1 Conjugate functions \cite{Sivaraman, 15 points}

Many thanks to Yuandong Tian whose solution we have used for this part.

Derive the conjugates of the following functions (3 points each):

(a) Max function. \( f(x) = \max_{i=1, \ldots, n} x_i \) on \( \mathbb{R}^n \).

By taking \( x_1 = x_2 = \ldots = x_n = x \), we have

\[
    f^*(\vec{y}) = \sup_{\vec{x}} (\vec{y}^T \vec{x} - f(\vec{x})) \geq x \sum_i y_i - x \tag{1}
\]

If \( \sum_i y_i \neq 1 \), then by taking \( x \) to be the same sign as \( \sum_i y_i - 1 \), we can make \( x \sum_i y_i - x \) arbitrarily large and thus \( f^*(\vec{y}) = +\infty \).

For \( \vec{1}^T \vec{y} = 1 \), we have:

\[
    f^*(\vec{y}) = \sup_{\vec{x}} (\vec{y}^T \vec{x} - f(\vec{x})) = \sup_{\vec{x}} (\sum_i y_i(x_i - x_{[1]})) \tag{2}
\]

where \( x_{[1]} \) is the largest element of \( \vec{x} \). Since \( x_i \leq x_{[1]} \), it is easy to see that if any \( y_i < 0 \), then \( f^*(\vec{y}) = +\infty \). Otherwise \( y_i(x_i - x_{[1]}) \leq 0 \) and \( f^*(\vec{y}) = 0 \) (0 is attainable by setting \( x_1 = x_2 = \ldots = x_n \)).

Summarizing, we have

\[
    f^*(\vec{y}) = \begin{cases} 
        0 & \vec{y} \geq 0, \vec{1}^T \vec{y} = 1 \\
        +\infty & \text{o/w}
    \end{cases} \tag{3}
\]

(b) Sum of largest elements. \( f(x) = \sum_{i=1}^r x_{[i]} \) on \( \mathbb{R}^n \). (where \( x_{[i]} \) denotes the \( i \)-th largest element in \( x \)).

\[\star \text{ SOLUTION:}\] Step 1: Prove \( \vec{1}^T \vec{y} = r \) for finite \( f^*(\vec{y}) \). By taking \( x_1 = x_2 = \ldots = x_n = x \), we have

\[
    f^*(\vec{y}) = \sup_{\vec{x}} (\vec{y}^T \vec{x} - f(\vec{x})) \geq x \sum_i y_i - rx \tag{4}
\]

If \( \sum_i y_i \neq r \), then by taking \( x \) to be the same sign as \( \sum_i y_i - r \), we can make \( x \sum_i y_i - rx \) arbitrarily large and thus \( f^*(\vec{y}) = +\infty \).

Step 2: Prove \( 0 \leq \vec{y} \leq \vec{1} \) for finite \( f^*(\vec{y}) \). If any \( y_i > 1 \), then we take \( x_i \) arbitrarily large and other \( x_j \)'s to be zero, \( f^*(\vec{y}) = +\infty \). If \( \vec{y} \) contains any negative entries \( y_i < 0 \), we can set \( x_i \) arbitrarily negative and others \( 0 \), then \( f(\vec{x}) = 0 \) and \( \vec{y}^T \vec{x} \) can be arbitrarily large. As a result \( f^*(\vec{y}) = +\infty \).

For \( \vec{1}^T \vec{y} = r \) and \( \vec{y} \leq \vec{1} \) we have:

\[
    f^*(\vec{y}) = \sup_{\vec{x}} (\vec{y}^T \vec{x} - f(\vec{x})) = \sup_{\vec{x}} \left( \sum_{i=1}^r (y_{[i]} - 1)x_{[i]} + \sum_{i=r+1}^n y_{[i]}x_{[i]} \right) \tag{5}
\]
Since $0 \leq y_i \leq 1$ and $x_i \geq x_r \geq x_j$ for all $1 \leq i \leq r < j$, we have $(y_i - 1)x_i \leq (y_i - 1)x_r$ and $y_jx_j \leq y_jx_r$. Therefore,

$$f^*(\vec{y}) = \sup_{\vec{x}} \left[ \sum_{i=1}^{r} (y_i - 1)x_i + \sum_{j=r+1}^{n} y_jx_j \right]$$

(6)

$$\leq \sup_{\vec{x}} \left[ x_r \left( \sum_{i=1}^{r} y_i - r \right) + x_r \sum_{j=r+1}^{n} y_j \right]$$

(7)

$$\leq \sup_{\vec{x}} \left[ x_r(r - r) \right]$$

(8)

$$= 0$$

(9)

hence $f^*(\vec{y}) \leq 0$. On the other hand 0 is attainable by setting $x_1 = x_2 = \ldots = x_n$.

So finally we have:

$$f^*(\vec{y}) = \begin{cases} 
0 & \vec{y} \geq 0, \vec{y} = \vec{1}, \vec{1}^T \vec{y} = r \\
+\infty & o/w
\end{cases}$$

(10)

(c) Power function. $f(x) = x^p$ on $\mathbb{R}_{++}$, where $p > 1$.

★ SOLUTION: If $y \geq 0$, then taking the derivative of $L(x, y) = yx - x^p$ w.r.t $x$, we get:

$$\frac{\partial L}{\partial x} = y - px^{p-1}$$

(11)

Set it to zero, and we get $x^* = y^{1/(p-1)}/p^{1/(p-1)}$. Note second derivative is $-p(p-1)x^{p-2} < 0$ for $x \in \mathbb{R}_{++}$, thus it is indeed unique minimum. So the dual function $f^*(y)$ is:

$$f^*(y) = yx^* - (x^*)^p = \frac{y^{p/(p-1)}}{p^{1/(p-1)}} - \frac{y^{p/(p-1)}}{p^{p/(p-1)}} = (p - 1) \left( \frac{y}{p} \right)^{p/(p-1)}$$

(12)

If $y < 0$, then both $yx$ and $-x^p$ are monotonously decreasing, and $\sup_{x>0} L(x, y) = 0$. Thus we have

$$f^*(y) = \begin{cases} 
(p - 1) \left( \frac{y}{p} \right)^{p/(p-1)} & y \geq 0 \\
0 & y < 0
\end{cases}$$

(13)

(d) Geometric mean. $f(x) = - (\prod_{i=1}^{n} x_i)^{1/n}$ on $\mathbb{R}_{++}^n$.

