CSE 251A: Homework 3 Solutions

1. *Checking convexity/concavity.*

   (a) $f(x) = e^{ax}$ is convex.
   Proof: The second partial derivative $H(x) = f''(x) = a^2 e^{ax} \geq 0$ for all x and a.

   (b) $f(x) = |x|$ is convex.
   Proof: $\forall a, b \in \mathbb{R}$ and $\theta \in (0, 1)$,

   $$f(\theta a + (1-\theta)b) = |\theta a + (1-\theta)b| \leq |\theta a| + |(1-\theta)b| = \theta|a| + (1-\theta)|b| = \theta f(a) + (1-\theta)f(b)$$

   (c) $f(x) = \ln x$ is concave.
   Proof: $-f(x) = -\ln x$ is convex because the second derivative

   $$H(x) = -f''(x) = \frac{1}{x^2} \geq 0$$

   (d) $f(x) = x^a \quad (x > 0)$. Here we only consider $x > 0$ because $f(x)$ is not always defined when x is negative. $f(x)$ is convex when $a \geq 1$ and $a \leq 0$, and is concave when $0 < a < 1$.
   Proof: The second derivative

   $$H(x) = a(a-1)x^{a-2}$$

   When $0 < a < 1$, $H(x) < 0$, which means the second derivative of $-f(x)$ is positive, so in this case $f(x)$ is concave. When $a \geq 1$ or $a \leq 0$, $H(x) \geq 0$, so in this case $f(x)$ is convex.

2. *Showing convexity.*

   (a) The Hessian of $f(x) = x^T M x$ is $H(x) = 2M$. Since M is positive semidefinite, so is $2M$; so f is convex.

   (b) The Hessian of $f(x) = e^{u^\top x}$ is

   $$H(x) = e^{u^\top x} u u^\top,$$

   which can also be written as $vv^\top$, where $v = (e^{u^\top x}/2)u$. For any vector w, $w^\top(vv^\top)w = (w^\top v)^2 \geq 0$ and hence the Hessian $H(x)$ is PSD. This implies that $f(x)$ is convex.

   (c) We prove this from first principles using our basic definition of convexity. Since $f(x) = \max(f_1(x), \ldots, f_k(x))$, where the individual f_i are all convex, we have that for all $x_1, x_2 \in \mathbb{R}$ and $t \in (0, 1)$,

   $$\begin{aligned}
   & f(tx_1 + (1-t)x_2) \\
   = \;& \max\left(f_1(tx_1 + (1-t)x_2), f_2(tx_1 + (1-t)x_2), \ldots, f_k(tx_1 + (1-t)x_2)\right) \\
   \leq \;& \max\left(tf_1(x_1) + (1-t)f_1(x_2), tf_2(x_1) + (1-t)f_2(x_2), \ldots, tf_k(x_1) + (1-t)f_k(x_2)\right) \\
   \leq \;& t\max\left(f_1(x_1), f_2(x_1), \ldots, f_k(x_1)\right) + (1-t)\max\left(f_1(x_2), f_2(x_2), \ldots, f_k(x_2)\right) \\
   = \;& tf(x_1) + (1-t)f(x_2)
   \end{aligned}$$

   Therefore, $f(x)$ is convex.

3. *Entropy.* The negation of the entropy, $N(p) = -H(p)$, has Hessian with entries

   $$\frac{\partial N}{\partial p_i \partial p_j} = \begin{cases} 0 & \text{if } i \neq j, \\ \frac{1}{p_i \ln 2} & \text{if } i = j \end{cases}$$

   This is a diagonal matrix with positive values on the diagonal. Thus the Hessian is P.S.D., whereupon N is convex and H is concave.

4. *Regression problem.*

   (a) Let

$$X = \begin{pmatrix} \leftarrow & x^{(1)} & \rightarrow \\ \leftarrow & x^{(2)} & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & x^{(n)} & \rightarrow \end{pmatrix}$$

   Then we can write the Hessian as

$$H(w) = 2\sum_{i=1}^{n} x^{(i)} \left(x^{(i)}\right)^T + 2\lambda I = 2X^T X + 2\lambda I$$

   (b) For all $z \in \mathbb{R}^d$

$$z^T H z = z^T(2X^T X + 2\lambda I)z = 2(z^T X^T X z + \lambda z^T I z) = 2||Xz||^2 + 2\lambda||z||^2 \geq 0$$

   Therefore, $H(w)$ is P.S.D, which means $L(w)$ is convex.

5. *Convex sets.*

   (a) The circle is not a convex set: for any two points on the circle, the line joining them does not lie on the circle.

   (b) The ball is convex.

   (c) Hyperplanes are convex.

   (d) k-sparse points are not convex: lines joining two such points can be upto $(2k)$-sparse.

