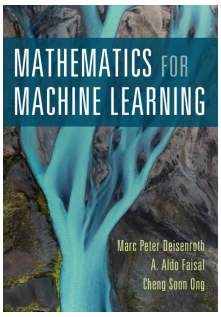
Mathematics for Machine Learning (Deisenroth, Faisal, Ong)

Chapter 7: Continuous Optimization

Full solutions to exercises 7.1 - 7.11



CONTENTS (clickable links)

7.1) $f(x) = x^3 + 6x^2 - 3x - 5$

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

$f''(x) = 6x + 12.$

- Stationary points when $f'(x) = 0$:

$f'(x) = 0$

$\Leftrightarrow 3x^2 + 12x - 3 = 0 \quad \Leftrightarrow x^2 + 4x - 1 = 0$

$\Leftrightarrow x = \frac{-4 \pm \sqrt{16 - 4(1)(-1)}}{2} = -2 \pm \sqrt{5}.$

- Are they minima or maxima?

Clearly $f''(x) < 0 \Leftrightarrow x < -2 \Leftrightarrow x$ is maximum.

$f''(x) > 0 \Leftrightarrow x > -2 \Leftrightarrow x$ is minimum.

Therefore; $-2 + \sqrt{5}$ is a minimum

$-2 - \sqrt{5}$ is a maximum.

7.2) $\theta_{i+1} = \theta_i - \gamma_i \nabla L_1(\theta_i)$ (can assume the L_n we chose is L_1.)

7.3 a) True. Let S, T be 2 convex sets. Let $a, b \in S \cap T$.

As $a, b \in S$, the line segment $ta + (1-t)b$, $t \in [0,1] \in S$

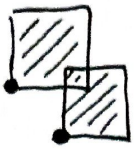
As $a, b \in T$, the line segment $ta + (1-t)b$, $t \in [0,1] \in T$

Therefore, the line segment $ta + (1-t)b$, $t \in [0,1] \in S \cap T$.

i.e., $S \cap T$ is convex.

b) False. e.g.,

is convex



is not (consider line between the bottom left corners of the 2 squares.)

c) False e.g.,

and there's no line between points in the left rectangle & the right rectangle.

7.4a) True.

Let f, g be convex functions

i.e., $f(ta+(1-t)b) \leq tf(a) + (1-t)f(b)$ for all $t \in [0,1]$, $a, b \in \mathbb{R}$.

and similarly for g.

Then:

$$(f+g)(ta+(1-t)b) = f(ta+(1-t)b) + g(ta+(1-t)b)$$

$$\leq tf(a) + (1-t)f(b) + tg(a) + (1-t)g(b)$$

$$= t(f+g)(a) + (1-t)(f+g)(b)$$

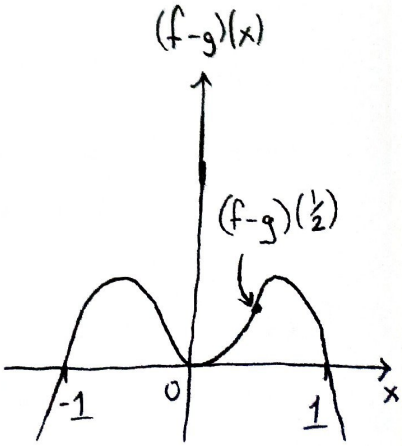
for all $t \in [0,1]$, $a, b \in \mathbb{R}$.

b) False.

Consider $f(x) = x^2$, $g(x) = x^4$

Then f, g are convex (look like a bowl)

But $f-g$ is not convex. e.g., $a=0$, $b=1$, $t=\frac{1}{2}$.

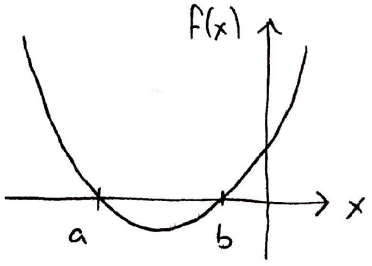


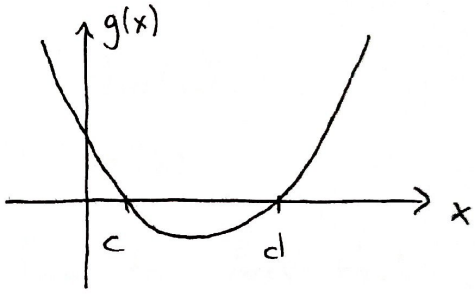
c) False.

