Optimization for Data Science, FS22 (Bernd Gärtner and Niao He)
Graded Assignment 1

Student Name, Legi Number: Deepak Narayanan, 21-959-622

Exercise 1 - Learning 3-SAT formulas

We first observe as per the question that we have 0 empirical risk, that is $l_n(\tilde{f}_n) = 0$. We need to prove the following:

$$\lim_{d \to \infty} P(l(\tilde{f}_n) > \epsilon) = 0$$

which is the same as

$$\lim_{d \to \infty} P(l(\tilde{f}_n) - l_n(\tilde{f}_n) > \epsilon) = 0$$

We note that Theorem 1.6 from the lecture notes states the following.

$$P(\sup_{H \in \mathcal{H}} |l_n(H) - l(H)| > \epsilon) \leq 4\mathcal{H}(2n)e^{-\frac{\epsilon^2 n}{8}}$$

For our condition, the empirical risk is zero for the hypotheses that we consider. Also, we have non-negative 0-1 loss. Therefore, we have the following for this particular problem:

$$P(\sup_{H \in \mathcal{H}} l(H) > \epsilon) \leq 4\mathcal{H}(2n)e^{-\frac{\epsilon^2 n}{8}}$$

The probability above is bounded by the growth function $\mathcal{H}(2n)$. For our question, if we can show that the above probability decays with d, we are done, since the theorem given us the bound for the worst case error. The question therefore boils down to computing the growth function $\mathcal{H}(2n)$ as a function of d.

Let us look at the definition of the growth function, as defined in Lecture Notes

$$\mathcal{H}(n) := \max\{|\mathcal{H} \cap \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}| : \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{D}\}$$

$$\mathcal{H} \cap \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} := \{H \cap \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} : H \in \mathcal{H}\}$$

$$\mathcal{H}(n) = \max\{|\{H \cap \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} : H \in \mathcal{H}\}| : \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{D}\}$$

The following quantity: $|\{H \cap \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} : H \in \mathcal{H}\}|$ is maximized when every hypothesis $H \in \mathcal{H}$ has a non-empty intersection with a given set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{D}$.

$$\implies \mathcal{H}(n) \leq \max\{|\mathcal{H}| : \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{D}\} = |\mathcal{H}|$$

Therefore our growth function is upper bounded by the size of our hypothesis class. That is,

$$\mathcal{H}(n) \leq |\mathcal{H}|$$

For our probability bounds, we can upperbound $\mathcal{H}(2n)$ by $|\mathcal{H}|$. We now try to compute this quantity $|\mathcal{H}|$. We are interested in 3-SAT formulas with $2d$ literals including their negations. To construct a single clause, we need 3 literals. Therefore, we have a total of $2d \times 2d \times 2d = 8d^3$ possible options for each clause. We therefore have a set of size $8d^3$. However, in our 3-SAT formula, we may include any subset of these clauses. If we count the total number of subsets of this set, it is 2^{8d^3}. Therefore, we have a total of 2^{8d^3} 3-SAT formulas. This is the size of our hypothesis class. Therefore, $|\mathcal{H}| = 2^{8d^3}$.

Going back to our probabilities, we therefore have:

$$P(\sup_{H \in \mathcal{H}} l(H) > \epsilon) \leq 4\mathcal{H}(2n)e^{-\frac{\epsilon^2 n}{8}} \leq 4|\mathcal{H}|e^{-\frac{\epsilon^2 n}{8}} = 4 \times 2^{8d^3} e^{-\frac{\epsilon^2 n}{8}} \leq 4 \times e^{8d^3} e^{-\frac{\epsilon^2 n}{8}} = 4 \times e^{8d^3 - \frac{\epsilon^2 n}{8}}$$

Now for our probability to small we firstly need $8d^3 - \frac{\epsilon^2 n}{8} < 0 \implies n > \frac{64d^3}{\epsilon^2}$. We still require decay in d. To achieve the same, we set $n = \frac{64d^3}{\epsilon^2} + poly(d)$, where $poly(d)$ is some polynomial with degree at least 1. For our example, we set $poly(d) = \frac{64d^2}{\epsilon^2}$. Therefore we have $n = p(d) = \frac{64d^3+64d^2}{\epsilon^2}$. Now plugging n back into our above equation gives,

$$P(\sup_{H \in \mathcal{H}} l(H) > \epsilon) \leq 4 \times e^{8d^3 - \frac{\epsilon^2 \frac{64d^3+64d^2}{\epsilon^2}}{8}} = 4 \times e^{8d^3 - 8d^3 - 8d^2} = 4e^{-8d^2}$$

