098311 Optimization 1 Spring 2018
HW 8

Chen Tessler 305052680
Orr Krupnik 302629027

June 7, 2018

Solution 1.

$\Rightarrow$ Direction 1: x^* is an optimal solution to the problem $\min\{f(x) : x \in C\}$. Therefore, $f(x^*) \leq f(x)$ for any $x \in C$. Since f is convex over C, it obeys the gradient inequality. Using the gradient inequality and the optimality of x^* we have:

$$\nabla f(x)^T(x^* - x) \leq f(x^*) - f(x) \leq 0, \, \forall x \in C$$
$$\Rightarrow \langle \nabla f(x), x^* - x \rangle \leq 0, \, \forall x \in C$$

$\Leftarrow$ Direction 2: Assume x^* is not an optimal solution of the optimization problem above. Then, there exists some $x^{**} \in C$ such that $f(x^{**}) < f(x^*)$. Now, let us define a new function $g(\lambda) = f(\lambda x^* + (1-\lambda)x^{**})$. Since f is convex, g is defined for any $\lambda \in [0,1]$. Additionally, since f is convex and $f(x^{**}) < f(x^*)$, $g(\lambda)$ is non-decreasing. In particular, there exists some $\tilde{\lambda} \in [0,1]$ such that $g'(\tilde{\lambda}) > 0$. Now, by definition:

$$0 < g'(\tilde{\lambda}) = \lim_{\epsilon \to 0} \frac{g(\tilde{\lambda}+\epsilon) - g(\tilde{\lambda})}{\epsilon} = \lim_{\epsilon \to 0} \frac{f((\tilde{\lambda}+\epsilon)x^* + (1-\tilde{\lambda}-\epsilon)x^{**}) - f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**})}{\epsilon} =$$
$$= \lim_{\epsilon \to 0} \frac{f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**} + \epsilon(x^* - x^{**})) - f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**})}{\epsilon} =$$
$$= \lim_{\theta \to 0} \frac{1}{1-\tilde{\lambda}} \frac{f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**} + \theta(x^* - (\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**})) - f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**})}{\theta} =$$
$$= \frac{1}{1-\tilde{\lambda}} \underbrace{\nabla f(\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**})^T(x^* - (\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**}))}_{\leq 0}$$

Since $\frac{1}{1-\tilde{\lambda}} > 0$, and since C is convex (and therefore $\tilde{\lambda}x^* + (1-\tilde{\lambda})x^{**} \in C$) this contradicts the fact that $\langle \nabla f(x), x^* - x \rangle \leq 0, \, \forall x \in C$. Hence, x^* must be an optimal solution to the problem $\min\{f(x) : x \in C\}$.

Solution 2.

$\Rightarrow$ Direction 1: Assume x^* is a stationary point of $(P1)$. Then, we have by the condition for stationarity:

$$\nabla f(x^*)^T(x - x^*) \geq 0, \, \forall x \in \Delta_n$$
$$\iff \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*)(x_i - x_i^*) \geq 0$$

Since this holds true for any $x \in \Delta_n$, in particular it holds true for $e_j, \; j \in \{1, ..., n\}$. Therefore, we have:

$$\forall j : \frac{\partial}{\partial x_j} f(x^*)(1 - x_j^*) \geq \sum_{i \neq j} \frac{\partial}{\partial x_i} f(x^*) x_i^*$$
$$\iff \frac{\partial}{\partial x_j} f(x^*) \geq \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i^* = \underbrace{\sum_{k: x_k^* > 0}^{n} \frac{\partial}{\partial x_k} f(x^*) x_k^*}_{(a)} \geq \min_{k: x^* k > 0} \frac{\partial}{\partial x_k} f(x^*) \; (1)$$

The inequality holds for all $j \in \{1, ..., n\}$, and particularly for $j \in argmin_{k: x_k^* > 0} \frac{\partial}{\partial x_k} f(x^*)$. Since the weighted average (a) is both greater or equal than and smaller or equal than its minimal value, it must equal the minimal value, meaning:

$$\forall j, k \in \{i : x_i^* > 0\} : \frac{\partial}{\partial x_k} f(x^*) = \frac{\partial}{\partial x_j} f(x^*) \triangleq \mu$$

Inequality (1) above holds for $j \notin \{k : x_k^* > 0\}$ as well, therefore for any such j (using $\sum_{k: x^* k > 0} x_k^* = 1$):

$$\frac{\partial}{\partial x_j} f(x^*) \geq \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i^* = \sum_{k: x_k^* > 0}^{n} \frac{\partial}{\partial x_k} f(x^*) x_k^* = \mu \sum_{k: x_k^* > 0}^{n} x_k^* = \mu$$

We have shown, then, that for a stationary point x^* we have $\frac{\partial}{\partial x_i} f(x^*) = \mu$ for some $\mu \in \mathbb{R}$ for $x_i^* > 0$ and $\frac{\partial}{\partial x_i} f(x^*) \geq \mu$ for $x_i^* = 0$.