- One idea is to create a function which "wiggles" i.e., ~~\/\/~~ by choosing convex functions which are $\pm$ at different parts of the domain.

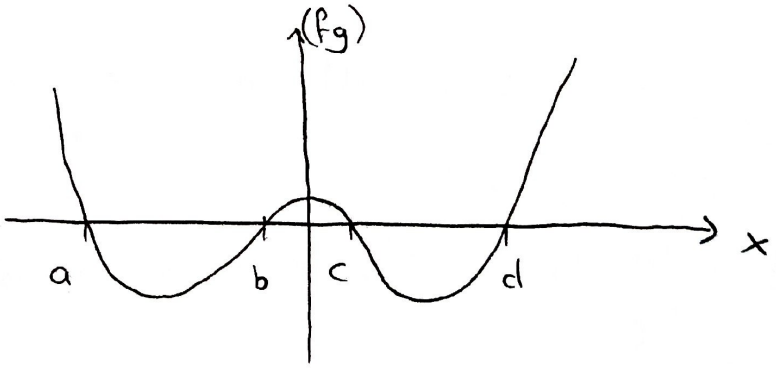
interval	$(-\infty, a)$	(a, b)	(b, c)	(c, d)	(d, ∞)
sign of f	+	−	+	+	+
sign of g	+	+	+	−	+
sign of fg	+	−	+	−	+ .

Example: $f(x) = (x+1)^2 - \frac{1}{2}$, $g(x) = (x-1)^2 - \frac{1}{2}$





then $(fg)(x) = \left((x+1)^2 - \frac{1}{2}\right)\left((x-1)^2 - \frac{1}{2}\right)$ has graph

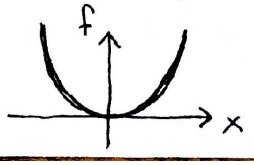


f, g are convex. (fg) is not. (e.g. consider the line between the two local minima.)

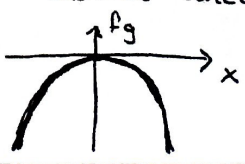
- Much easier example (just realised that constant functions are convex.)

$f(x) = x^2$

$g(x) = -1$



product of f w/ g. →



d) True.

Let f, g be two convex functions.

i.e $f(ta+(1-t)b) \leq tf(a) + (1-t)f(b)$ for all $t \in [0,1]$, $a, b \in \mathbb{R}$.

& similarly for g.

Let $m := \max\{f, g\}$, i.e., $m(x) = \max\{f(x), g(x)\}$ for all $x \in \mathbb{R}$.

Then, $m(ta+(1-t)b) = \begin{cases} f(ta+(1-t)b), & \text{if } f(ta+(1-t)b) > g(ta+(1-t)b) \\ g(ta+(1-t)b), & \text{otherwise.} \end{cases}$

However, we have:

$\rightarrow f(ta+(1-t)b) \leq tf(a)+(1-t)f(b) \leq tm(a)+(1-t)m(b)$

$\rightarrow g(ta+(1-t)b) \leq tg(a)+(1-t)g(b) \leq tm(a)+(1-t)m(b)$

with the last inequality following immediately from the fact that $m(a) = \max\{f(a), g(a)\}$, $m(b) = \max\{f(b), g(b)\}$.

Therefore, $m(ta+(1-t)b) \leq tm(a)+(1-t)m(b)$

7.5) $p^T x + \xi = p_0 x_0 + p_1 x_1 + 1 \cdot \xi = (p_0, p_1, 1)\begin{pmatrix} x_0 \\ x_1 \\ \xi \end{pmatrix} = \begin{pmatrix} p_0 \\ p_1 \\ 1 \end{pmatrix}^T \begin{pmatrix} x_0 \\ x_1 \\ \xi \end{pmatrix}$

The constraints $\xi \geq 0$ $(\Leftrightarrow -\xi \leq 0)$, $x_0 \leq 0$, $x_1 \leq 3$ can be written as:

$$\begin{aligned} 1x_0 + 0x_1 + 0\xi &\leq 0 \\ 0x_0 + 1x_1 + 0\xi &\leq 3 \\ 0x_0 + 0x_1 - 1\xi &\leq 0 \end{aligned} \quad \Leftrightarrow \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}\begin{pmatrix} x_0 \\ x_1 \\ \xi \end{pmatrix} \leq \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix}.$$