We now take $d \to \infty$ and we have

$$\lim_{d \to \infty} P(\sup_{H \in \mathcal{H}} l(H) > \epsilon) \leq 4e^{-8d^2} = 0$$

We now observe that $l(H) > \epsilon$ for some $H \in \mathcal{H} \implies \sup_{H \in \mathcal{H}} l(H) > \epsilon$. This implies P($l(H) > \epsilon$) for some $H \in \mathcal{H} \leq P(\sup_{H \in \mathcal{H}} l(H) > \epsilon)$.
This implies,

$$\lim_{d \to \infty} P(l(H) > \epsilon) \leq \lim_{d \to \infty} P(\sup_{H \in \mathcal{H}} l(H) > \epsilon)$$

We also know that probabilties cannot be negative.

$$0 \leq \lim_{d \to \infty} P(l(H) > \epsilon) \leq \lim_{d \to \infty} P(\sup_{H \in \mathcal{H}} l(H) > \epsilon) \leq 0$$

$$\implies 0 \leq \lim_{d \to \infty} P(l(H) > \epsilon) \leq 0 \implies \lim_{d \to \infty} P(l(H) > \epsilon) = 0$$

Therefore, we have, for some $H \in \mathcal{H}$

$$\lim_{d \to \infty} P(l(H) > \epsilon) = 0$$

And more specifically, for $H = \tilde{f}_n$,

$$\lim_{d \to \infty} P(l(\tilde{f}_n) > \epsilon) = 0$$

Exercise 2 - Convexity

Part (a)

1

We use the second order characterization for convexity in this exercise and provide a proof by contradiction. Suppose $f(\mathbf{x}) = \sum_{i=1}^{d} x_i e^{x_i}$ be convex. Then second order characterization tells that f is convex if and only if the $\mathbf{dom}(f)$ is convex and $\forall \mathbf{x} \in \mathbf{dom}(f)$, the hessian evaluated at $\mathbf{x}$ is positive semidefinite. The domain of f in this case, $\mathbb{R}^d$ is convex. The hessian is defined as follows.

$$\nabla^2 f(\mathbf{x})_{i,j} := \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$$

We first compute $\frac{\partial f}{\partial x_i}(\mathbf{x})$.

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = e^{x_i} + x_i e^{x_i}$$

Therefore,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} 2e^{x_i} + x_i e^{x_i} & \text{if } i = j \\ 0, \text{else} \end{cases}$$

For hessian to be positive semidefinite, we require $\forall \mathbf{v} \in \mathbb{R}^d$, $\forall \mathbf{x} \in \mathbf{dom}(f)$, $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq 0$. Consider $\mathbf{x'} = [-3, 0, \ldots 0]^T \in \mathbb{R}^d$, with -3 at the first coordinate and 0 at the other coordinates. Clearly $\mathbf{x'} \in \mathbf{dom}(f)$. Let $\mathbf{v} = [v_1, v_2, v_3, \ldots v_d]^T$. Then clearly, we have $\mathbf{v}^T \nabla^2 f(\mathbf{x'}) \mathbf{v} = v_1(2e^{-3} - 3e^{-3})v_1 = v_1^2(-1e^{-3}) = -v_1^2 e^{-3} < 0$ if $v_1 \neq 0$. However for positive semidefiniteness we require nonnegativity for all $\mathbf{v} \in \mathbb{R}^d$. This is a thus contradiction to f being convex. Therefore, f is not convex.

2

Firstly, we show Δ_d is convex. Let us take $\mathbf{x}$, $\mathbf{y} \in \Delta_d$. Let $\lambda \in [0, 1]$. We then have $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} = [\lambda x_1 + (1-\lambda)y_1, \ldots, \lambda x_d + (1-\lambda)y_d]^T$. Now summing over individual coordinates (we do this since $\forall x \in \Delta_d, \sum_{i=1}^{d} x_i = 1$) gives us, $\lambda \sum_{i=1}^{d} x_i + (1-\lambda) \sum_{i=1}^{d} y_i = \lambda + 1 - \lambda = 1$, since $\sum_i x_i = \sum_i y_i = 1$. Also, each element of $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ lies between 0 and 1. Therefore Δ_d is convex.

