7. Continuous Optimization (Exercises Only)

Exercises:

7.1

- Consider the univariate function:

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

- Find its stationary points and indicate whether they are maximum, minimum, or saddle points
- To find its stationary points we find the derivative and set it equal to 0:

$$f'(x) = 3x^2 + 12x - 3 = 0$$

- Now we solve for x. Unfortunately this doesn't work out nicely and we need to use the quadratic formula, but we can start by dividing both sides of the equation by 3:

$$x^2 + 4x - 1 = 0$$

$$x = \frac{-4 \pm \sqrt{4^2 - 4(1)(-1)}}{2}$$

$$x = \frac{-4 \pm 2\sqrt{5}}{2}$$

$$x = -2 \pm \sqrt{5}$$

- Therefore, the stationary points are:

$$x_1 = -2 + \sqrt{5} \qquad x_2 = -2 - \sqrt{5}$$

- Now, to find out if these are maximum, minimum, or saddle points we need to do the second derivative test:

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

$$f''(x_1) = 6\sqrt{5}$$

- $f''(x_1)$ is positive therefore this is a **minimum**

$$f''(x_2) = -6\sqrt{5}$$

- $f''(x_2)$ is negative therefore this is a **maximum**

7.2

- Consider the update equation for stochastic gradient descent:

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

- Write down the update when we use a minibatch size of one:

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

7.3

- Consider whether the following statements are true or false:
  a. The intersection of any two convex sets is convex
  - This is **true**
  b. The union of any two convex sets is convex
  - This is **false**. The union of two convex sets could form a shape which is nonconvex.
  c. The difference of a convex set A from a convex set B is convex
  - This is **false**. Consider a convex set A taking a "bite" out of convex set B. B may no longer be a convex set.

7.4

- Consider whether the following statements are true or false:
  a. The sum of any two convex functions is convex.
  - This is **true**. Consider two convex functions f_1 and f_2:

$$f_1(\theta x + (1-\theta)y) \leq \theta f_1(x) + (1-\theta) f_1(y)$$
$$f_2(\theta x + (1-\theta)y) \leq \theta f_2(x) + (1-\theta) f_2(y)$$

```
   - Adding these functions gives us:
```

$$f_1(\theta x + (1-\theta)y) + f_x(\theta x + (1-\theta)y) \leq \theta f_1(x) + \theta f_2(x) + (1-\theta) f_1(y) + (1-\theta) f_2(y)$$

```
 - Rearranging the right side we see that this satisfies the definition of convexity:
```

$$f_1(\theta x + (1-\theta)y) + f_x(\theta x + (1-\theta)y) \leq \theta(f_1(x) + f_2(x)) + (1-\theta)(f_1(y) + f_2(y))$$

```
b. The difference of any two convex functions is convex
   - This is **false**.
```

```
c. The product of any two convex functions is convex
    - This is **false**
d. The maximum of any two convex functions is convex.
    - I don't even understand the question. The maximum is not a function.
```

7.5

- Express the following optimization problem as a standard linear program in matrix notation

$$\max_{x \in \mathbb{R}^2, \zeta \in \mathbb{R}} p^T x + \zeta$$

- Subject to the constraints that:
  1. $\zeta \geq 0$
  2. $x_0 \leq 0$
  3. $x_1 \leq 3$
- We need to rewrite the constraints in the form: $Ay \leq b$

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

$$y = \begin{bmatrix} \zeta \\ x_0 \\ x_1 \end{bmatrix}$$

$$b = \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$$

7.6

- Consider the linear program illustrated by Figure 7.9:

$$\min_{x \in \mathbb{R}^2} - \begin{bmatrix} 5 \\ 3 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\text{subject to} \begin{bmatrix} 2 & 2 \\ 2 & -4 \\ -2 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 33 \\ 8 \\ 5 \\ -1 \\ 8 \end{bmatrix}$$