So optimisation problem is:

$$\max_{\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}\in\mathbb{R}^3} \begin{pmatrix}p_0\\p_1\\1\end{pmatrix}^T\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}, \text{ subject to } \begin{pmatrix}1&0&0\\0&1&0\\0&0&-1\end{pmatrix}\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}\leq\begin{pmatrix}0\\3\\0\end{pmatrix}$$

$$\Updownarrow$$

$$\min_{\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}\in\mathbb{R}^3} \begin{pmatrix}-p_0\\-p_1\\-1\end{pmatrix}^T\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}, \text{ subject to } \begin{pmatrix}1&0&0\\0&1&0\\0&0&-1\end{pmatrix}\begin{pmatrix}x_0\\x_1\\s\end{pmatrix}\leq\begin{pmatrix}0\\3\\0\end{pmatrix}.$$

7.6) let $c=\begin{pmatrix}5\\3\end{pmatrix}$, $A=\begin{pmatrix}2&2\\2&-4\\-2&1\\0&-1\\0&1\end{pmatrix}$, and $b=\begin{pmatrix}33\\8\\5\\-1\\8\end{pmatrix}$

Then the primal problem is:

$$\min_{\underline{x}} -c^T\underline{x}, \text{ subject to } A\underline{x}-b\leq\underline{0}$$

This means each component $(A\underline{x}-b)_i$ is ≤ 0.

The Lagrangian is:

$$h(\underline{x},\lambda) = -c^T\underline{x}+\lambda^T(A\underline{x}-b) = (\lambda^TA-c^T)\underline{x}-\lambda^Tb.$$
$$= (A^T\lambda-c)^T\underline{x}-\lambda^Tb.$$

We require $D(\lambda)=\min_{\underline{x}} h(\underline{x},\lambda)$.

However, $h(\underline{x},\lambda)$ is differentiable so minimum is easily found.

We have $\frac{\partial}{\partial\underline{x}}(h(\underline{x},\lambda)) = (A^T\lambda-c)^T$

so minimum occurs when $A^T\lambda-c=\underline{0}$.

- The dual problem is:

$$\max_{\lambda} -\lambda^Tb=-b^T\lambda, \text{ subject to } \lambda\geq\underline{0}$$
$$c=A^T\lambda.$$

7.7) Let $Q = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$, $c = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$, $A = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

Then the quadratic program is:

$$\min_{\underline{x}} \ \frac{1}{2}\underline{x}^T Q \underline{x} + c^T \underline{x}, \text{ subject to } A\underline{x} - b \leq \underline{0}.$$

↖ This means each component $(A\underline{x}-b)_i \leq 0$.

The Lagrangian is:

$$\begin{aligned} h(\underline{x}, \lambda) &= \frac{1}{2}\underline{x}^T Q \underline{x} + c^T \underline{x} + \lambda^T (A\underline{x} - b) \\ &= \frac{1}{2}\underline{x}^T Q \underline{x} + (c + A^T\lambda)^T \underline{x} - \lambda^T b. \end{aligned}$$

As $h(\underline{x}, \lambda)$ is differentiable w.r.t $\underline{x}$, $\min_{\underline{x}} h(\underline{x}, \lambda)$ can easily be found.

$$\frac{d}{d\underline{x}}\left(h(\underline{x}, \lambda)\right) = \underline{x}^T Q + (c + A^T\lambda)^T. \qquad \text{(using Eq. 5.105 \& 5.107)}$$

Setting this equal to $\underline{0}$ (& transposing along with $Q = Q^T$):

$$Qx = -(c + A^T\lambda) \quad \Rightarrow \quad x = -Q^{-1}(c + A^T\lambda).$$

Subbing this into $h(\underline{x}, \lambda)$ yields:

$$\begin{aligned} D(\lambda) := \min_{\underline{x}} h(\underline{x}, \lambda) &= \frac{1}{2}(c + A^T\lambda)^T Q^{-1} Q Q^{-1}(c + A^T\lambda) \\ &\quad - (c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b \\ &= -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b. \end{aligned}$$

So, the dual problem is;

$$\max_{\lambda} \left(-\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - b^T\lambda \right), \text{ subject to } \lambda \geq \underline{0}.$$

7.8) $\min_{\omega} \frac{1}{2}\omega^T\omega$, subject to $1 - \omega^T x \leq 0$.

