EE227BT Discussion Section #1

Exercise 1 (Quadratics And Least Squares) Consider the two dimensional quadratic function, $f : \mathbb{R}^2 \to \mathbb{R}$ given by:

$$f(w) = w^\top Aw - 2b^\top w + c$$

where $A \in \mathbb{S}^2_+$, $b \in \mathbb{R}^2$ and $c \in \mathbb{R}$.

1. Explain why the function f is convex.
2. Assume $c = 0$. Give a concrete example of a matrix $A \succ 0$ and a vector b such that the point $w^* = \begin{bmatrix}-1 & 1\end{bmatrix}^\top$ is the unique minimizer of the quadratic function $f(w)$.
3. Assume $c = 0$. Give a concrete example of a matrix $A \succeq 0$, and a vector b such that the quadratic function $f(w)$ has infinitely many minimizers and all of them lie on the line $w_1 + w_2 = 0$.
4. Assume $c = 0$. Give a concrete example of a **non-zero** matrix $A \succeq 0$ and a vector b such that the quadratic function $f(w)$ tends to $-\infty$ as we follow the direction defined by the vector $\begin{bmatrix}1 & 0\end{bmatrix}^\top$.
5. Say that we have the data set $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ of features $x^{(i)} \in \mathbb{R}^2$ and values $y^{(i)} \in \mathbb{R}$. Define $X = \begin{bmatrix}x^{(1)} & \dots & x^{(n)}\end{bmatrix}^\top$ and $y = \begin{bmatrix}y^{(1)} & \dots & y^{(n)}\end{bmatrix}^\top$. In terms of X and y, find a matrix A, a vector b and a scalar c, so that we can express the sum of the square losses $\sum_{i=1}^{n}(w^\top x^{(i)} - y^{(i)})^2$ as the quadratic function $f(w) = w^\top Aw - 2b^\top w + c$.
6. Which of the following can be true for the minimization of the sum of the square losses of part (5):
   (a) It can have a unique minimizer.
   (b) It can have infinitely many minimizers, all of them lying on a single line.
   (c) It can be unbounded from below, i.e. there is some direction so that if we follow this direction the loss tends asymptotically to $-\infty$.

Solution 1

1. Consider any line $u + tv$, parametrized by $t \in \mathbb{R}$. Let $g(t)$ be the restriction of f on the line, i.e. $g(t) \doteq f(u + tv)$. Then

   $$g(t) = (v^\top Av)t^2 - 2(b^\top v - u^\top Av)t + (c + u^\top Au - 2b^\top u)$$

   which is a convex univariate quadratic, since $v^\top Av \geq 0$.
2. $A = I, b = w^*$.

3. $A = ee^\top, b = 0$, where $e = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$.
4. $A = e_2 e_2^\top, b = e_1$, where $e_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$ and $e_2 = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$.
5. $A = X^\top X, b = X^\top y, c = y^\top y$.
6. (a) This will be the case when $X^\top X \succ 0$.
   (b) This will be the case when $\lambda_{\min}(X^\top X) = 0$.
   (c) This can not happen since $\sum_{i=1}^{n} (w^\top x^{(i)} - y^{(i)})^2 = \|Xw - y\|_2^2 \geq 0$.

Exercise 2 (Solving Least Squares with CVX) 1. Use the standard normal distribution in order to generate a random 16×8 matrix X, and a random 16×1 vector y. Then use CVX in order to solve the least squares problem:

$$\min_{w \in \mathbb{R}^8} \; \|Xw - y\|_2^2$$

Check your answer by comparing with the analytic least squares solution.

2. Now assume that we are interested in finding a binary valued vector w for the least squares problem, i.e. we would like to solve

$$p^* = \min_{w \in \mathbb{R}^8} \; \|Xw - y\|_2^2 \; : \; w_i \in \{0, 1\}, \; i = 1, \dots, 8$$

Note that this problem is not convex, but we can form the following convex relaxation

$$p^*_{\text{int}} = \min_{w \in \mathbb{R}^8} \; \|Xw - y\|_2^2 \; : \; 0 \leq w_i \leq 1, \; i = 1, \dots, 8$$

Use CVX to find p^*_{int}. What is the relation between p^* and p^*_{int}?