- The dual is:

$$\mathcal{D}(\lambda) = -\begin{bmatrix} 33 \\ 8 \\ 5 \\ -1 \\ 8 \end{bmatrix}^T \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}$$

$$\text{subject to} \quad \begin{array}{c} c + A^T\lambda = 0 \\ \lambda \geq 0 \end{array}$$

7.7

- Consider the quadratic program illustrated in Figure 7.4:

$$\min_{x \in \mathbb{R}^2} \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\text{subject to} \quad \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

- Derive the dual quadratic program using Lagrange duality
- Starting from the general form of a quadratic program:

$$\min_{x \in \mathbb{R}} \frac{1}{2} x^T Q x + c^T x$$

$$\text{subject to} \quad Ax \leq b$$

- We start by forming the Lagrangian:

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

- distributing the λ^T:

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

- Factoring x out of the linear terms:

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

- Now we can differentiate this with respect to x and set it equal to zero:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial}{\partial x} (\frac{1}{2} x^T Q x) + \frac{\partial}{\partial x} [(c + \lambda^T A) x] - \frac{\partial}{\partial x} (\lambda^T b) = 0$$

- Differentiating the first term:

$$\frac{\partial}{\partial x} = Qx$$

- And the second term:

$$\frac{\partial}{\partial x}[(c + \lambda^T A)x] = (c + \lambda^T A)$$

- And the third term:

$$\frac{\partial}{\partial x}(\lambda^T b) = 0$$

- Therefore:

$$\frac{\partial \mathcal{L}}{\partial x} = Qx + (c + \lambda^T A) = 0$$

- Now we can solve for x:

$$\begin{aligned} Qx &= -(c + \lambda^T A) \\ x &= -Q^{-1}(c + \lambda^T A) \end{aligned}$$

- Now we can plug this back into our Lagrangian equation to form the Lagrangian Dual:

$$\mathcal{D}(\lambda) = \frac{1}{2}\left[-Q^{-1}(c + \lambda^T A)\right]^T Q \left[-Q^{-1}(c + \lambda^T A)\right] + (c + \lambda^T A)^T \left[-Q^{-1}(c + \lambda^T A)\right] - \lambda^T b$$

- To make simplifying this easier let:

$$z = (c + \lambda^T A)$$

$$\begin{aligned} \mathcal{D}(\lambda) &= \frac{1}{2}\left(-Q^{-1}z\right)^T Q \left(-Q^{-1}z\right) + c^T \left(-Q^{-1}z\right) - \lambda^T b \\ &= \frac{1}{2}z^T Q^{-1} Q(Q^{-1}z) + z^T Q^{-1} z - \lambda^T b \\ &= \frac{1}{2}z^T Q^{-1} z + z^T Q^{-1} z - \lambda^T b \\ &= -\frac{1}{2}z^T Q z - \lambda^T b \end{aligned}$$

- Now substituting z back in we get the final form of our Lagrangian dual:

$$\mathcal{D}(\lambda) = -\frac{1}{2}(c + \lambda^T A)Q^{-1}(c + \lambda^T A) - \lambda^T b$$

$$\text{subject to} \quad \begin{aligned} Ax - b = 0 \\ \lambda \geq 0 \end{aligned}$$

- Looking at our original equation, let:

$$c = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

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

$$b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$Q^{-1} = \frac{1}{6}\begin{bmatrix} -4 & 1 \\ 1 & -2 \end{bmatrix}$$

7.8

- Consider the following convex optimization problem:

$$\min_{w \in \mathbb{R}^D} \frac{1}{2} w^T w$$

$$\text{subject to} \quad w^T x \geq 1$$

- Derive the Lagrangian dual by introducing the Lagrange multiplier λ
- First thing's first we need to get the inequality constraint in the standard form: $g(x) \leq 0$:

$$\begin{aligned} w^T x &\geq 1 \\ w^T x - 1 &\geq 0 \\ 1 - w^T x &\leq 0 \end{aligned}$$

- Now we can form the Lagrangian:

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

- Differentiating with respect to w and setting the gradient equal to zero:

$$\frac{\partial \mathcal{L}}{\partial w} = w - \lambda x = 0$$

- Solving for w:

$$w = \lambda x$$

- Plugging this equation for w back into the Lagrangian gives us the dual:

$$\begin{aligned} \mathcal{D}(\lambda, x) &= \frac{1}{2}(\lambda x)^T(\lambda x) + \lambda(1 - (\lambda x)^T x) \\ &= \frac{\lambda^2}{2} x^T x + \lambda - \lambda^2 x^T x \\ &= -\frac{\lambda}{2} x^T x + \lambda \end{aligned}$$

- So our dual optimization problem becomes:

$$\max_{x \in \mathbb{R}^D, \lambda \geq 0} -\frac{\lambda}{2} x^T x + \lambda$$

$$\text{subject to} \quad \begin{aligned} w - \lambda x = 0 \\ \lambda \geq 0 \end{aligned}$$

7.9

- Consider the negative entropy of $x \in \mathbb{R}^D$

$$f(x) = \sum_{d=1}^{D} x_d \log(x_d)$$

- Derive the conjugate function $f^*(s)$ by assuming the standard dot product
- First we form the convex conjugate

$$f^*(s) = \langle s, x \rangle - f(x)$$

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

- Now we differentiate with respect to a single x value, x_d and set it equal to zero:

$$\frac{\partial}{\partial x_d} = s_d - \log(x_d) + 1 = 0$$

- Now, we solve for x_d:

$$s_d + 1 = \log(x_d)$$

$$x_d = \exp(s_d + 1)$$

- Now we plug this value back into the conjugate function $f^*(s)$:

$$\begin{aligned} f^*(s) &= \sum_{d=1}^{D} s_d \exp(s_d + 1) - \sum_{d=1}^{D} \exp(s_d + 1) \log(\exp(s_d + 1)) \\ &= \sum_{d=1}^{D} s_d \exp(s_d + 1) - \sum_{d=1}^{D} \exp(s_d + 1)(s_d + 1) \\ &= \sum_{d=1}^{D} s_d \exp(s_d + 1) - \sum_{d=1}^{D} s_d \exp(s_d + 1) + \sum_{d=1}^{D} \exp(s_d + 1) \\ &= \sum_{d=1}^{D} \exp(s_d + 1) \end{aligned}$$

7.10

- Consider the function:

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c$$

- Where A is positive definite, which means it's invertible
- Derive the convex conjugate of $f(x)$

$$f^*(x) = \langle s, x \rangle - f(x)$$

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

- Differentiating with respect to x and setting equal to zero:

$$\frac{\partial f^*}{\partial x} = s - Ax - b = 0$$

- Solving for x:

$$x = A^{-1}(s - b)$$

- Plugging this value for x back into the conjugate:

$$\begin{aligned}
f^*(s) &= s^T(A^{-1}(s-b)) - \frac{1}{2}(A^{-1}(s-b))^T A(A^{-1}(s-b)) - b^T(A^{-1}(s-b)) - c \\
&= s^T A^{-1}(s-b) - \frac{1}{2}(s-b)A^{-1}(s-b) - b^T A^{-1}(s-b) - c \\
&= s^T A^{-1}s - s^T A^{-1}b - b^T A^{-1}s - b^T A^{-1}b - \frac{1}{2}(s-b)A^{-1}(s-b) - c \\
&= s^T A^{-1}s - 2s^T A^{-1}b - b^T A^{-1}b - \frac{1}{2}(s-b)A^{-1}(s-b) - c \\
&= (s-b)A^{-1}(s-b) - \frac{1}{2}(s-b)A^{-1}(s-b) - c \\
&= \frac{1}{2}(s-b)A^{-1}(s-b) - c
\end{aligned}$$

7.11

- The hinge loss is given by:

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

- Compute the convex conjugate of the hinge loss $L^*(\beta)$
- Add a l_2 proximal term and compute the conjugate of the resulting function:

$$L^*(\beta) + \frac{\gamma}{2}\beta^2$$

- Where γ is the given hyperparameter