For showing convexity of f on Δ_d, we reuse some calculations from the first sub-part of this question. Particularly, we will reuse our derivations of the hessian. We know the following:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} 2e^{x_i} + x_i e^{x_i} & \text{if } i = j \\ 0, \text{else} \end{cases}$$

Now for any vector $\mathbf{v} = [v_1, v_2, \ldots, v_d] \in \mathbb{R}^d$, let us consider, $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v}$. We first compute $\mathbf{v}^T \nabla^2 f(\mathbf{x})$.

$$\mathbf{v}^T \nabla^2 f(\mathbf{x}) = [v_1(2e^{x_1} + x_1 e^{x_1}), v_2(2e^{x_2} + x_2 e^{x_2}), \ldots, v_d(2e^{x_d} + x_d e^{x_d})]$$

Now, we compute $\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v}$.

$$\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} = \sum_{i=1}^{d} v_i^2(2e^{x_i} + x_i e^{x_i}) = \sum_{i=1}^{d} v_i^2 e^{x_i}(2 + x_i)$$

For this quantity to be less than 0, it is clear that at least one of the $x_i < -2$, since $e_i^x > 0$ and $v_i^2 \geq 0$ for any choice of v_i and x_i. Naturally, the hessian is positive semidefinite if we allow $\mathbf{x} \in (-2, \infty)^d$. By second order characterization, f is convex on $(-2, \infty)^d$. In particular, observe that $\Delta_d \subset (-2, \infty)^d$ and convex. Since f is convex over a set containing Δ_d and Δ_d itself is convex, f is convex over Δ_d.

3

We know that e^x is a convex function (it's second derivative is again $e^x > 0 \forall x \in \mathbb{R}$ (Using Second Order Characterization it is convex). Using Jensen's Inequality (Lemma 2.12 from lecture notes), we have the following:

$$e^{\sum_{i=1}^{d} \lambda_i x_i} \leq \sum_{i=1}^{d} \lambda_i e^{x_i}$$

where $\sum_{i=1}^{d} \lambda_i = 1$. Observe that $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_d]^T \in \Delta_d$. The above inequality holds for all choices of λ, where $\lambda_i \geq 0$ and $\sum_{i=1}^{d} \lambda_i = 1$. In particular it holds for the following choice of λ.
Set $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_d]^T = [x_1, x_2, \ldots, x_d]^T$. This implies,

$$e^{\sum_{i=1}^{d} x_i . x_i} \leq \sum_{i=1}^{d} x_i e^{x_i}$$

We thus have

$$e^{\sum_{i=1}^{d} x_i^2} = e^{||\mathbf{x}||^2} \leq \sum_{i=1}^{d} x_i e^{x_i}$$

Observe that the RHS is the function we are given in this question. Minimizing over both sides on Δ_d, we have,

$$\min_{\mathbf{x} \in \Delta_d} e^{||\mathbf{x}||^2} \leq \min_{\mathbf{x} \in \Delta_d} \sum_{i=1}^{d} x_i e^{x_i}$$

We are given $\min_{\mathbf{x} \in \Delta_d} ||\mathbf{x}||^2 = \frac{1}{d}$. e^x is monotonically increasing function. Therefore the minimum of $e^{||\mathbf{x}||^2}$ occurs when $||\mathbf{x}||^2 = \frac{1}{d}$. This implies

$$\min_{\mathbf{x} \in \Delta_d} e^{||\mathbf{x}||^2} = e^{\frac{1}{d}}$$

This further implies

$$e^{\frac{1}{d}} \leq \min_{\mathbf{x}\in\Delta_d} \sum_{i=1}^{d} x_i e^{x_i}$$

Now we have a lower bound on the minimum value of our function. We can compute the functional value for choice of $\mathbf{x} = [\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]^T \in \Delta_d$. We then have

$$f([\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]^T) = \sum_{i=1}^{d} \frac{1}{d} e^{\frac{1}{d}} = \frac{1}{d} \sum_{i=1}^{d} e^{\frac{1}{d}} = e^{\frac{1}{d}}$$