★ SOLUTION: If any $y_i > 0$, then setting $x_i$ arbitrarily large and others $\epsilon > 0$ fixed, then $f^*(\vec{y}) = +\infty$ since both $y_i x_i$ and $-f(\vec{x}) = (\prod_{i=1}^{n} x_i)^{1/n}$ are increasing.

For $\vec{y} \leq 0$, setting $\vec{w} = -\vec{y} \geq 0$ and using arithematic-geometric inequality, we have

$$\vec{w}^T \vec{x} \geq n \left( \frac{1}{n} \sum_{i=1}^{n} w_i x_i \right)^{1/n}$$

(14)

$$\geq n \left( \prod_{i=1}^{n} w_i x_i \right)^{1/n}$$

(15)

$$= n \left( \prod_{i=1}^{n} w_i \right)^{1/n} \left( \prod_{i=1}^{n} x_i \right)^{1/n}$$

(16)

$$= H(-\vec{y})(-f(\vec{x}))$$

(17)
Note the equality holds if we have \(x_1 = x_2 = \ldots = x_n\). So if we have \(H(-\bar{y}) = H(\bar{w}) = n (\prod_{i=1}^n w_i)^{1/n} \geq 1\), then \(\bar{g}^T \bar{x} = -\bar{w}^T \bar{x} \leq f(\bar{x})\) (Note \(f(\bar{x})\) is negative) and \(\sup_{\bar{x}} (\bar{g}^T \bar{x} - f(\bar{x})) = 0\) (0 is attainable by setting \(x_1 = x_2 = \ldots = x_n = \epsilon > 0\) but \(\epsilon\) is arbitarily close to zero) Conversely, if we have \(H(-\bar{y}) < 1\), then by setting \(x_1 = x_2 = \ldots = x_n\) arbitrary large we get \(f^*(\bar{g}) = +\infty\).

Thus the dual function is as follows:
\[
f^*(\bar{g}) = \begin{cases} 0 & \bar{y} \leq 0, H(-\bar{y}) = n [\prod_{i=1}^n -y_i]^{1/n} = n [(1)^n \prod_{i=1}^n y_i]^{1/n} \geq 1 \\
+\infty & \text{otherwise}
\end{cases}
\] (18)

(e) Negative generalized logarithm for second-order cone. \(f(x, t) = -\log(t^2 - x^T x)\) on \((x, t) \in \mathbb{R}^n \times \mathbb{R}^{||x||_2} < t\).

\[\star\] **SOLUTION:** Denote the second order cone \(K = \{(x, t)||x||_2 \leq t\}\). Denote the dual variable \(\bar{y} = (\bar{z}, -u)\). If \(u < 0\), then taking \(\bar{x} = \bar{z}\) and \(t = 2||\bar{x}||_2\), we get
\[
f^*(\bar{y}) = \sup_{\bar{x}, t} (-ut + \bar{z}^T \bar{x} + \log(t^2 - \bar{x}^T \bar{x})) \geq -ut + t^2/4 + \log(3/4t^2) \to +\infty
\] (19)

for \(t \to +\infty\) and thus \(-ut \to +\infty\).

If \(u \geq 0\), then the dual function is
\[
f^*(\bar{y}) = \sup_{\bar{x}, t} (-ut + \bar{z}^T \bar{x} + \log(t^2 - \bar{x}^T \bar{x}))
\] (20)

Denote \(c = \sup_{||x||_2 = 1} \bar{z}^T \bar{x} = ||\bar{z}||_2 \geq 0\) and \(\bar{z}\) that achieves this supremum. For any ray \(\bar{x}^t = \lambda \bar{z}\) with \(\lambda \geq 0\), we have:
\[
f^*(\bar{y}) \geq g(\lambda, t) = -ut + \lambda c + \log(t^2 - \lambda^2)
\] (21)

Note \(f^*(\bar{y}) = \sup_{\lambda, t} g(\lambda, t)\).

If \(0 < u < c\), then setting \(\lambda = \beta t\) so that \(c\beta > u\) makes \(g \to +\infty\). If \(0 < c \leq u\), then taking the partial derivative of \(g\) with respect to \(\lambda\) and \(t\), and setting them to zero, we get:
\[
\frac{\partial g}{\partial t} = u + \frac{2t}{t^2 - \lambda^2} = 0
\]
(22)
\[
\frac{\partial g}{\partial \lambda} = c - \frac{2\lambda}{t^2 - \lambda^2} = 0
\]
(23)

Solving the equations, we get two sets of stationary points under the constraints that \(\{t \geq 0, \lambda \geq 0, t \geq \lambda\}\):
\[
t_1 = \lambda_1 = 0 \quad t_2 = \frac{2u}{u^2 - c^2}, \quad \lambda_2 = \frac{2c}{u^2 - c^2}
\] (24)
\[t_1 = \lambda_1 = 0\] leads \(-\infty\); while
\[
g(\lambda_2, t_2; \bar{x}) = -2 + \log \frac{4}{u^2 - c^2}
\] (25)

While the boundary \(t = \lambda\) gives \(-\infty\), boundary \(\lambda = 0\) gives \(-2 + \log \frac{4}{u^2} \leq g(\lambda_2, t_2; \bar{x})\) for optimal \(t\). Thus Eqn. 25 is indeed the supremum and finally
\[
f^*(\bar{y}) = f^*((\bar{z}, -u)) = -2 + \log \frac{4}{u^2 - c^2}
\] (26)

In one word, given \(\bar{y} = (\bar{z}, -u)\), denote \(c = \sup_{||x||_2 = 1} \bar{z}^T \bar{x} = ||\bar{z}||_2\) and the dual function is the following:
\[
f^*(\bar{y}) = f^*((\bar{z}, -u)) = \begin{cases} +\infty & \bar{y} \notin K' \\
-2 + \log \frac{4}{u^2 - ||\bar{z}||_2^2} & \bar{y} \in K'
\end{cases}
\] (27)

where \(K' = \{\bar{y} = (\bar{z}, -u) | u > ||\bar{z}||_2\}\).
2 The Lagrange dual function and conjugate functions [Yi, 5 points]

The Lagrange dual (of an optimization problem) and the conjugate (of the objective function \( f \)) are closely related. This question will show a simple example.

Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{s.t.} & \quad Ax \leq b \\
& \quad Cx = d
\end{align*}
\]

with variables \( x \in \mathbb{R}^n \). Derive the Lagrange dual of this problem, and express it using the conjugate \( f^* \) of function \( f \).

⋆ SOLUTION: This is a very simple question, designed to enhance our understanding of Lagrange dual and function’s dual (conjugate). The answer is exactly eq. (5.10) and eq.(5.11) on the page 221 of Boyd and Vandenberghe’s textbook.

3 Convexity, Strong Duality and KKT Conditions [Yi, 25 points]

3.1 A Convex Problem with Strong Duality Failed

This question is Problem 5.21 from Boyd & Vandenberghe.

Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad e^{-x} \\
\text{s.t.} & \quad x^2/y \leq 0
\end{align*}
\]

with variables \( x \) and \( y \), and domain \( D = \{(x, y) | y > 0\} \).

1. [3 pts] Verify that this is a convex optimization problem and check whether Slater’s condition holds or not.

⋆ SOLUTION: The domain \( D = \{(x, y) | y > 0\} \) is a convex set; the objective function \( f_0(x) = e^{-x} \) is convex since \( f''_0(x) = e^{-x} \geq 0 \); the function \( f_1(x, y) = x^2/y \) is convex because it is actually the perspective function of \( g(x) = x^2 \) and \( g(x) \) is convex (or, we can directly check the Hessian of \( f_1(x, y) \) to verify the convexity of the function).

Since \( x^2/y \leq 0 \) is not affine and there is not a point in relint\( (D) \) to satisfy \( x^2/y < 0 \), the Slater’s condition does not hold.

2. [4 pts] Give the Lagrange dual problem, and find the optimal solution \( \lambda^* \) and optimal value \( d^* \) of the dual problem. What is the optimal duality gap?

⋆ SOLUTION: It’s trivial to see that the primal optimal is \( p^* = 1 \) (when \( x^* = 0 \)). The Lagrangian dual problem can be formulated as

\[
\begin{align*}
\text{maximize} & \quad g(\lambda) \quad \text{s.t.} \quad \lambda \geq 0
\end{align*}
\]

where \( g(\lambda) = \inf_{x \in \mathbb{R}, y \geq 0} \{ e^{-x} + \lambda x^2/y \} \). For any chosen \( \lambda \), we can let \( x \to \infty \) and \( \lambda x^2/y \to 0 \) so that \( g(\lambda) = 0 \). Therefore the dual optimal is \( d^* = 0 \) and the duality gap is 1.
3.2 KKT Conditions for Non-convex Problems

This question is Problem 5.29 from Boyd & Vandenberghe.

Consider this non-convex optimization problem:

\[
\begin{align*}
\min_{x_1, x_2, x_3} & \quad -3x_1^2 + x_2^2 + 2x_3^2 + 2(x_1 + x_2 + x_3) \\
\text{such that} & \quad x_1^2 + x_2^2 + x_3^2 = 1
\end{align*}
\]

We will see that although it is non-convex, strong duality still holds.

1. [3 pts] Derive and state the KKT conditions.

★ SOLUTION: The Lagrangian of this problem is

\[\mathcal{L}(x_1, x_2, x_3, \lambda) = -3x_1^2 + x_2^2 + 2x_3^2 + 2(x_1 + x_2 + x_3) + \lambda(x_1^2 + x_2^2 + x_3^2 - 1)\]

Taking the derivatives with respect to each variable and setting the derivatives to zero gives the KKT conditions:

\[
\begin{align*}
-6x_1 + 2 + 2\lambda x_1 &= 0 \\
2x_2 + 2 + 2\lambda x_2 &= 0 \\
4x_3 + 2 + 2\lambda x_3 &= 0 \\
x_1^2 + x_2^2 + x_3^2 &= 1
\end{align*}
\]

2. [3 pts] Find all solutions \(x, \lambda\) (where \(\lambda\) is the dual variable) which satisfy the KKT conditions. You can use any tool for this question.

★ SOLUTION: We can use the first three KKT conditions above to find expressions for \(x_1, x_2, x_3\) in terms of \(\lambda\):

\[
\begin{align*}
x_1 &= \frac{1}{3 - \lambda} \\
x_2 &= \frac{-1}{1 + \lambda} \\
x_3 &= \frac{-1}{2 + \lambda}
\end{align*}
\]

Plugging these values into the last KKT condition gives:

\[
\frac{1}{(3 - \lambda)^2} + \frac{1}{(1 + \lambda)^2} + \frac{1}{(2 + \lambda)^2} = 1
\]
Solving this (which you can do with e.g. Mathematica), we get four real solutions for $\lambda$, from which we can compute values for $x_1, x_2, x_3$ and the objective:

- $\lambda = -3.14929$, $(x_1, x_2, x_3) = (0.16262, 0.46527, 0.870103)$
  - objective = 4.64728
- $\lambda = 0.223509$, $(x_1, x_2, x_3) = (0.360167, -0.817321, -0.44974)$
  - objective = -1.1304
- $\lambda = 1.89190$, $(x_1, x_2, x_3) = (0.902445, -0.345794, -0.256944)$
  - objective = -5.36549
- $\lambda = 4.03523$, $(x_1, x_2, x_3) = (-0.965973, -0.198601, -0.165694)$
  - objective = -5.36549

3. [3 pts] Which pair $(x, \lambda)$ corresponds to the optimal solution, and what is the optimal value of the objective?

