \\ +\infty \text{ otherwise} \end{cases}$$

Let us now define function $M(\alpha) = L^*(\beta) + \frac{\gamma}{2}\beta^2$ and calculate its convex conjugate function (we assume that $\gamma > 0$):

$$M^*(\alpha) = \sup_{\beta\in\mathbb{R}} \alpha\beta - L^*(\beta) - \frac{\gamma}{2}\beta^2 = \sup_{\beta\in[-1,0]} \alpha\beta - L^*(\beta) - \frac{\gamma}{2}\beta^2 = \sup_{\beta\in[-1,0]} \alpha\beta - \beta - \frac{\gamma}{2}\beta^2,$$

where the second equality is because outside of interval $[-1, 0]$ we have $-L^*(\beta) = -\infty$.

In interval $[-1, 0]$ parabolic function $\alpha\beta - \beta - \frac{\gamma}{2}\beta^2 = -\frac{\gamma}{2}\left(\beta - \frac{\alpha-1}{\gamma}\right)^2 + \frac{(1-\alpha)^2}{2\gamma}$ has a maximim at $\beta = \frac{\alpha-1}{\gamma}$, if this point is inside this interval, and at one of the interval edges otherwise. More precisely

$$M^(\alpha) = \begin{cases} 0, \text{ if } \alpha > 1 \\ \frac{(1-\alpha)^2}{2\gamma}, \text{ if } 0 \le \alpha \le 1 \\ 1 - \alpha - \frac{\gamma}{2}, \text{ if } \alpha > 1 \end{cases}$$