We therefore have,

$$e^{\frac{1}{d}} \leq \min_{\mathbf{x}\in\Delta_d} \sum_{i=1}^{d} x_i e^{x_i} \leq_{\mathbf{x}\in\Delta_d} f(\mathbf{x}) \implies e^{\frac{1}{d}} \leq \min_{\mathbf{x}\in\Delta_d} \sum_{i=1}^{d} x_i e^{x_i} \leq f([\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]^T) = e^{\frac{1}{d}}$$

$$\implies e^{\frac{1}{d}} \leq \min_{\mathbf{x}\in\Delta_d} \sum_{i=1}^{d} x_i e^{x_i} \leq e^{\frac{1}{d}} \implies \min_{\mathbf{x}\in\Delta_d} \sum_{i=1}^{d} x_i e^{x_i} = e^{\frac{1}{d}}$$

We therefore see that our function $f(\mathbf{x})$ attains a minimum value of $e^{\frac{1}{d}}$ and this precisely happens at the point $[\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]^T$. Therefore, the uniform distribution $[\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}]^T$ indeed minimizes f over Δ_d.

Part (b)

We use Lemma 3.11 from the Lecture Notes for this exercise. The following statements are equivalent. Note that the below statements are modified for the univariate case.

1. f is strongly convex with parameter μ.
2. $g(x) := f(x) - \frac{\mu}{2}x^2$ is convex over $\mathbf{dom}(\mathrm{g}) = \mathbf{dom}(\mathrm{f})$.

Given, $f(x) = \ln(x + \sqrt{1+x^2}) + x^2$. We show $g(x)$ defined as above is convex and prove the required result. $g(x) = \ln(x + \sqrt{1+x^2}) + x^2 - \frac{1}{2}x^2 = \ln(x + \sqrt{1+x^2}) + \frac{1}{2}x^2$, since $\mu = 1$. Note that the domain of f and g are both $\mathbb{R}$. Since g is differentiable twice, we use the second order characterization of convexity to show that g is convex. Specifically, we want to show that the second derivative of our univariate function g is non-negative.

$$\frac{dg}{dx} = \frac{1}{x + \sqrt{1+x^2}}.(1 + \frac{2x}{2\sqrt{1+x^2}}) + x = \frac{1}{x + \sqrt{1+x^2}}.\frac{x + \sqrt{1+x^2}}{\sqrt{1+x^2}} + x = \frac{1}{\sqrt{1+x^2}} + x$$

$$\frac{d^2g}{dx^2} = \frac{-x}{(1+x^2)^{\frac{3}{2}}} + 1$$

For $x < 0$, we have $\frac{d^2g}{dx^2} > 0$. Now let us consider the case when $x \geq 0$. For $x \geq 0$, consider the following function: $h(x) = 1 + x^6 + 2x^2 + 3x^4$. Note that the $h(x) \geq 0 \forall x \geq 0$.

$1 + x^6 + 2x^2 + 3x^4 \geq 0 \implies 1 + x^6 + 2x^2 + 3x^4 + x^2 \geq x^2 \implies (1+x^2)^3 \geq x^2 \implies \frac{x^2}{(1+x^2)^3} \leq 1$. Now taking square root on both sides, we have $\frac{x}{(1+x^2)^{\frac{3}{2}}} \leq 1$(Since $x \geq 0$, we do not need absolute values). This is precisely the term that we subtract from 1 in $\frac{d^2g}{dx^2}$. Therefore, $\frac{d^2g}{dx^2} \geq 0 \forall x \geq 0$. Combining both cases, we have $\frac{d^2g}{dx^2} \geq 0 \forall x \in \mathbb{R}$, proving g is convex. Using Lemma 3.11, we conclude that $f(x) = \ln(x + \sqrt{1+x^2}) + x^2$ is strongly convex with parameter $\mu = 1$.