Langrangian is

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

$$= \frac{1}{2}\omega^T\omega - \lambda\omega^T x + \lambda.$$

This is differentiable w.r.t. ω, doing so gives:

$$\frac{\partial \mathcal{L}(\omega, \lambda)}{\partial \omega} = \omega^T - \lambda x^T$$

So minimum obtained when $(\omega - \lambda x)^T = 0 \iff \omega = \lambda x$.

Subbing this into $\mathcal{L}(\omega, \lambda)$ gives:

$$D(\lambda) = \frac{\lambda^2}{2} x^T x + \lambda(1 - \lambda x^T x) = -\frac{\lambda^2}{2} x^T x + \lambda.$$

$\therefore$ the dual problem is:

$$\max_{\lambda} \left(-\frac{\lambda^2}{2} x^T x + \lambda\right), \text{ subject to } \lambda \geq 0.$$

7.9) $f(\underline{x}) = \sum_{d=1}^{D} x_d \log(x_d) = \sum_{d=1}^{D} g(x_d)$, where $g: \mathbb{R} \to \mathbb{R}$, $t \mapsto t\log t$.

i.e. f is a sum of functions.

Following Example 7.8,

$$f^*(\underline{s}) = \sum_{d=1}^{D} g^*(s_d)$$

We just need to find g^* (& this is easy since g is convex & differentiable.)

$$g^*(u) = \sup_t \left(\langle u, t \rangle - g(t) \right)$$

$$= \sup_t (ut - t\log t)$$

Differentiating w.r.t. t & solving "$= 0$" will give maximum.

$$\frac{d}{dt}(ut - t\log t) = u - \log t - 1.$$

We have $u - \log t - 1 = 0 \iff t = e^{u-1}$.

Subbing this into expression for $g^*(u)$ yields:

$$g^*(u) = ue^{u-1} - e^{u-1}(u-1) = e^{u-1}.$$

Therefore, $f^*(\underline{s}) = \sum_{d=1}^{D} g^*(s_d) = \sum_{d=1}^{D} e^{s_d - 1}$.

7.10) $f^*(\underline{s}) = \sup_{\underline{x}} \left(\underline{s}^T \underline{x} - f(\underline{x}) \right) = \sup_{\underline{x}} \left(\underline{s}^T \underline{x} - \frac{1}{2} \underline{x}^T A \underline{x} - b^T \underline{x} - c \right)$

To find the maximum, we differentiate and set "$= 0$".

$$\frac{d}{d\underline{x}} \left(\underline{s}^T \underline{x} - f(\underline{x}) \right) = \underline{s}^T - \underline{x}^T A - b^T$$

Setting equal to $\underline{0}$ and solving for $\underline{x}$ yields:

$$\underline{s} - A\underline{x} - b = 0 \iff \underline{x} = A^{-1}(\underline{s} - b)$$

Subbing into $f^*(\underline{s})$ gives:

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

$$= \frac{1}{2}(\underline{s} - b)^T A^{-1}(\underline{s} - b) - c.$$

7.11) $L(\alpha) = \max\{0, 1-\alpha\}$.

So $L^*(\beta) = \sup_\alpha \left(\alpha\beta - \max\{0, 1-\alpha\}\right)$ — 2 cases $\alpha \geq 1$ & $\alpha \leq 1$

$$= \max \begin{cases} \sup_\alpha(\alpha\beta), \text{ with } \alpha \geq 1, \\ \sup_\alpha(\alpha\beta + \alpha - 1), \text{ with } \alpha \leq 1 \end{cases}$$

(i.e., to find the sup overall, find $\sup_{\alpha\geq 1}$ & $\sup_{\alpha\leq 1}$ and then take the maximum of these 2 values.)

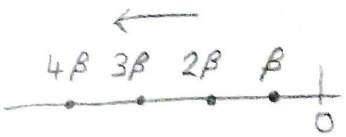
$$L^*(\beta) = \max \begin{cases} \sup_\alpha(\alpha\beta) \text{ with } \alpha \geq 1, \\ \sup_\alpha\left(\alpha(\beta+1) - 1\right) \text{ with } \alpha \leq 1. \end{cases}$$

There are different cases to consider, depending on the signs of β and $(\beta+1)$:

- $\beta > 0$: $L^*(\beta) = \max\left\{ \begin{matrix} \infty, \\ (\beta+1)-1 = \beta \end{matrix} \right\} = \infty$.