   (e) The set of positive semidefinite matrices is closed under addition and multiplication by positive scalars; therefore it is convex.

6. *Norms.*

   (a) We can check that ℓ_1 is a norm by going through the definition, one property at a time:

      i. $\|x\|_1 = \sum_{i=1}^d |x_i| \geq 0$.

      ii. If $x = 0$, then $\|x\|_1 = 0$. If $\exists i, x_i \neq 0$, then $\|x\|_1 \geq |x_i| > 0$. Therefore, $\|x\|_1 = 0$ if and only if $x = 0$.

      iii. For any real-valued t, we have $\|tx\|_1 = \sum_{i=1}^d |tx_i| = |t| \sum_{i=1}^d |x_i| = |t| \, \|x\|_1$

      iv. $\|x + y\|_1 = \sum_{i=1}^d |x_i + y_i| \leq \sum_{i=1}^d |x_i| + |y_i| = \sum_{i=1}^d |x_i| + \sum_{i=1}^d |y_i| = \|x\|_1 + \|y\|_1$

   (b) Invoking homogeneity and the triangle inequality, we have that for any norm f,

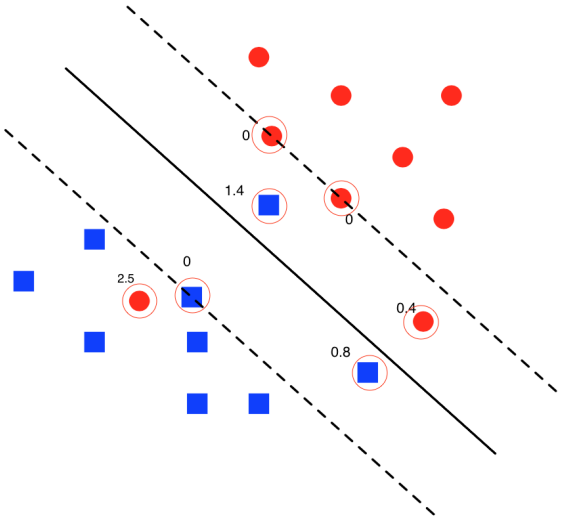
$$f(\theta x + (1-\theta)y) \leq f(\theta x) + f((1-\theta)y) = |\theta| f(x) + |1-\theta| f(y) = \theta f(x) + (1-\theta) f(y).$$

   Thus any norm is a convex function.

   (c) Various inequalities relating $\|x\|_1$, $\|x\|$, and $\|x\|_\infty$:

      i. $\|x\|_1 = \sqrt{(\sum_{i=1}^d |x_i|)^2} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d |x_i||x_j|} \geq \sqrt{\sum_{i=1}^d x_i^2} = \|x\|$.
      $\|x\| = \sqrt{\sum_{i=1}^d x_i^2} \geq \sqrt{\max_i x_i^2} = \max_i |x_i| = \|x\|_\infty$

      ii. Let vector $a = (|x_1|, |x_2|, \dots, |x_d|), b = (1, 1, \dots, 1)_d$
      $\|x\|_1 = \sum_{i=1}^d |x_i| = |a \cdot b| \leq \|a\| \, \|b\| = \sqrt{\sum_{i=1}^d x_i^2} \sqrt{\sum_{i=1}^d 1^2} = \|x\| \cdot \sqrt{d}$.
      $\|x\| = \sqrt{\sum_{i=1}^d x_i^2} \leq \sqrt{d \cdot \max_i x_i^2} = \|x\|_\infty \cdot \sqrt{d}$.
      Therefore, $\|x\|_1 \leq \|x\| \cdot \sqrt{d} \leq \|x\|_\infty \cdot d$.



(d) The unit ball $\{x : x^T Ax \leq 1\}$ is an ellipsoid.

7. *A lower bound for the perceptron.* Pick any $\gamma > 0$. We will now try to construct a dataset where perceptron must use $\frac{1}{\gamma^2}$ updates. Consider the following data set in $\mathbb{R}^d$, where $d = 1/\gamma^2$:

   - There are d points, each corresponding to one coordinate direction: $e_1, e_2, \ldots, e_d$, where e_i is the vector with all zeros except for a 1 at position i.
   - All points have label $+1$.

   These points are correctly classified by the vector $w^* = (\gamma, \gamma, \ldots, \gamma)$, which has unit length and has margin $\min_i(w^* \cdot e_i) = \gamma$.

   Now suppose the perceptron algorithm is run on this data set, and that it produces a linear separator w. If perceptron does not update on e_i, then $w_i = 0$ and w will not correctly classify e_i. Therefore, there must be at least one update for every data point: a total of $1/\gamma^2$ updates.

8. *Support vectors.* See Figure for the pictorial solution. Observe that any point within the two dashed lines, as well as any point whose label is predicted incorrectly is a support vector and has a non-zero dual value. Without loss of generality suppose that the blue side corresponds to the positive side – that is, the side where $w^\top x + b \geq 0$.