Part (c)

We need to show that the function $f(\mathbf{x}) = \sqrt{1 + ||\mathbf{x}||^2}$ is smooth with parameter 1. We use Lemma 3.3 from lecture notes for this. According to Lemma 3.3, a function f being smooth with parameter L is equivalent to another function g defined as $g(\mathbf{x}) = \frac{L}{2}\mathbf{x}^T\mathbf{x} - f(\mathbf{x})$ is convex over $\mathbf{dom}(g) = \mathbf{dom}(f)$. Note that $\mathbf{dom}(g) = \mathbf{dom}(f) = \mathbb{R}^d$ in this case. In this case $L = 1$. Therefore we need to show that the following function is convex.

$$g(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{x}}{2} - \sqrt{1 + ||\mathbf{x}||^2}$$

Let's take two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We use the basic definition of convexity for g. For $\lambda \in [0, 1]$, we need to show that,

$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda)(g(\mathbf{y}))$$

Let us define another function $h : \mathbb{R}_+ \to \mathbb{R}$, such that $h(x) = \frac{x^2}{2} - \sqrt{1+x^2}$. We can observe that g is actually composition of h and the $||.||_2$ norm, i.e, $g(\mathbf{x}) = h(||\mathbf{x}||)$. Here $\mathbb{R}_+ := \mathbb{R} - (-\infty, 0)$ For convexity, we can then do the following:

$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = h(||\lambda\mathbf{x} + (1-\lambda)\mathbf{y}||)$$

We observe that h is monotonically increasing in it's domain. To verify this, let us look at the derivative of h.

$$\frac{dh}{dx} = x - \frac{x}{\sqrt{1+x^2}} = x\left(1 - \frac{1}{\sqrt{1+x^2}}\right)$$

In the domain of h, $x \geq 0$, and $\frac{1}{\sqrt{1+x^2}} \leq 1$. This implies $\frac{dh}{dx} \geq 0$. This combined with the triangle inequality for norms gives us:

$$g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = h(||\lambda\mathbf{x} + (1-\lambda)\mathbf{y}||) \leq h(\lambda||x|| + (1-\lambda||y||))$$

We now observe once more that h is convex over it's domain. It's domain $\mathbb{R}_+$ is convex. To verify convexity, we check for second order characteristization of convexity for h.

$$\frac{d^2h}{dx^2} = 1 - \frac{\left(\sqrt{1+x^2} - x\frac{x}{\sqrt{1+x^2}}\right)}{1+x^2} = 1 - \frac{1}{(1+x^2)^{\frac{3}{2}}}$$

The second term $\frac{1}{(1+x^2)^{\frac{3}{2}}} \leq 1$ on the domain. This implies $\frac{d^2h}{dx^2} \geq 0$. Therefore h is convex. Now that h is convex, we can use the definition of convexity to obtain,

$$g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq h(\lambda||\mathbf{x}|| + (1-\lambda||\mathbf{y}||)) \leq \lambda h(||\mathbf{x}||) + (1-\lambda)h(||\mathbf{y}||).$$

But g($\mathbf{x}$) = h($||\mathbf{x}||$) as pointed out. This implies

$$g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda h(||\mathbf{x}||) + (1-\lambda)h(||\mathbf{y}||) = \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})$$
$$\text{We therefore have, } g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})$$

This proves the convexity of g. By Lemma 3.3, f is smooth with parameter $L = 1$.

Exercise 3 - Lagrange Duality

We have the following problem:

$$\min_{x_1,x_2} f(x_1, x_2) = x_1^2 + 3x_2^2 - x_1x_2$$

subject to

$$x_1 + 2x_2 \geq 2; 3x_1 - 2x_2 = 1$$

We can rewrite the constraints as

$$2 - x_1 - 2x_2 \leq 0 \ (f_1); 3x_1 - 2x_2 - 1 = 0 \ (h_1)$$

Part (a)

Yes, the above problem is indeed a convex program. For a convex program, we need a convex objective function (f) over a convex set (our domain is $\mathbb{R}^2$, which is convex), f_1 is convex and h_1 is affine.