★ SOLUTION: The optimal solution is $(x_1, x_2, x_3, \lambda) = (−0.965973, -0.198601, -0.165694, 4.03523)$, which gives an optimal objective value of $−5.36549$.

Note that multiple solutions satisfy the KKT conditions, but only one is optimal. For non-convex problems, a solution satisfying the KKT conditions is not necessarily optimal. However, if the objective and constraint functions are differentiable, strong duality implies that any optimal point must satisfy the KKT conditions.

### 3.3 KKT & Supporting Hyperplanes for Convex Problems

This question is based on Problem 5.31 from Boyd & Vandenberghe.

Consider this problem where all $f_i$, $i = 0, \ldots, m$, are convex and differentiable:

$$\min_x f_0(x) \quad \text{such that} \quad f_i(x) \leq 0, \; i = 1, \ldots, m \tag{31}$$

Assume $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^m$ satisfy the KKT conditions:

- $f_i(x^*) \leq 0, \; i = 1, \ldots, m$
- $\lambda_i^* \geq 0, \; i = 1, \ldots, m$
- $\lambda_i^* f_i(x^*) = 0, \; i = 1, \ldots, m$
- $\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) = 0$

1. [5 pts] Show that $\nabla f_0(x^*)^T (x - x^*) \geq 0$ for all feasible $x$. (Note that this has a nice geometric interpretation: if $\nabla f_0(x^*) \neq 0$, then $\nabla f_0(x^*)$ defines a supporting hyperplane to the feasible set at $x^*$.)

★ SOLUTION: Suppose $x$ is a feasible point. Multiplying the last KKT condition by $(x - x^*)$ gives

$$\nabla f_0(x^*)^T (x - x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*)^T (x - x^*) = 0$$
Since each $f_i$ is convex, $f_i(a) \geq f_i(b) + \nabla f_i(b)^T(a - b)$ for all $x, y$ in the domain of $f_i$. For $a = x$ and $b = x^*$, this means that $\nabla f_i(x^*)^T(x - x^*) \leq f_i(x) - f_i(x^*)$. Combining this with the above equality gives

$$\nabla f_0(x^*)^T(x - x^*) + \sum_{i=1}^m \lambda_i^* f_i(x) - \lambda_i^* f_i(x^*) \geq 0$$

The third KKT condition reduces this to

$$\nabla f_0(x^*)^T(x - x^*) + \sum_{i=1}^m \lambda_i^* f_i(x) \geq 0$$

Since $x$ is feasible, we know $f_i(x) \leq 0$, so the second KKT condition gives $\lambda_i^* f_i(x) \leq 0$. This lets us reduce the above inequality to

$$\nabla f_0(x^*)^T(x - x^*) \geq 0$$

Since $f_0$ is convex, we know that for all $x, y$ in the domain of $f_0$,

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^T(y - x)$$

So, if $\nabla f_0(x^*)^T(y - x) \geq 0$, then $f_0(y) \geq f_0(x)$. Thus, we have shown that if $x^*$ satisfies the KKT conditions, $f_0(y) \geq f_0(x^*)$ for all feasible $y$; i.e. $x^*$ is an optimal solution.

To get a better idea of this geometric interpretation, you’ll plot the supporting hyperplane to the feasible set at $x^*$ for the following simple problem:

$$\min_x x_1^2 + x_2^2 \text{ such that } 2x_1 - 3(x - 2) - y \leq 0;$$

2. [2 pts] Find the optimal solution $x^*$ for this problem. (You may do this by hand or using e.g. CVX, but either way, show your work or your code.)

★ SOLUTION: The following code in CVX computes an optimal value of $(x_1^*, x_2^*) = (1.1820, 0.5203)$:

```matlab
cvx_begin
  variable x(1);
  variable y(1);
  minimize( x^2+y^2 );
  subject to
    (2*x-3)*(x-2)-y <= 0;
  cvx_end
x1star = x;
x2star = y;
```

3. [2 pts] Plot the feasible region for this problem (with $x_1$ on the horizontal axis and $x_2$ on the vertical axis), as well as the supporting hyperplane for the feasible region defined by $\nabla f_0(x^*)^T(x - x^*) \geq 0$.

★ SOLUTION: Since $\nabla f_0(x^*) = (2x_1^*, 2x_2^*) = (2.3640, 1.0406)$, the supporting hyperplane is defined by:

$$\begin{pmatrix} 2.3640 & 1.0406 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 1.1820 \\ 0.5203 \end{pmatrix} \geq 0$$

i.e. $x_2 \geq 3.2055 - 2.2718x_1$

The feasible region and supporting hyperplane are plotted in Figure 1.
Feasible region (epigraph of parabola) and supporting hyperplane defined by the optimal solution.

4 Optimality conditions for QCQP [Yi, 10 points]

[Ex. 5.26, B&V]. Consider the following Quadratic Constraints Quadratic Programming (QCQP) problem

\[
\begin{align*}
\text{minimize} & \quad x_1^2 + x_2^2 \\
\text{subject to} & \quad (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1 \\
& \quad (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1
\end{align*}
\]

with variables \( x \in \mathbb{R}^2 \).

1. [3 pts] Sketch the feasible set and level sets of the objective. Find the optimal point \( x^* \) and optimal value \( p^* \).

\[\star \text{ SOLUTION:} \]

The level sets of the objective are circles centered at the origin. The feasible set is the intersection of two circles: one with radius 1 centered at \((1,1)\), and the other with radius 1 centered at \((1,-1)\). Clearly the only feasible point (and thus the optimal point) is \( x_1 = 1, x_2 = 0 \). The optimal value is \( p^* = 1 \).