- $\beta = 0$: $L^*(\beta) = \max\left\{ \begin{matrix} 0, \\ \sup_\alpha(\alpha-1) \text{ with } \alpha \leq 1 \end{matrix} \right\} = 0. = \beta$.

- $-1 < \beta < 0$: $L^*(\beta) = \max\left\{ \begin{matrix} \beta, \\ (\beta+1)-1 = \beta \end{matrix} \right\} = \beta$.



- $\beta = -1$: $L^*(\beta) = \max\left\{ \begin{matrix} \sup_\alpha(-\alpha), \alpha \geq 1 \\ -1 \end{matrix} \right\} = -1 = \beta$.

- $\beta < -1$: $L^*(\beta) = \max\left\{ \begin{matrix} \sup_\alpha(-|\beta|\alpha), \alpha \geq 1 \\ \sup_\alpha(-|\beta+1|\alpha - 1), \alpha \leq 1 \end{matrix} \right\} = \max\left\{ \begin{matrix} \beta \\ \infty \end{matrix} \right\} = \infty$

consider $\alpha \to -\infty$.

So $L^*(\beta) = \begin{cases} \infty & \text{if } \beta > 0 \\ \beta & \text{if } -1 \le \beta \le 0 \\ \infty & \text{if } \beta < -1 \end{cases}$.

$$(L^*)^*(u) = \sup_{\beta} \left(u\beta - L^*(\beta) - \frac{\gamma}{2}\beta^2 \right).$$

if $\beta \notin [-1, 0]$ then $-L^*(\beta) = -\infty$; - very bad news.

$$= \sup_{\beta \in [-1,0]} \left(u\beta - \beta - \frac{\gamma}{2}\beta^2 \right).$$

$$= \sup_{\beta \in [-1,0]} \left(-\frac{\gamma}{2}\beta^2 + (u-1)\beta \right).$$

- If $\gamma > 0$, then quadratic, $-\frac{\gamma}{2}\beta^2 + (u-1)\beta$, has $\cap$ shape.

Find local maximum: $\frac{d}{d\beta}\left(-\frac{\gamma}{2}\beta^2 + (u-1)\beta \right) = -\gamma\beta + u - 1.$

so local maximum at $\beta = \frac{u-1}{\gamma}$.

Find values at extremes of interval, -1 & 0: $\left(-\frac{\gamma}{2}\beta^2 + (u-1)\beta \right)\Big|_{\beta=-1} = 1 - u - \frac{\gamma}{2}.$

$$\left(-\frac{\gamma}{2}\beta^2 + (u-1)\beta \right)\Big|_{\beta=0} = 0.$$

→ If local maxima $\beta = \frac{u-1}{\gamma} \in [-1, 0]$ $\left(\Leftrightarrow -\gamma + 1 \le u \le 1 \right)$

then $(L^*)^*(u) = -\frac{\gamma}{2}\left(\frac{u-1}{\gamma}\right)^2 + (u-1)\left(\frac{u-1}{\gamma}\right) = \frac{1}{2}\frac{(u-1)^2}{\gamma}.$

→ If $\beta > 0$ $(\Leftrightarrow u > 1)$ then supremum occurs at $\beta = 0$.

Thus, $(L^*)^*(u) = 0$.

→ If $\beta < -1$ $(\Leftrightarrow u < -\gamma + 1)$ then supremum occurs at $\beta = -1$.

Thus $(L^*)^*(u) = 1 - u - \frac{\gamma}{2}$.

In summary;

If $\gamma > 0$, then $(L^*)^*(u) = \begin{cases} 1 - u - \frac{\gamma}{2}, & \text{if } u < -\gamma + 1 \\ \frac{1}{2}\frac{(u-1)^2}{\gamma}, & \text{if } -\gamma + 1 \leq u \leq 1 \\ 0, & \text{if } u \geq 1 \end{cases}$

- If $\gamma \leq 0$, then $-\frac{\gamma}{2}\beta^2 + (u-1)\beta$ looks like $\cup$

(or $\backslash$ or $/$ in the case $\gamma = 0$).

In any case, there is no local extrema and so

$$(L^*)^*(u) = \max\left\{0, 1 - u - \frac{\gamma}{2}\right\}$$

values at the interval extrema, $\beta = -1, 0$.

Remark:

If $\gamma = 0$, then $(L^*)^* = L$.