3. Finally use CVX to solve the LASSO problem

$$\min_{w \in \mathbb{R}^8} \; \|Xw - y\|_2^2 + \lambda \|w\|_1$$

where $\lambda > 0$ is a hyper-parameter. Use values of λ in the interval $[10^{-4}, 10^6]$, and create a plot of each coordinate w_i of the optimal vector w versus the corresponding hyper-parameter λ.

Solution 2

cvx_leastsq

August 29, 2017

In [1]:
```
import cvxpy as cvx
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
# Random Instance Generation
n = 16
d = 8
X = np.random.rand(n,d)
y = np.random.rand(n)
```

In [3]:
```
# Least Squares
w = cvx.Variable(d)
objective = cvx.Minimize(cvx.sum_entries(cvx.square(X*w - y)))
prob = cvx.Problem(objective)
print("Optimal value", prob.solve())
print("Optimal var")
print(w.value)

K = np.dot(X.T, X)
detK = np.linalg.det(K)
wopt = np.linalg.solve(K, np.dot(X.T, y))
print("Optimal var using normal equations")
print(wopt)
```

```
Optimal value 0.7269248580530685
Optimal var
[[-0.03744191]
 [ 0.29492623]
 [-0.03587683]
 [ 0.09939876]
 [-0.25274529]
 [ 0.66684273]
 [ 0.19989569]
 [ 0.04706958]]
Optimal var using normal equations
[-0.03744375  0.29492624 -0.03587498  0.09939864 -0.25274482  0.6668402
  0.1998968   0.04707078]
```

```
In [4]: # Interval Constrained Least Squares
        w = cvx.Variable(d)
        objective = cvx.Minimize(cvx.sum_entries(cvx.square(X*w - y)))
        constraints = [0 <= w, w <= 1]
        prob = cvx.Problem(objective, constraints)
        print("Optimal value", prob.solve())
        print("Optimal var")
        print(w.value)
```

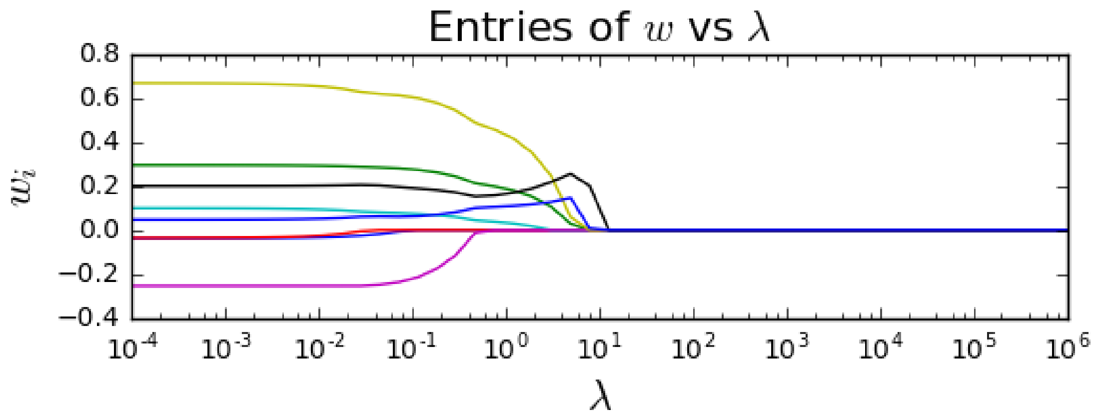
```
Optimal value 0.8530015312527397
Optimal var
[[  1.41279015e-10]
 [  2.28935370e-01]
 [  1.42273570e-10]
 [  5.17538320e-02]
 [  3.25442467e-11]
 [  5.24910648e-01]
 [  1.37646559e-01]
 [  9.76480218e-02]]
```

```
In [5]: # LASSO
        lam = cvx.Parameter(sign="positive")
        w = cvx.Variable(d)
        error = cvx.sum_squares(X*w - y)
        obj = cvx.Minimize(error + lam*cvx.norm(w, 1))
        prob = cvx.Problem(obj)

        sq_penalty = []
        l1_penalty = []
        w_values = []
        lam_vals = np.logspace(-4, 6)
        for val in lam_vals:
            lam.value = val
            prob.solve()
            sq_penalty.append(error.value)
            l1_penalty.append(cvx.norm(w, 1).value)
            w_values.append(w.value)

        # Plot entries of w vs lam
        plt.subplot(212)
        for i in range(d):
            plt.plot(lam_vals, [wi[i,0] for wi in w_values])
        plt.xlabel(r'$\lambda$', fontsize=16)
        plt.ylabel(r'$w_{i}$', fontsize=16)
        plt.xscale('log')
        plt.title(r'Entries of $w$ vs $\lambda$', fontsize=16)
```

```
plt.tight_layout()
plt.show()
```