2. [3 pts] Give the KKT conditions. Do there exist Lagrange multipliers \( \lambda_1^* \) and \( \lambda_2^* \) that prove that \( x^* \) is optimal?

\[\star \text{ SOLUTION:} \]

The Lagrange is:

\[
L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + x_2^2 + \lambda_1[(x_1 - 1)^2 + (x_2 - 1)^2 - 1] + \lambda_2[(x_1 - 1)^2 + (x_2 + 1)^2 - 1]
\]
The KKT conditions for optimal solution $x^*_1 = 1, x^*_2 = 0, \lambda^*_1, \lambda^*_2$ are:

$$(x^*_1 - 1)^2 + (x^*_2 - 1)^2 \leq 1 \text{ (primal constraint)}$$

$$(x^*_1 - 1)^2 + (x^*_2 + 1)^2 \leq 1 \text{ (primal constraint)}$$

$$\lambda^*_1 \geq 0 \text{ (dual constraint)}$$

$$\lambda^*_2 \geq 0 \text{ (dual constraint)}$$

$$\lambda^*_1[(x^*_1 - 1)^2 + (x^*_2 - 1)^2 - 1] = 0 \text{ (complementary slackness)}$$

$$\lambda^*_2[(x^*_1 - 1)^2 + (x^*_2 + 1)^2 - 1] = 0 \text{ (complementary slackness)}$$

$$x^*_1 + \lambda^*_1(x^*_1 - 1) + \lambda^*_2(x^*_1 - 1) = 0 \left( \frac{\partial L}{\partial x_1} \text{ at } x^*_1 \text{ is 0} \right)$$

$$x^*_2 + \lambda^*_1(x^*_2 - 1) + \lambda^*_2(x^*_2 + 1) = 0 \left( \frac{\partial L}{\partial x_2} \text{ at } x^*_2 \text{ is 0} \right)$$

Note that the condition $x^*_1 + \lambda^*_1(x^*_1 - 1) + \lambda^*_2(x^*_1 - 1) = 0$ is 1 = 0, which never holds. So, there does not exist $\lambda^*_1, \lambda^*_2$, together with $x^*_1 = 1, x^*_2 = 0$, satisfying the KKT conditions.

3. [4 pts] Derive and solve the Lagrange dual problem. Does strong duality hold?

**SOLUTION:** We already have the Lagrange in previous answer. Setting the derivatives w.r.t. $x_1$ and $x_2$ to zeros, we have:

$$x_1 + \lambda_1(x_1 - 1) + \lambda_2(x_1 - 1) = 0$$

$$x_2 + \lambda_1(x_2 - 1) + \lambda_2(x_2 + 1) = 0$$

which gives:

$$x_1 = \frac{\lambda_1 + \lambda_2}{1 + \lambda_1 + \lambda_2}$$

$$x_2 = \frac{\lambda_1 - \lambda_2}{1 + \lambda_1 + \lambda_2}$$

Plugging them into the Lagrange, the dual problem becomes:

$$\max_{\lambda_1, \lambda_2 \geq 0} \frac{\lambda_1 + \lambda_2}{1 + \lambda_1 + \lambda_2} - \frac{(\lambda_1 - \lambda_2)^2}{1 + \lambda_1 + \lambda_2}$$

Note that $\frac{\lambda_1 + \lambda_2}{1 + \lambda_1 + \lambda_2} - \frac{(\lambda_1 - \lambda_2)^2}{1 + \lambda_1 + \lambda_2} = 1 - \frac{1}{1 + \lambda_1 + \lambda_2} - \frac{(\lambda_1 - \lambda_2)^2}{1 + \lambda_1 + \lambda_2}$, whose supremum is 1 as $\lambda_1 = \lambda_2 = \infty$. So the dual objective value can be infinitely close to 1, but can not attain 1 by any finite $\lambda_1$ and $\lambda_2$. This is consistent with the fact that no solution can satisfies the KKT conditions.

5 Dual of SOCP [Yi, 10 points]

Show that the dual of the SOCP

$$\min f^T x$$

$$s.t. \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m$$

with variables $x \in \mathbb{R}^n$ can be expressed as

$$\max \sum_{i=1}^m (b_i^T u_i + d_i v_i)$$

$$s.t. \sum_{i=1}^m (A_i^T u_i + c_i v_i) + f = 0$$

$$\|u_i\|_2 \leq -v_i, \quad i = 1, \ldots, m,$$
with variables \( u_i \in \mathbb{R}^n, v_i \in \mathbb{R} \), \( i = 1, \ldots, m \). The problem data are \( f \in \mathbb{R}^n, A_i \in \mathbb{R}^{n \times n}, b_i \in \mathbb{R}^{n}, c_i \in \mathbb{R}^n \) and \( d_i \in \mathbb{R}, i = 1, \ldots, m \). Derive the dual in the following two ways.

(a) [5 points] Introduce new variables \( y_i \in \mathbb{R}^{n_i} \) and \( t_i \in \mathbb{R} \) and equalities \( y_i = A_i x + b_i, t_i = c_i^T x + d_i \), and derive the Lagrange dual.

★ SOLUTION: We restate the problem with new variables \( y_i \) and \( t_i \):

\[
\begin{align*}
\min_{x, y, t} f^T x \\
\text{s.t.} \\
y_i &= A_i x + b_i \quad (i \in \mathbb{R}) \\
t_i &= c_i^T x + d_i \\
\|y_i\|_2 &\leq t_i \\
t_i &\geq 0
\end{align*}
\]

Now the Lagrangian is:

\[
L(x, y, t, u, v, \alpha, \beta)_{\alpha_i \geq 0, \beta_i \geq 0} = f^T x + \sum_{i=1}^{m} u_i^T (A_i x + b_i - y_i) + \sum_{i=1}^{m} v_i (c_i^T x + d_i - t_i) + \sum_{i=1}^{m} \alpha_i (\|y_i\|_2 - t_i) + \sum_{i=1}^{m} \beta_i t_i
\]

Use the derivatives of the Lagrangian to get dual constraints.

