Homework 3

1.1

- 2.1) We prove by induction:

  - Base case, $K=1$ we get that $\theta_1 x_1 \in C$ since $\theta_1 = 1$ and $x_1 \in C$. ✓
  - Assume that $\sum_{i=1}^{K} \theta_i x_i \in C$ with $\theta_i \geq 0$, $\sum_{i=1}^{K} \theta_i = 1$.
  - Take $\sum_{i=1}^{K+1} \theta_i x_i$ with $\sum_{i=1}^{K+1} \theta_i = 1$, $\theta_i \geq 0$.

Then $\sum_{i=1}^{K+1} \theta_i x_i = \sum_{i=1}^{K} \theta_i x_i + \theta_{K+1} x_{K+1}$

(1) $$= (1-\theta_{K+1}) \sum_{i=1}^{K} \frac{\theta_i}{(1-\theta_{K+1})} x_i + \theta_{K+1} x_{K+1}$$

Now $\sum_{i=1}^{K} \frac{\theta_i}{1-\theta_{K+1}} = 1$ since $\sum_{i=1}^{K} \theta_i = 1 - \theta_{K+1}$

and thus by induction hypothesis $\sum_{i=1}^{K} \frac{\theta_i}{1-\theta_{K+1}} x_i \in C$.

Say $\sum_{i=1}^{K} \frac{\theta_i}{1+\theta_{K+1}} = x^* \in C$.

Then (1) reduces to $(1-\theta_{K+1}) x^* + \theta_{K+1} x_{K+1}$ which is in C by definition of convexity.

Thus $\sum_{i=1}^{K+1} \theta_i x_i \in C$ $\square$

2.5) The distance between the two hyperplanes will be perpendicular to both hyperplanes.

Then take a point x_1 in the first hyperplane: $a^T x_1 = b_1$

Now from x_1 we go in a direction perpendicular until we hit x_2 in the second hyperplane. So $x_2 = x_1 + at$ for some $t \in \mathbb{R}$.

Plugging into the second hyperplane we get $a^T(x_1 + at) = b_2$.

$$a^T x_1 + a^T a t = b_2, \qquad t = \frac{b_2 - a^T x_1}{a^T a} = \frac{b_2 - b_1}{a^T a}$$

So $x_2 = x_1 + \frac{b_2 - b_1}{a^T a} a$ and thus the distance is

$$\|x_2 - x_1\| = \left\| \frac{b_2 - b_1}{a^T a} a \right\| = \frac{|b_2 - b_1|}{\|a\|^2} \|a\| = \frac{|b_2 - b_1|}{\|a\|} \qquad \square$$

2.12) (a) This is the intersection of two (convex) half-spaces:

$$a^T x \leq \beta \quad \text{and} \quad a^T x \geq \alpha$$

And thus is convex

(b) This is the intersection of $2n$ half-spaces:

$$x_i \leq \beta_i \quad \text{and} \quad x_i \geq \alpha_i \quad \text{for every } i$$

And thus is convex

(c) This is the intersection of two halfspaces:

$$a_1^T x \leq b_1 \quad \text{and} \quad a_2^T x \leq b_2$$

And thus is convex

(d) First note that the set of points closer to a point than another point is a half-space:
- Geometrically take the perpendicular bisector of the two points and it's one side of it
- Algebraically note that:

$$\|x - x_0\| \leq \|x - x_1\|$$

$$\|x - x_0\|^2 \leq \|x - x_1\|^2$$

$$(x - x_0)^T (x - x_0) \leq (x - x_1)^T (x - x_1)$$

$$\cancel{x^T x} - 2x_0^T x + x_0^T x_0 \leq \cancel{x^T x} - 2x_1^T x + x_1^T x_1$$

$$2(x_1 - x_0)^T x \leq x_1^T x_1 - x_0^T x_0$$

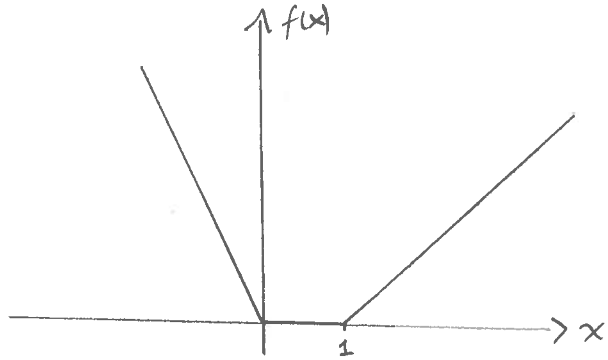
is a half-space (with $a = 2(x_1 - x_0)$, $b = x_1^T x_1 - x_0^T x_0$)

