SOLUTIONS HW 2

1 Problem 1

1. The function f does not have maximum over $\mathbb{R}^3$ because $f(x_1, 0, 0) = 2x_1^2 - 2x_1 + 5$ is not bounded. The function f has a unique minimum. Indeed,
$$\nabla f = [4x_1 - 2, 4x_2 + 2x_3 - 2, 2x_3 + 2x_2 - 2]^T \tag{1}$$
and $\nabla f(x) = 0 \Rightarrow (x_1, x_2, x_3) = (0.5, 0, 1)$. Since,
$$\nabla^2 f = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 2 \\ 0 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{pmatrix} \tag{2}$$
is PD, we conclude the result.

2. Since Q is PD we have $\nabla f(x) = Qx$. We consider $g(a_k) = f(x_k - a_k\, Q\, x_k)$ and we minimize g
$$g(a_k) = f((I - a_k\, Q)\, x_k) = \frac{1}{2} x_k^T Q x_k - (x_k^T Q^2 x_k) a_k + \frac{1}{2}(x_k^T Q^3 x_k) a_k^2 \tag{3}$$
Hence $g(a_k)$ is minimized when $a_k = \frac{x_k^T Q^2 x_k}{x_k^T Q^3 x_k}$

3. Let us consider matrix A whose i^{th} row is a_i and the column vector $b = (b_1, .., b_n)^T$, then
$$\begin{aligned} f(x) &= \frac{1}{n}(Ax - b)^T (Ax - b) + \frac{\lambda}{2} x^T I x \\ &= \frac{1}{n}(x^T A^T A x + b^T b - b^T A x - x^T A^T b) + \frac{\lambda}{2} x^T I x \end{aligned} \tag{4}$$
We have
$$\nabla f = \frac{1}{n}(2A^T A x - 2A^T b) + \lambda x = \left(\frac{2}{n} A^T A + \lambda I\right) x - \frac{2}{n} A^T b \tag{5}$$
and $\nabla f = 0 \Rightarrow x^* = \left(A^T A + \frac{n}{2}\lambda I\right)^{-1} A^T b$. Also, $\nabla^2 f = \frac{2}{n} A^T A + \lambda I$ is PD because $x^T \nabla^2 x = \frac{2}{n}(Ax)^T (Ax) + \lambda x^T x > 0$ for all $x \neq 0$. Hence, the optimal solution x^* is unique. It is worth mentioning that $A^T A = \sum_{i=1}^n a_i a_i^T$ and $A^T b = \sum_{i=1}^n a_i^T b_i$.

2 Problem 2

1. Let us fix y_1, $y_2 \in \mathbb{R}^n$, x_1, $x_2 \in \mathbb{R}^n$, $a \in [0, 1]$. First, we assume that f is convex and we prove that this is also the case for g. Indeed,
$$\begin{aligned} g(ax_1 + (1-a)x_2) &= f((ax_1 + (1-a)x_2)(y_1 - y_2) + y_2), \quad \text{by definition of g} \\ &= f(a(x_1(y_1 - y_2) + y_2) + (1-a)(x_2(y_1 - y_2) + y_2)) \\ &\leq a f(x_1(y_1 - y_2) + y_2) + (1-a) f(x_2(y_1 - y_2) + y_2), \quad \text{by convexity of f} \\ &= a g(x_1) + (1-a) g(x_2) \end{aligned} \tag{6}$$
Next, we assume that g is convex and we prove that this is also the case for f. Indeed,
$$\begin{aligned} f(ay_1 + (1-a)y_2) &= f(a(y_1 - y_2) + y_2) \\ &= g(a) \\ &\leq a\, g(1) + (1-a)\, g(0), \quad \text{by convexity of g} \\ &= a f(y_1) + (1-a) f(y_2), \quad \text{by definition of g} \end{aligned} \tag{7}$$

2. Yes. The function $f(x) = x\log(x)$, $x > 0$ is convex since $f''(x) = \frac{1}{x} > 0$. We also prove that the function $g(x, y) = x\log(x) + y\log(y)$ is convex. Indeed, let us fix x_1, x_2, y_1, $y_2 \in \mathbb{R}^+$ and $a \in [0, 1]$, then

$$\begin{aligned} g(ax_1 + (1-a)x_2, ay_1 + (1-a)y_2) &= f(ax_1 + (1-a)x_2) + f(ay_1 + (1-a)y_2) \\ &\leq af(x_1) + (1-a)f(x_2) + af(y_1) + (1-a)f(y_2) \\ &= ag(x_1, y_1) + (1-a)g(x_2, y_2) \end{aligned} \tag{8}$$

As a result the set $\mathcal{S} \equiv \{(x_1, x_2) : x_1, x_2 > 0, \quad g(x_1, x_2) \leq 4\}$ is convex.