First, we have \( \frac{\partial L}{\partial x} = f + \sum_{i=1}^{m} (A_i^T u_i + c_i v_i) = 0 \), which gives

\[
\sum_{i=1}^{m} (A_i^T u_i + c_i v_i) + f = 0
\]

Second, we have \( \frac{\partial L}{\partial y_i} = -u_i + \alpha_i \frac{y_i}{\sqrt{y_i^T y_i}} = 0 \) so that \( u_i = \alpha_i \frac{\sqrt{y_i^T y_i}}{y_i} \), which gives

\[
\|u_i\|_2 = \left( \frac{\alpha_i y_i}{\sqrt{y_i^T y_i}} \right)^T \left( \frac{\alpha_i y_i}{\sqrt{y_i^T y_i}} \right) = \alpha_i \sqrt{\frac{y_i^T y_i}{y_i^T y_i}} = \alpha_i
\]

Third, we have \( \frac{\partial L}{\partial t_i} = -v_i - \alpha_i + \beta_i = 0 \) so that \( -v_i + \beta_i = \alpha_i \), which gives:

\[
v_i \geq \alpha_i
\]

Now, plugging into all conditions about \( \frac{\partial L}{\partial x}, \frac{\partial L}{\partial y_i}, \frac{\partial L}{\partial t_i} \), the Lagrangian dual is:

\[
\begin{align*}
\min_{x, y, t} L(x, y, t, u, v, \alpha, \beta) &= \sum_{i=1}^{m} u_i^T b_i + \sum_{i=1}^{m} v_i d_i \\
&= \sum_{i=1}^{m} (u_i^T b_i + v_i d_i)
\end{align*}
\]

Thus we have the following dual problem:

\[
\max_{u, v} \sum_{i=1}^{m} (u_i^T b_i + v_i d_i)
\]
\[ \sum_{i=1}^{m} (A_i^T u_i + c_i v_i) + f = 0 \]
\[ \|u_i\|_2 \leq -v_i \quad \forall i \]

(b) [5 points] Start from the conic formulation of the SOCP and use the conic dual. Use the fact that the second-order cone is self-dual.

★ SOLUTION: As we just did, deriving the dual of second-order cone programs using KKT conditions is not very easy. Let’s consider the conic formulation of the primal problem, and use the fact that second-order cone is self-dual to derive the dual problem.

Let define \( K = \{(x, t) \mid \|x\|_2 \leq t\} \) as the second-order cone. Also, we define

\[ E_i = (A_i \quad c_i^T), \quad f_i = (b_i \quad d_i) \]

Now the primal problem can be rewritten as:

\[ \text{minimize } f^T x \]
\[ \text{s.t. } E_i x + f_i \in K, \quad i = 1, \ldots, m \]

Let’s take \( y_i = [u_i; v_i] \in -K \). Since \( K \) is self-dual, we have \( y_i^T (E_i x + f_i) \leq 0 \), which gives:

\[ (u_i^T A_i + v_i c_i^T)x + (u_i^T b_i + v_i d_i) \leq 0, \quad i = 1, 2, \ldots, m \]

Summing all of them together, we have:

\[ -\sum_{i=1}^{m} (u_i^T A_i + v_i c_i^T)x \geq \sum_{i=1}^{m} (u_i^T b_i + v_i d_i) \]

In order to obtain a lower bound for the primal objective \( f^T x \), we let \(-\sum_{i=1}^{m} (u_i^T A_i + v_i c_i^T) = f^T \), which gives the constraint:

\[ \sum_{i=1}^{m} (A_i^T u_i + c_i v_i) + f = 0 \]

With this constraint, \( \sum_{i=1}^{m} (u_i^T b_i + v_i d_i) \) becomes a lower bound for the objective \( f^T x = -\sum_{i=1}^{m} (u_i^T A_i + v_i c_i^T)x \).

Note that since \( y_i = [u_i; v_i] \in -K \), we also have constraints:

\[ \|u_i\|_2 \leq -v_i, \quad i = 1, 2, \ldots, m \]

Finally, to get the dual problem, we maximize the lower bound \( \sum_{i=1}^{m} (u_i^T b_i + v_i d_i) \) of the objective, subject to all constraints we have:

\[ \text{maximize } \sum_{i=1}^{m} (u_i^T b_i + v_i d_i) \]
\[ \text{s.t. } \sum_{i=1}^{m} (A_i^T u_i + c_i v_i) + f = 0 \]
\[ \|u_i\|_2 \leq -v_i, \quad i = 1, \ldots, m, \]
6 Newton’s method [Sivaraman, 40 points]

Many thanks to Yuandong Tian whose solutions we used for the theory part and Bin Zhao whose figures we used for the implementation part.

In this question you will implement Newton’s method for multi-class logistic regression.

Consider, the expression for K-class logistic regression,

\[
P(y = k | X = x) = \frac{\exp(w_k^T x)}{\sum_{k=1}^{K} \exp(w_k^T x)}
\]

where each \(w_i\) is a p-dimensional vector (where \(p\) is the number of features). More generally you would consider a bias term for each class, but you can ignore the bias term for this problem.

6.1 Theory

1. [2 pts] Consider you are given a training set \((X, Y) = \{(X_1, y_1), \ldots, (X_n, y_n)\}\). Write an expression for the log-likelihood of the training set, and cast the problem of maximizing this log-likelihood with a ridge penalty of the form \(\frac{\lambda}{2} \sum_{k=1}^{K} ||w_k||^2\) on the parameters as an optimization problem.

**SOLUTION:** The log-likelihood for a single training sample \((\vec{x}_i, y_i)\) is:

\[
\log P(y_i | \vec{x}_i) = \sum_{k=1}^{K} I(y_i = k)(w_k^T \vec{x}_i) - \log \sum_{k=1}^{K} \exp(w_k^T \vec{x}_i)
\]

So the log-likelihood for the entire training set is:

\[
\log L = \sum_{i=1}^{n} \log P(y_i | \vec{x}_i) = \sum_{k=1}^{K} w_k^T (\sum_{y_j = k} \vec{x}_j) - \sum_{i=1}^{n} \log \sum_{k=1}^{K} \exp(w_k^T \vec{x}_i)
\]

The optimization problem of maximizing this log-likelihood with a ridge penalty is the follows:

\[
\max \{\vec{w}_k\} \sum_{k=1}^{K} w_k^T (\sum_{y_j = k} \vec{x}_j) - \sum_{i=1}^{n} \log \sum_{k=1}^{K} \exp(w_k^T \vec{x}_i) - \frac{\lambda}{2} \sum_{k=1}^{K} ||w_k||^2
\]

2. [5 pts] Derive analytic expressions for the gradient, and the hessian of the problem.