Thus the set of points closer to a given point than a set is an ifinite intersection of half-spaces and thus is convex:

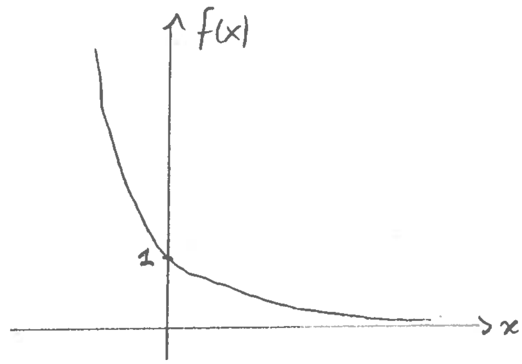
$$\bigcap_{y \in S} \{x \mid \|x - x_0\| \leq \|x - y\|\}$$

2.1

(i) Take $f(x) = \max(-x, 0, x-1)$, then as seen in the graph below f has a unique minimum value at zero but multiple minimizers. Any $x \in [0, 1]$ is a minimizer:



(ii) Take $f(x) = e^{-x}$ for $x \in \mathbb{R}$. Then f is convex since $f''(x) = e^{-x} > 0$ for all x, but f has no minimum value since we can make $f(x) = e^{-x}$ as close to zero as desired but never zero:



2.2

(i) For every y, we have that:

$$\frac{\partial L(h_w(x),y)}{\partial x} = L'(h_w(x),y)\cdot w$$

$$\frac{\partial^2 L(h_w(x),y)}{\partial x^2} = L''(h_w(x),y)\cdot w^2$$

using the property:

Also note that since $h_w(x) = w\cdot x$ is an affine mapping and $L(h,y)$ is convex, then $L(h_w(x),y)$ is convex by properties of convexity

Since $L(h,y)$ is convex then $L''(h_w(x),y) \geq 0 \quad \forall x,y$ and thus $L''(h_w(x),y)\cdot w^2 \geq 0 \quad \forall x,y$ so $L(h_w(x),y)$ is convex.

(ii) $\hat{L}(w) = \frac{1}{M}\sum_{i=1}^{M} L(h_w(x_i),y_i) = \sum_{i=1}^{M} \frac{1}{M} L(h_w(x_i),y_i)$ is a weighted sum of convex functions and thus is convex.

3.1

To reduce the optimality gap by a factor of 10 we need:

(for $C_f = 3$)

$$c^k \leq \frac{1}{10} \Rightarrow \left(1-\frac{m}{M}\right)^k \leq \frac{1}{10}$$

$$\left(\frac{2}{3}\right)^k \leq \frac{1}{10} \quad \text{(since the condition number } M/m = 3 \text{ so } \frac{m}{M} = \frac{1}{3}\text{)}$$

$$k \geq \log_{2/3}\left(\frac{1}{10}\right)$$

$$k \geq 5.6$$

So we need 6 iterations.

If the condition number is 100 instead, we get:

$$\left(1-\frac{1}{100}\right)^k \leq \frac{1}{10}$$

$$\left(\frac{99}{100}\right)^k \leq \frac{1}{10}$$

$$k \geq \log_{99/100}\left(\frac{1}{10}\right)$$

$$k \geq 229.1$$

So we need 230 iterations

4.1

① $\hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} (w - y_i)^2/2$

$\hat{L}'(w) = 1/m \sum_{i=1}^{m} (w - y_i)$

setting $\hat{L}'(w)$ to zero we get: $\sum_{i=1}^{m} (w^* - y_i) = 0$

$$m w^* = \sum_{i=1}^{m} y_i$$

$$w^* = 1/m \sum_{i=1}^{m} y_i = \bar{y}$$

② From above $\hat{L}'(w) = 1/m \sum_{i=1}^{m} (w - y_i)$

$$= w - 1/m \sum_{i=1}^{m} y_i$$

$$= w - \bar{y}$$

③ With learning rate $h=1$, initial $w = w_0$ we get that:

$$w_1 = w_0 - \hat{L}'(w_0) = w_0 - (w_0 - \bar{y}) = \bar{y}$$

so $w_1 = \bar{y} = w^*$ and GD converges in one step.