$\Leftarrow$ Direction 2: Assume there exsists some $\mu \in \mathbb{R}$ such that $\frac{\partial}{\partial x_i} f(x^*) = \begin{cases} = \mu & x_i^* > 0 \\ \geq \mu & x_i^* = 0 \end{cases}$.

Then:

$$\nabla f(x^*)^T(x - x^*) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*)(x_i - x_i^*) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i - \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i^* =$$
$$= \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i - \sum_{k: x_k^* > 0} \frac{\partial}{\partial x_k} f(x^*) x_k^* = \underbrace{\sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x^*) x_i}_{\geq \mu} - \mu \underbrace{\sum_{k: x_k^* > 0} x_k^*}_{=1}$$

where the first sum is greater or equal μ since it is a weighted sum (with the weights summing to 1 since $x \in \Delta_n$) of the partial derivatives of $f(x^*)$, all of which are greater or equal μ, and the second sum is equal 1 since $x^* \in \Delta_n$. Concluding, the last part of the inequality above is greater or equal zero, which taking the whole inequality gives us the stationarity condition for x^* (since it holds for any $x \in \Delta_n$).

Solution 3.

$\Rightarrow$ Direction 1: Assume x^* satisfying $a^T x^* = 1$ is a stationary point of $(P2)$. We shall show that $\frac{\frac{\partial f}{\partial x_i}(x^*)}{a_i} = \frac{\frac{\partial f}{\partial x_j}(x^*)}{a_j}$ for any $i, j \in \{1, ..., n\}$.

Assume that for some $i \neq j$: $\frac{1}{a_i}\frac{\partial f}{\partial x_i} > \frac{1}{a_j}\frac{\partial f}{\partial x_j}$. Define x as follows (for some $\alpha > 0$:

$$x_k = \begin{cases} x_k^* & k \notin \{i, j\} \\ x_k^* - \frac{\alpha}{a_k} & k = i \\ x_k^* + \frac{\alpha}{a_k} & k = j \end{cases}$$

Note:

$$a^T x = \sum_{k \neq i,j} a_k x_k^* + a_i(x_i^* - \frac{\alpha}{a_i}) + a_j(x_j^* + \frac{\alpha}{a_j}) = \sum_i a_i x_i^* = a^T x^* = 1$$

Therefore, x satisfies the constraints and therefore should satisfy the stationarity condition with x^*, however:

$$\begin{aligned} \nabla f(x^*)^T(x - x^*) &= \sum_{k=1}^n \frac{\partial f}{\partial x_k}(x^*)(x_k - x_k^*) \\ &= \sum_{k \neq i,j} \frac{\partial f}{\partial x_k}(x^*)(x_k - x_k^*) + \frac{\partial f}{\partial x_i}(x^*)(x_i - x_i^*) + \frac{\partial f}{\partial x_j}(x^*)(x_j - x_j^*) \\ &= \sum_{k \neq i,j} \frac{\partial f}{\partial x_k}(x^*)(x_k^* - x_k^*) - \frac{\partial f}{\partial x_i}(x^*)\frac{\alpha}{a_i} + \frac{\partial f}{\partial x_j}(x^*)\frac{\alpha}{a_j} \\ &= -\frac{\partial f}{\partial x_i}(x^*)\frac{\alpha}{a_i} + \frac{\partial f}{\partial x_j}(x^*)\frac{\alpha}{a_j} = \alpha\left(\frac{1}{a_j}\frac{\partial f}{\partial x_j}(x^*) - \frac{1}{a_i}\frac{\partial f}{\partial x_i}(x^*)\right) < 0 \end{aligned}$$

which contradicts the stationarity assumption of x^*. Thus we conclude that $\frac{1}{a_i}\frac{\partial f}{\partial x_i}(x^*) = \frac{1}{a_j}\frac{\partial f}{\partial x_j}(x^*)\ \ \forall i, j$.

$\Leftarrow$ Direction 2: Assume $\frac{1}{a_j}\frac{\partial f}{\partial x_j}(x^*) = \frac{1}{a_i}\frac{\partial f}{\partial x_i}(x^*)$ for some x^* satisfying $a^T x^* = 1$ and all $i, j \in \{1, ..., n\}$. We show x^* is a stationary point of $(P2)$, or $\nabla f(x^*)^T(x - x^*) \geq 0$ for any x satisfying $a^T x = 1$.