   Any blue point on the positive side of the dashed line on the blue side and any red point on the negative side of the dashed line on the red side will have slack variable value zero – as in these cases $y^{(i)}(w^\top x^{(i)} + b) > 1$. Additionally, a blue point $(x^{(i)}, y^{(i)})$ that lies on the dashed line on the blue side satisfies $w^\top x^{(i)} + b = 1$ which implies that the corresponding slack variable is 0. The same applies to red points that lie on the dashed line on the red side.

   A point lying on the decision boundary satisfies $y^{(i)}(w^\top x^{(i)} + b) = 0$ – hence, the slack variable value is 1. Thus, a red point $x^{(i)}$ lying on the red side but in between the decision boundary and the dashed line will have a slack variable value between 0 and 1; we visually estimate the value from the figure. The same applies to blue points.

   Finally, a blue point that lies on the negative side of the decision boundary has $y^{(i)}(w^\top x^{(i)} + b) < 0$, and so its slack variable value is more than 1. If this point lies between the decision boundary and the dashed line on the negative side, then $y^{(i)}(w^\top x^{(i)} + b) > -1$ and hence the slack variable value is between 1 and

2. If it is to the negative side of the dashed line on the red side, then $y^{(i)}(w^\top x^{(i)} + b) < -1$ and hence the slack variable is more than 2. We use these principles and visually estimate the values in the figure.

The margin decreases if the factor C is increased.

9. (a) To solve by substitution, we use $x = 4/y$ and solve the optimization problem: $\min \frac{4}{y} + 5y$, subject to $y \geq 0$. Through simple calculus, the solution turns out to be $x = 2\sqrt{5}$, and $y = \frac{2}{\sqrt{5}}$. Plugging this in we get that the optimal solution value is $2\sqrt{5} + 2\sqrt{5} = 4\sqrt{5}$.

   (b) This problem can be done by simply using the formulas discussed in class. The Lagrangean is: $L(x, \lambda, \nu) = x + 5y - \lambda_1 x - \lambda_2 y + \nu(xy - 4)$. The KKT conditions are:

$$\begin{aligned} 1 - \lambda_1 + \nu y &= 0 && (1)\\ 5 - \lambda_2 + \nu x &= 0 && (2)\\ xy &= 4 && (3)\\ x \geq 0, \quad & y \geq 0 && (4)\\ \lambda_1 x &= 0 && (5)\\ \lambda_2 y &= 0 && (6)\\ \nu(xy - 4) &= 0 && (7) \end{aligned}$$

   Due to complementary slackness, and because $x, y \geq 0$ at the optimal solution, $\lambda_1 = \lambda_2 = 0$. Moreover, $\nu = -1/y^* = -\sqrt{5}/2$.

10. (a) Let K denote the intersection of halfspaces given by $w_1, w_2, \ldots \in \mathbb{R}^d$ and $b_1, b_2, \ldots \in \mathbb{R}$:

$$K = \bigcap_i \{x : w_i \cdot x \leq b_i\}.$$

    For any $x, y \in K$ and $0 < \theta < 1$,

$$w_i \cdot (\theta x + (1-\theta)y) = \theta w_i \cdot x + (1-\theta) w_i \cdot y \leq \theta b_i + (1-\theta) b_i = b_i, \quad \text{for } i = 1, 2, ...$$

    Therefore, $\theta x + (1-\theta)y \in K$; and K is a convex set.

   (b) The unit ball in $\mathbb{R}^d$ can be written as

$$\bigcap_{\|w\|=1} \{x : w \cdot x \leq 1\}.$$

11. P_1 and P_2 are polyhedra that are intersections of finitely many halfspaces. Let the halfspaces for P_1 be given by $u_1, \ldots, u_m \in \mathbb{R}^d$ and $b_1, \ldots, b_m \in \mathbb{R}$:

$$P_1 = \bigcap_{i=1}^{m} \{x : u_i \cdot x \leq b_i\}.$$

Likewise, let P_2 be given by $v_1, \ldots, v_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$:

$$P_2 = \bigcap_{i=1}^{n} \{x : v_i \cdot x \leq c_i\}.$$

We wish to find the point $x_1 \in P_1$ and $x_2 \in P_2$ that are closest to one another. Let us write $z = x_1 - x_2$. Here is the optimization problem:

$$\begin{gathered} \min \ \|z\|^2 \\ u_i \cdot x_1 \leq b_i, \quad i = 1, 2, \ldots, m \\ v_i \cdot x_2 \leq c_i, \quad i = 1, 2, \ldots, n \\ z = x_1 - x_2 \end{gathered}$$

The constraints are are all linear, and the objective function is convex, so this is a convex optimization problem.