To check for convexity of our objective function, we can use the second order characterization of convexity.

$$\frac{\partial f}{\partial x_i} = \begin{cases} 2x_1 - x_2, & \text{if } i = 1, \\ 6x_2 - x_1, & \text{if } i = 2 \end{cases}$$

$$\frac{\partial^2 f}{\partial x_i^2} = \begin{cases} 2, & \text{if } i = 1, \\ 6, & \text{if } i = 2 \end{cases}$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1} = -1$$

We now have the entries of the hessian. For f to be convex it is sufficient that the hessian is positive semidefinite that is for any $\mathbf{v}^T = [v_1, v_2] \in \mathbb{R}^2$, $\mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} \geq 0$. For

our hessian, we have $\mathbf{v}^T\nabla^2 f(\mathbf{x})\mathbf{v} = 2v_1^2 - 2v_1v_2 + 6v_2^2 = v_1^2 + (v_1 - v_2)^2 + 5v_2^2 \geq 0$ as all terms are individually ≥ 0. Therefore f is convex.
We will now check for convexity of f_1.

$$f_1 = 2 - x_1 - 2x_2 = 2 - [1, -2]^T\mathbf{x} = 2 - \mathbf{a}^T\mathbf{x}$$

where $\mathbf{a} = [1, -2]^T$. Consider $\mathbf{u}^T = [u_1, u_2], \mathbf{v}^T = [v_1, v_2]$, $\lambda \in [0, 1]$.

For convexity, we need $f_1(\lambda\mathbf{u} + (1-\lambda)\mathbf{v}) \leq \lambda f_1(\mathbf{u}) + (1-\lambda)f_1(\mathbf{v})$.

We have $f_1(\lambda\mathbf{u} + (1-\lambda)\mathbf{v}) = 2 - \mathbf{a}^T(\lambda\mathbf{u} + (1-\lambda)\mathbf{v}) = 2 - \lambda\mathbf{a}^T\mathbf{u} - (1-\lambda)\mathbf{a}^T\mathbf{v} = \lambda(2 - \mathbf{a}^T u) + (1-\lambda)(2 - \mathbf{a}^T v) = \lambda f_1(\mathbf{u}) + (1-\lambda)f_1(\mathbf{v})$ (since, (1 - $\lambda + \lambda$)2 = 2).
Therefore f_1 is convex.

h_1 can be rewritten as $h_1(x) = [3, 2]^T\mathbf{x} - 1 = \mathbf{a}^T\mathbf{x} - 1$, where $\mathbf{a}^T = [3, 2], \mathbf{x} \in \mathbb{R}^2$. h_1 has the same form as affine functions (a linear function plus a constant) and is therefore affine.

Since all conditions for a convex program are satisifed, f is a convex program.

Part (b)

The lagrangian for the above problem is:

$$L(\mathbf{x}, \lambda, \mu) = x_1^2 + 3x_2^2 - x_1x_2 + \lambda(2 - x_1 - 2x_2) + \mu(3x_1 - 2x_2 - 1).$$

The Lagrange Dual is the function:

$$g(\lambda, \mu) = \inf_{\mathbf{x}\in\mathbb{R}^2} L(\mathbf{x}, \lambda, \mu)$$

For strong duality, it is sufficient if we have a slater point, that is a point $\mathbf{x} = [x_1, x_2]^T$, such that equality constraints are satisfied and inequality constraints are *strictly* satisfied i.e., $f_1(\mathbf{x}) < 0$. Consider the point $\mathbf{x} = [3, 4]^T$. Clearly, the equality constraints are satisfied, since 3.3 - 2.4 - 1 = 0. The inequality constraints are also strictly satisfied, since 2 - 3 - 2.4 = -9 < 0. Therefore the above convex program and its lagrange dual have a slater point, therefore satisfying the condition for strong duality. This is as mentioned in Theorem 2.47 in the Lecture Notes.

Part (c)

We use the KKT conditions for this part. The KKT condition gives us

$$\nabla_{x_1,x_2} L(x, \mu, \lambda) = 0$$

Using this, we obtain the following:

$$\nabla_{x_1} L(x, \mu, \lambda) = 0 \implies 2x_1 - x_2 - \lambda + 3\mu = 0 \implies 2x_1 - x_2 = \lambda - 3\mu$$

$$\nabla_{x_2} L(x, \mu, \lambda) = 0 \implies 6x_2 - x_1 - 2\lambda - 2\mu = 0 \implies 6x_2 - x_1 = 2\lambda + 2\mu$$