**SOLUTION:** The analytic expression for the gradient is:

\[
\frac{\partial J}{\partial w_k} = \left(\sum_{y_j = k} \vec{x}_j\right) - \sum_{i=1}^{n} \frac{\exp(w_k^T \vec{x}_i) \vec{x}_i}{\sum_{l=1}^{K} \exp(w_l^T \vec{x}_i)} - \lambda \vec{w}_k
\]

for \(k = 1 \ldots K\). The Hessian matrix is thus given by:

\[
\frac{\partial^2 J}{\partial w_k \partial w_l} = -\sum_{i=1}^{n} \left[ \frac{\exp(w_k^T \vec{x}_i) \vec{x}_i \vec{x}_i^T}{\sum_{l=1}^{K} \exp(w_l^T \vec{x}_i)} - \frac{\exp(w_k^T \vec{x}_i) \exp(w_l^T \vec{x}_i) \vec{x}_i \vec{x}_i^T}{\left(\sum_{l=1}^{K} \exp(w_l^T \vec{x}_i)\right)^2} \right] - \lambda I_p
\]

\[
\frac{\partial^2 J}{\partial w_k \partial w_l} = \sum_{i=1}^{n} \frac{\exp(w_k^T \vec{x}_i) \exp(w_l^T \vec{x}_i) \vec{x}_i \vec{x}_i^T}{\left(\sum_{l=1}^{K} \exp(w_l^T \vec{x}_i)\right)^2} \quad l \neq k
\]
Here $I_p$ is a $p$ by $p$ identity matrix. Denote

$$c_{ik} = \frac{\exp(\mathbf{w}_k^T \mathbf{x}_i)}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_i)}$$  \hspace{1cm} (40)$$

Note $\sum_{k=1}^{K} c_{ik} = 1$. We can simplify the gradient and the Hessian to be the following:

$$g_k = \frac{\partial J}{\partial \mathbf{w}_k} = \left( \sum_{y_j=k} \mathbf{x}_j \right) - \sum_{i=1}^{n} c_{ik} \mathbf{x}_i - \lambda \mathbf{w}_k$$  \hspace{1cm} (41)$$

and

$$H_{kk} = \frac{\partial^2 J}{\partial \mathbf{w}_k \partial \mathbf{w}_k} = - \sum_{i=1}^{n} c_{ik}(1 - c_{ik}) \mathbf{x}_i \mathbf{x}_i^T - \lambda I_p$$  \hspace{1cm} (42)$$

$$H_{kl} = \frac{\partial^2 J}{\partial \mathbf{w}_k \partial \mathbf{w}_l} = \sum_{i=1}^{n} c_{ik} c_{il} \mathbf{x}_i \mathbf{x}_i^T \quad k \neq l$$  \hspace{1cm} (43)$$

Note that there are $Kp$ unknowns in $\{\mathbf{w}_k\}$, and each entry, for example, $H_{kk}$ or $H_{kl}$, is a $p \times p$ submatrix.

3. [3 pts] Show analytically that the hessian you derived is negative semidefinite. Also, show that if the weight on the ridge term $\lambda > 0$, then the objective is “strictly” concave. A function $f$ is strictly concave if $f(x) < f(x_0) + g(x_0)^T (x - x_0)$ for all $x$ in $\text{dom}(f)$ where $g$ is the gradient (or an element of the sub-gradient) of $f$.

Hint: A diagonally dominant symmetric matrix with real entries with positive diagonal entries is positive semidefinite. You can additionally use the fact that the Kronecker product of two, PSD matrices is PSD. Show that the Hessian can be written as the Kronecker product of a PSD matrix and a diagonally dominant one. Are the diagonal entries of the diagonally dominant matrix positive or negative? What can you conclude?

★ SOLUTION: First let us consider the following matrix $M^{(i)}$:

$$M_{kk}^{(i)} = -c_{ik}(1 - c_{ik})$$  \hspace{1cm} (44)$$

$$M_{kl}^{(i)} = c_{ik} c_{il} \quad k \neq l$$  \hspace{1cm} (45)$$

Then $M^{(i)}$ is diagonally dominant since (note $c_{ik} \geq 0$ and $\sum_{i=1}^{K} c_{il} = 1$):

$$\sum_{l \neq k} |M_{kl}^{(i)}| = c_{ik} \sum_{l \neq k} c_{il} = c_{ik}(1 - c_{ik}) \leq |M_{kk}^{(i)}|$$  \hspace{1cm} (46)$$

And $M_{kk}^{(i)} \leq 0$, so the matrix $M^{(i)}$ is negative semidefinite. Thus the Hessian

$$H = \sum_{i=1}^{n} M^{(i)} \otimes \mathbf{x}_i \mathbf{x}_i^T - \lambda I_{Kp}$$  \hspace{1cm} (47)$$

is also negative semidefinite, since

• For any $i$, $\mathbf{x}_i \mathbf{x}_i^T$ is a positive semidefinite matrix.

• For $A$ negative semidefinite and $B$ positive semidefinite, $A \otimes B$ is also negative semidefinite, where $\otimes$ is the Kronecker product.