④ With learning rate $h = 1/2$ and initial $w = w_0$ we get that:

$$w_1 = w_0 - 1/2\, \hat{L}'(w_0) = w_0 - 1/2(w_0 - \bar{y}) = 1/2\, w_0 + 1/2\, \bar{y}$$

(1) $\Rightarrow w_1 - \bar{y} = 1/2\,(w_0 - \bar{y})$

And $w_{k+1} = w_k - 1/2\, \hat{L}'(w_k) = w_k - 1/2(w_k - \bar{y}) = 1/2\, w_k + 1/2\, \bar{y}$

(2) $\Rightarrow w_{k+1} - \bar{y} = 1/2\,(w_k - \bar{y})$

From (1) and (2) by induction we get that $(w_k - \bar{y}) = (1/2)^k (w_0 - \bar{y})$

(which implies $|w_k - \bar{y}| \leq (1/2)^k |w_0 - \bar{y}|$)

4.2

$$m\hat{L}'(\omega_0) = 0 - 1 - 1 - 1 - 1 - 1 = -5$$

$6\hat{L}'(-3) = -5$ (note that in this case $M = 6$)

$\hat{L}'(-3) = -5/6$

Then $\omega_1 = \omega_0 - 6\hat{L}'(\omega_0) = -3 - 6\left(\frac{-5}{6}\right) = 2$ (1 iteration)

$$m\hat{L}'(\omega_1) = +1 + 1 + 1 + 1 + 0.5 - 1 = 3.5$$

$6L'(2) = 3.5$

$L'(2) = 7/12$

Then $\omega_2 = \omega_1 - 6\hat{L}'(\omega_1) = 2 - 6(7/12) = -3/2$ (2 iterations)

$$m\hat{L}'(\omega_2) = +1 + 0.5 - 1 - 1 - 1 - 1 = -2.5$$

$6\hat{L}'(-3/2) = -5/2$

$\hat{L}'(-3/2) = -5/12$

Then $\omega_3 = \omega_2 - 6\hat{L}'(\omega_2) = -3/2 - 6(-5/12) = 1$ (3 iterations)

4.3

$$\hat{L}(w) = 1/M \sum_{i=1}^{M} (w \cdot x_i - w^* \cdot x_i)^2/2$$

$$\frac{\partial \hat{L}(w)}{\partial w_j} = 1/M \sum_{i=1}^{M} (w \cdot x_i - w^* \cdot x_i)(x_i)_j$$

$$= 1/M \sum_{i=1}^{M} (w - w^*) \cdot x_i (x_i)_j$$

Hence $\nabla \hat{L}(w) = 1/M \sum_{i=1}^{M} (w - w^*) \cdot x_i \, x_i$

To express this as a matrix equation $\nabla \hat{L}(w) = H(w - w^*)$ we see that the j^{th} component of $H(\omega-\omega^*)$ is the j^{th} row of H dotted with $(\omega-\omega^*)$.

Comparing this with $\nabla\hat{L}(\omega) = \frac{1}{M}\sum_{i=1}^{M}(\omega-\omega^*)\cdot x_i\, x_i$ we see that the j^{th} component of $\nabla\hat{L}(\omega)$ is $\frac{1}{M}\sum_{i=1}^{M}(\omega-\omega^*)\cdot x_i (x_i)_j$

$$= \frac{1}{M}(\omega-\omega^*)\cdot\sum_{i=1}^{M} x_i (x_i)_j$$

Thus the j^{th} row of H must be $\frac{1}{M}\sum_{i=1}^{M}(x_i)_j\, x_i^T$ $(\star)$ but this is the j^{th} row of $\frac{1}{M}X^TX$ so $H = \frac{1}{M}X^TX$

From $\star$ we can also see that if the l^{th} row of H is $\frac{1}{M}\sum_{i=1}^{M}(x_i)_l\, x_i^T$ then $H_{kl} = \frac{1}{M}\sum_{i=1}^{M}(x_i)_l(x_i)_k = \frac{1}{M}\sum_{i=1}^{M}(x_i)_k(x_i)_l$

For $M=4$ and $x_i = (1, i)$ we have that:

$$H = \tfrac{1}{4}X^TX = \tfrac{1}{4}\begin{pmatrix}1&1&1&1\\1&2&3&4\end{pmatrix}\begin{pmatrix}1&1\\1&2\\1&3\\1&4\end{pmatrix}$$

$$H = \tfrac{1}{4}\begin{pmatrix}4&10\\10&30\end{pmatrix}$$

$\square$