The KKT conditions also gives us complementary slackness, which is

$$\lambda(2 - x_1 - 2x_2) = 0$$

In the above equation either $\lambda = 0$ or $(2 - x_1 - 2x_2) = 0$. Suppose $\lambda \neq 0$.

$$\implies 2 - x_1 - 2x_2 = 0 \qquad 3x_1 - 2x_2 = 1 \text{ (From equality constraint)}$$

Upon solving, we get $x_1 = \dfrac{3}{4}$ $x_2 = \dfrac{5}{8}$

We can now plug these values in the equations obtained by using vanishing gradients of the lagrangian and we have

$$\lambda - 3\mu = \frac{7}{8} \qquad\qquad 2\lambda + 2\mu = 3$$

$$\implies \lambda = \frac{43}{32} \qquad\qquad \mu = \frac{5}{32}$$

Since $\lambda > 0$, it is a feasible solution for the dual. We have a feasible solutions for the primal and it's lagrange dual satisfying the KKT conditions. Further, we do not have any duality gap due to strong duality which implies, $\mathbf{x} = [x_1^*, x_2^*]^T = [\frac{3}{4}, \frac{5}{8}]^T$ is the minimizer of the given convex program, due to Theorem 2.52 from the lecture notes.

Exercise 4

We have $g(x) = (f(x) - 1)^2$ Our update rule is:

$$x_{t+1} = x_t - \frac{1}{L}\nabla g(x_t)$$

We note that

$$\nabla g(x) = 2(f(x) - 1)\nabla(f(x))$$

We consider the change in value of g because of gradient descent. Let us consider two consecutive steps, and observe the change in value of g. Specifically, we are interested in $g(x_{t+1}) - g(x_t)$. Since g is L-smooth, and learning rate here is $\frac{1}{L}$, we use Lemma 3.7 from lecture notes, which states the following:

$$g(x_{t+1}) - g(x_t) \leq -\frac{1}{2L}||\nabla g(x_t)||^2$$

Using this condition, and plugging in the value for $\nabla g(x_t)$, we have:

$$g(x_{t+1}) - g(x_t) \leq -\frac{1}{2L}4(f(x_t) - 1)^2||\nabla f(x_t)||^2 = -\frac{2}{L}g(x_t)||\nabla f(x_t)||^2$$

We know that in the domain that we consider $||\nabla f(x_t)|| \geq \beta \implies -||\nabla f(x_t)|| \leq -\beta$ Therefore,

$$g(x_{t+1}) - g(x_t) \leq -\frac{2}{L} g(x_t) ||\nabla f(x_t)||^2 \leq -\frac{2}{L} g(x_t) \beta^2$$

$$g(x_{t+1}) \leq g(x_t) - \frac{2}{L} g(x_t) \beta^2 = (1 - \frac{2\beta^2}{L}) g(x_t)$$

We have a neat recursive update now. If we repeat gradient descent for T iterations, we have,

$$g(x_T) \leq \left(1 - \frac{2\beta^2}{L}\right)^T g(x_0) \leq \left(1 - \frac{2\beta^2}{L}\right)^T \alpha$$

We want to compute the number of iterations to reach error ϵ off from the minimum functional value. To always guarantee this, we effectively need our iterate at iteration T to be less than equal to ϵ, since $g(x) \geq 0 \implies g(x^*) \geq 0$ and hence $g(x_T) - g(x*) \leq \epsilon$. This implies,

$$\left(1 - \frac{2\beta^2}{L}\right)^T \alpha \leq \epsilon \implies T \log \left(1 - \frac{2\beta^2}{L}\right) \leq \log \frac{\epsilon}{\alpha} \implies T \log \frac{1}{1 - \frac{2\beta^2}{L}} \geq \log \frac{\alpha}{\epsilon}$$

$$\implies T \geq \frac{\log \frac{\alpha}{\epsilon}}{\log \frac{1}{1 - \frac{2\beta^2}{L}}}$$

Therefore we can set $T = \frac{\log \frac{\alpha}{\epsilon}}{\log \frac{1}{1 - \frac{2\beta^2}{L}}} = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$. Therefore $T = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ iterations suffice, as after these many iterations, we have $g(x_T) \leq \epsilon$. And since $g(x) \geq 0$, we have either found a better point than the minimizer or we are definitely within ϵ of the minimizer.