Denote $\mu \triangleq \frac{1}{a_j}\frac{\partial f}{\partial x_j}(x^*) = \frac{1}{a_i}\frac{\partial f}{\partial x_i}(x^*)$, and then:

$$\begin{aligned} \nabla f(x^*)^T(x - x^*) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x^*)(x_i - x_i^*) = \sum_{i=1}^n \mu a_i(x_i - x_i^*) = \\ = \mu \sum_{i=1}^n a_i x_i - \mu \sum_{i=1}^n a_i x_i^* = \mu(a^T x - a^T x^*) = 0 \geq 0 \end{aligned}$$

therefore, x^* obeys the stationarity condition with any x satisfying $a^T x = 1$, and therefore it is a stationary point of $(P2)$.

Solution 4. The optimization problem $max\{x^T y : ||y||_1 \leq k, -e \leq y \leq e\}$ is equivalent to the linear program:

$$\begin{aligned} \max \;& x^T y \\ \text{s.t. } & ||y||_1 \leq k \\ & ||y||_\infty \leq 1 \end{aligned}$$

First, we show the constraint set is convex. For any $y^{(1)}, y^{(2)}$ in the constraint set and for any $\lambda \in [0,1]$:

$$\begin{aligned} ||\lambda y^{(1)} + (1-\lambda) y^{(2)}||_1 &= \sum_{i=1}^{n} |\lambda y_i^{(1)} + (1-\lambda) y_i^{(2)}| \leq \sum_{i=1}^{n} \lambda |y_i^{(1)}| + (1-\lambda)|y_i^{(2)}| = \\ &= \lambda ||y^{(1)}||_1 + (1-\lambda)||y^{(2)}||_1 \leq \max_{j=1,2} ||y^{(j)}||_1 \leq k \\ \forall i \in \{1, ..., n\} : \; & |\lambda y_i^{(1)} + (1-\lambda) y_i^{(2)}| \leq \lambda |y_i^{(1)}| + (1-\lambda)|y_i^{(2)}| \leq \max_{j=1,2} |y_i^{(j)}| \leq 1 \\ & \Rightarrow ||\lambda y^{(1)} + (1-\lambda) y^{(2)}||_\infty \leq 1 \end{aligned}$$

hence $\lambda y^{(1)} + (1-\lambda) y^{(2)}$ is in the constraint set, and it is therefore convex.
Now, since this is a linear program, with a maximization of a linear function over a convex set, at least one optimal solution is an extreme point (as we've seen in lecture 8 slide 5).
Now, let us find the extreme points of the constraint set.

Lemma. The set of extreme points of the set $C = \{y : ||y||_1 \leq k, ||y||_\infty \leq 1\}$ is the set $Z = \{z \in \{-1, 1, 0\}^n, ||z||_1 = k\}$.

Proof. Assume that some point $y \in C \cap Z^c$ (Z^c is the complement of Z) is an extreme point of C. We assume that y has $t \leq n$ non-zero elements. By definition of C, $\sum_i |y_i| \leq k$ and $|y_i| \leq 1 \;\; \forall i$.

Denote the set I as all the indices of y in which $0 < |y_i| < 1$. Additionally, denote the set J as all the indices of y in which $|y_i| = 1$.

We define $g = \sum_{j \in J} e_j \cdot \text{sign}(y_j) \in C$ (a vector in which all indices are zeros except those in which $|y_i| = 1$ where $g_i = \text{sign}(y_i)$), and $\forall i \in I \;\; y^i = e_i \cdot \text{sign}(y_i) + g \in C$ (y^i contains ± 1 at any index in which $|y_i| = 1$ in addition to the index i). Any y^i can be represented by two vectors $z^1, z^2 \in Z$ such that: $z_j^1 = z_j^2 = \text{sign}(y_i) \;\; \forall j \in J \cup I$, $z_j^1 = -z_j^2 \;\; \forall j \notin I \cup J$. Since C is a convex set, $y^i = \frac{z^1 + z^2}{2}$ is also in C.

Finally, since y can be represented as a convex composition of t points in C, specifically, $y^1, ..., y^t$:

$$y = \lambda_1 y^1 + (1-\lambda_1)\left(\lambda_2 y^2 + (1-\lambda_2)\left(\lambda_3 y^3 ...\right)\right)$$

where $\lambda_i \Pi_{j=1}^{i-1}(1-\lambda_j) = y_i$, this entails (by definition) that y is not an extreme point of C. This holds true for any $y \in C \cap Z^c$.