Exercise 3 (A Simple Case Of LASSO) Say that we have the data set $\{(x^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ of features $x^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$. Define $X = \begin{bmatrix} x^{(1)} & \dots & x^{(n)} \end{bmatrix}^\top$ and $y = \begin{bmatrix} y^{(1)} & \dots & y^{(n)} \end{bmatrix}^\top$. For the sake of simplicity, assume that the data has been centered and whitened so that each feature has mean 0 and variance 1 and the features are uncorrelated, i.e. $X^\top X = nI$.

Consider the linear least squares regression with regularization in the ℓ_1-norm, also known as LASSO:

$$w^* = \arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_1$$

1. Decompose this optimization problem in d univariate optimization problems.
2. If $w_i^* > 0$, then what is the value of w_i^*?
3. If $w_i^* < 0$, then what is the value of w_i^*?
4. What is the condition for w_i^* to be 0?
5. Now consider the case of ridge regression, which uses the the ℓ_2 regularization $\lambda \|w\|_2^2$.

   $$w^* = \arg\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

   What is the new condition for w_i^* to be 0? How does this differ from the condition obtained in part (4)? What does this suggest about LASSO?

Solution 3 1.

$$\|Xw - y\|_2^2 + \lambda \|w\|_1 = \sum_{i=1}^{d} \left[n w_i^2 - 2y^\top x_i w_i + \lambda |w_i| \right] + y^\top y$$

where $X = \begin{bmatrix} x_1 & \dots & x_d \end{bmatrix}$.

2. If $w_i^* > 0$, then the first order optimality conditions for w_i^* write

   $$2nw_i^* - 2y^\top x_i + \lambda = 0$$

   from which we obtain

   $$w_i^* = \frac{2y^\top x_i - \lambda}{2n}$$

   which is positive when

   $$y^\top x_i > \frac{\lambda}{2}$$

3. If $w_i^* < 0$, then the first order optimality conditions for w_i^* write

   $$2nw_i^* - 2y^\top x_i - \lambda = 0$$

   from which we obtain

   $$w_i^* = \frac{2y^\top x_i + \lambda}{2n}$$

   which is negative when

   $$y^\top x_i < -\frac{\lambda}{2}$$

4. From the previous parts $w_i^* = 0$, when $|y^\top x_i| \le \frac{\lambda}{2}$.
5. In the case of ridge regression the optimal weight vector w is given by

$$w_i^* = \frac{y^\top x_i}{n + \lambda},\ i = 1, \ldots, d$$

So the coordinate i is only zero when $y^\top x_i = 0$, in contrast to LASSO where the coordinate i is zero when $y^\top x_i \in \left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]$. This suggest that LASSO forces a lot of coordinates to be zero, i.e. induces sparsity to the optimal weight vector.