• $-\lambda I_{Kp}$ is negative semidefinite.
In the case of $\lambda > 0$, $H$ is negative definite matrix. According to the Taylor theorem, the objective $J$ can be written as the following:

$$J(\vec{w}) = J(\vec{w}_0) + \nabla g(\vec{w}_0)^T(\vec{w} - \vec{w}_0) + \frac{1}{2}(\vec{w} - \vec{w}_0)^T H(\vec{x}_i)(\vec{w} - \vec{w}_0)$$  \hspace{1cm} (48)$$

where $\vec{x}_i$ lies on the line segment connecting $\vec{w}$ and $\vec{w}_0$ in the $K_p$-dimensional space. Since $H$ is negative definite matrix in the entire $\text{dom}(f)$, we thus have $(\vec{w} - \vec{w}_0)^T H(\vec{x}_i)(\vec{w} - \vec{w}_0) < 0$ for $\vec{w} \neq \vec{w}_0$, and

$$J(\vec{w}) < J(\vec{w}_0) + \nabla g(\vec{w}_0)^T (\vec{w} - \vec{w}_0)$$  \hspace{1cm} (49)$$

6.2 Implementation

In this part of the question you will work with two data sets, the first is a synthetic dataset and the second is the UCI Iris dataset.

1. [10 pts] Implement Newton’s method for this problem. Your implementation should accept a value for $\lambda$, however don’t use values of $\lambda$ that are too small, or your hessian might not be invertible. Declare convergence whenever the L2-norm of the parameter vector (or matrix) changes by less than $10^{-6}$.

★ SOLUTION: See the file “newton_lr.m”

2. [7 pts] You should now download the synthetic dataset (“data_large.mat”) from the website. Train your logistic regressor on the training set, and use the learned weights to predict class labels for the test set. In your report you should show the weights you learned (you can print these out since this is a big vector or matrix but clearly indicate the class and feature for the weights). Report your accuracy on the test set. Use $\lambda = 0.01$ for this part.

★ SOLUTION: The accuracy is 0.792. The learned $\vec{w}$ is shown in Tbl. 1

3. [6 pts] Download “data_small.mat” from the website. Plot regularization paths for the weights (you should have only 4 weights), by varying $\lambda$ (values of $\{10^{-2}, 1, 10^2, 10^4, 10^6, 10^8, 10^{10}\}$ work well for me but you can use any values as long as you get a reasonable curve).

★ SOLUTION: Figure shows the regularization paths for the weights, where we set $\lambda$ to $\{10^{-2}, 1, 10^2, 10^4, 10^6, 10^8, 10^{10}\}$.

4. [7 pts] Download the UCI Iris dataset from the website (we have split the data for you into train, validation and test sets). Use the validation set to pick the best value of $\lambda$. In your report, clearly write down the value of $\lambda$ and your accuracy on the training, test and validation sets.

★ SOLUTION: Select $\lambda$ according to validation set accuracy and pick $\lambda = 1$. Accuracy on training set is 94.44%, on validation set is 100% and on test set is 96.67%.
<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Class 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7302</td>
<td>0.57296</td>
<td>-0.20573</td>
<td>-1.3408</td>
<td>-0.75662</td>
</tr>
<tr>
<td>-0.019672</td>
<td>-0.018552</td>
<td>0.021217</td>
<td>-0.01101</td>
<td>0.028017</td>
</tr>
<tr>
<td>0.33158</td>
<td>-0.5067</td>
<td>0.99726</td>
<td>1.4225</td>
<td>-2.2446</td>
</tr>
<tr>
<td>0.41914</td>
<td>-1.2377</td>
<td>0.55752</td>
<td>0.50817</td>
<td>-0.24716</td>
</tr>
<tr>
<td>-0.62499</td>
<td>0.48578</td>
<td>-0.05546</td>
<td>-0.16151</td>
<td>0.95617</td>
</tr>
<tr>
<td>-0.82304</td>
<td>-0.14361</td>
<td>0.66324</td>
<td>0.74045</td>
<td>-0.43704</td>
</tr>
<tr>
<td>0.0053212</td>
<td>-0.04786</td>
<td>0.10052</td>
<td>-0.15148</td>
<td>0.093121</td>
</tr>
<tr>
<td>-0.62524</td>
<td>1.8507</td>
<td>0.24449</td>
<td>-0.81728</td>
<td>-0.6527</td>
</tr>
<tr>
<td>0.019489</td>
<td>0.086413</td>
<td>-0.097003</td>
<td>0.039228</td>
<td>-0.048128</td>
</tr>
<tr>
<td>-0.23775</td>
<td>-1.9771</td>
<td>2.5328</td>
<td>-0.57894</td>
<td>0.26098</td>
</tr>
<tr>
<td>-0.6853</td>
<td>-0.72641</td>
<td>1.4177</td>
<td>-0.02333</td>
<td>0.017337</td>
</tr>
<tr>
<td>-0.64222</td>
<td>0.72766</td>
<td>-0.45098</td>
<td>-0.40163</td>
<td>0.76717</td>
</tr>
<tr>
<td>-1.2687</td>
<td>0.72086</td>
<td>-1.0369</td>
<td>1.3662</td>
<td>0.21858</td>
</tr>
<tr>
<td>-0.61435</td>
<td>0.38945</td>
<td>0.46138</td>
<td>1.4823</td>
<td>-1.7188</td>
</tr>
<tr>
<td>1.8635</td>
<td>-0.16089</td>
<td>-0.54633</td>
<td>-0.58598</td>
<td>-0.57026</td>
</tr>
<tr>
<td>0.0048084</td>
<td>-0.00021944</td>
<td>-0.073244</td>
<td>-0.0011321</td>
<td>0.069787</td>
</tr>
<tr>
<td>0.66707</td>
<td>0.42088</td>
<td>-1.628</td>
<td>-2.0273</td>
<td>2.5674</td>
</tr>
<tr>
<td>0.62611</td>
<td>0.39901</td>
<td>0.96616</td>
<td>-0.50918</td>
<td>-1.4821</td>
</tr>
<tr>
<td>-0.038506</td>
<td>0.024211</td>
<td>-0.052342</td>
<td>0.047536</td>
<td>0.019101</td>
</tr>
<tr>
<td>0.51283</td>
<td>-1.4555</td>
<td>1.0302</td>
<td>0.20898</td>
<td>-0.29651</td>
</tr>
</tbody>
</table>

Table 1: The learned $\mathbf{w}$ of Problem 6.2.2. Rows are feature 1 to feature 15, Columns are 5 classes.

![Figure 2: Regularization paths for the weights.](image-url)