We now show that any $z \in Z$ can't be represented as a convex combination of two distinct points in C. For the convex combination of any two points $y^1 \neq y^2 \in C$ to equal z, the following is required:

$$z = \lambda y^1 + (1-\lambda) y^2, \;\; \lambda \in (0,1)$$

however, as z has exactly k indices equal to ± 1 and by definition of C, any point $y \in C$ satisfies $|y_i| \leq 1$ and $\sum_i |y_i| \leq k$. Assume on the contrary that z can indeed be represented as a convex combination, then at least y^1 or y^2 have one index i such that $|z_i| = 1$ and either $y_i^1 \neq z_i$ or $y_i^2 \neq z_i$ and thus by definition for any $\lambda \in (0,1)$: $\lambda y_i^1 + (1-\lambda) y_i^2 \neq z_i$ and hence z_i can't be represented as a convex combination of two non-equal points in C. As such, any point $z \in Z$ is by definition an extreme point of C.

We conclude that as any point $y \in C \cap Z^c$ is not an extreme point and any point $z \in Z$ is an extreme point - then the set Z defines the set of extreme points of C.

$\square$

Since at least one optimal solution to the linear optimization problem is an extreme point, let us look at the extreme points of the constraint set. Such an extreme point has k elements which are ± 1, while the rest are 0. Denoting this set of k active indices as I_k, we have:

$$\max_{||y||_1 \leq k, ||y||_\infty \leq 1} x^T y = \max_{||y||_1 \leq k, ||y||_\infty \leq 1} \sum_{i=1}^{n} x_i y_i = \max_{||y||_1 \leq k, ||y||_\infty \leq 1} \sum_{i \in I_k} x_i y_i \leq \max_{||y||_1 \leq k, ||y||_\infty \leq 1} \sum_{i \in I_k} |x_i||y_i| =$$
$$= \max_{I_k \subseteq \{1,\ldots,n\}} \sum_{i \in I_k} |x_i| \leq \sum_{i \in 1..k} |x_{\langle i \rangle}| = f(x) \;\; \blacksquare$$

Solution 5. The following CVX based code solves the max-margin problem (based on the code given in the problem statement):

```
cvx_begin
    variable w(2);
    variable b;
minimize(pow_pos(norm(w, 2), 2) / 2)
subject to
    A1 * w + b <= -1 * ones(length(A1), 1);
    A2 * w + b >= ones(length(A2), 1);
cvx_end
```

The produced solution (w, β) is $w = (-43.960, 19.441)^T$ and $\beta = 12.417$. Figure 1 below shows the resulting separating line. Code is attached as an m file to this submission.

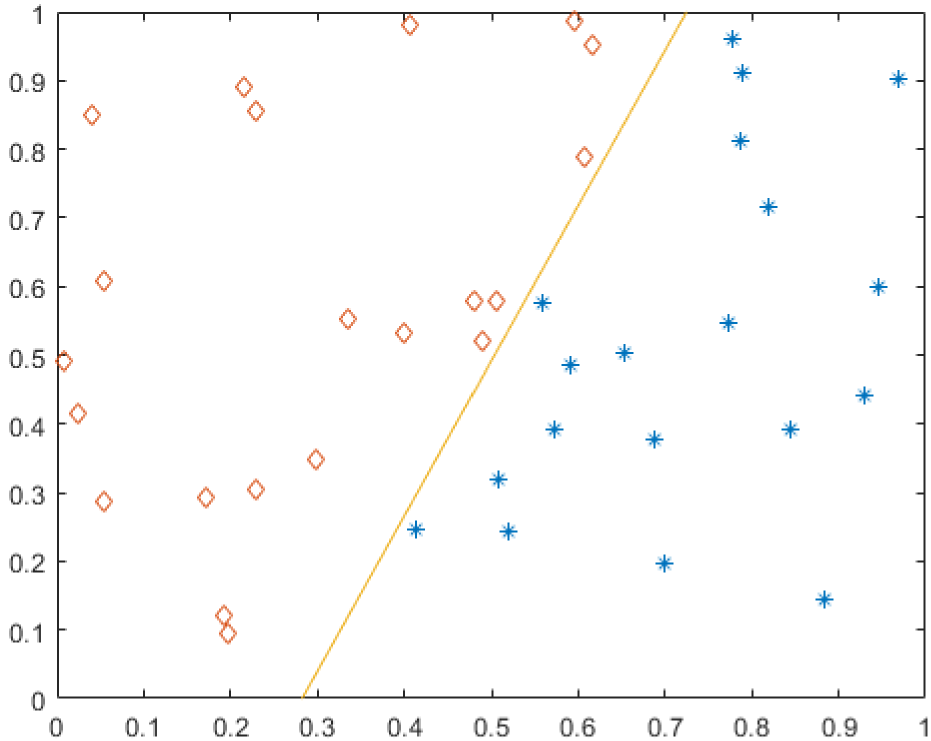

Figure 1: The generated points and resulting max-margin separator line