3. Let us fix x_1, $x_2 \in \mathbb{R}^n$ and $a \in [0, 1]$. By concavity of g it holds $g(ax_1 + (1-a)x_2) \geq ag(x_1) + (1-a)g(x_2)$. In order to prove that $f \circ g$ is concave, we proceed as follows

$$\begin{aligned} f(g(ax_1 + (1-a)x_2)) &\geq f(ag(x_1) + (1-a)g(x_2)), \quad \text{by concavity of g, \& the fact f is increasing} \\ &\geq af(g(x_1)) + (1-a)f(g(x_2)), \quad \text{by concavity of f} \end{aligned} \tag{9}$$

Hence, $f \circ g$ is concave.

3 Problem 3

We must find the minimum m such that

$$f(x_k + \beta^m \tilde{\alpha} d_k) \leq f(x_k) + \sigma \beta^m \tilde{\alpha} \nabla f^T d_k \tag{10}$$

where $\nabla f = [4x_1, 8x_2^3]^T$, and since we apply steepest decent we choose $d_k = -\nabla f$. Hence, by substitution we obtain

$$f(1 - 0.5^m 4, 0) = 2(1 - 0.5^m 4)^2 \leq 2 - 0.80.5^m \tag{11}$$

and the minimum m that satisfies the inequality is $m = 2$, which implies that $a_k = \tilde{\alpha}\beta^m = 1 \cdot 0.5^2 = 0.25$.

4 Problem 4

We have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq (\nabla f(x_k))^T \alpha\, D\, \nabla f(x_k) - \frac{L}{2}\|\alpha\, D\, \nabla f(x_k)\|_2^2 \\ &\geq \alpha\left(\lambda_{min} - \frac{L}{2}\alpha\lambda_{max}^2\right)\|\nabla f(x_k)\|^2 \end{aligned} \tag{12}$$

We know $\lambda_{min} - \frac{L}{2}\alpha\lambda_{max}^2 > 0$. We observe that

$$\alpha\left(\lambda_{min} - \frac{L}{2}\alpha\lambda_{max}^2\right)\sum_{k=0}^{n}\|\nabla f(x_k)\|^2 \leq f(x_0) - f(x_{n+1}) \leq f(x_0) - f_{min} \tag{13}$$

As a result for all $n \in \mathbb{N}$

$$\sum_{k=0}^{n}\|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f_{min}}{\alpha\left(\lambda_{min} - \frac{L}{2}\alpha\lambda_{max}^2\right)} \tag{14}$$

which implies that as $n \to \infty$ the series converges and as a result $\lim_{n\to\infty} \nabla f(x_n) = 0$.

5 Problem 5

1. We have

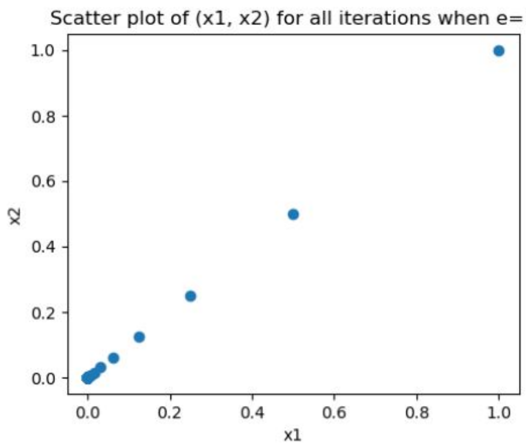
$$\nabla f = [2x_1 + 2\frac{1-\varepsilon}{1+\varepsilon}x_2, 2x_2 + 2\frac{1-\varepsilon}{1+\varepsilon}x_1]^T, \quad \text{and} \quad \nabla^2 f = \begin{pmatrix} 2 & 2\frac{1-\varepsilon}{1+\varepsilon} \\ 2\frac{1-\varepsilon}{1+\varepsilon} & 2 \end{pmatrix} \tag{15}$$

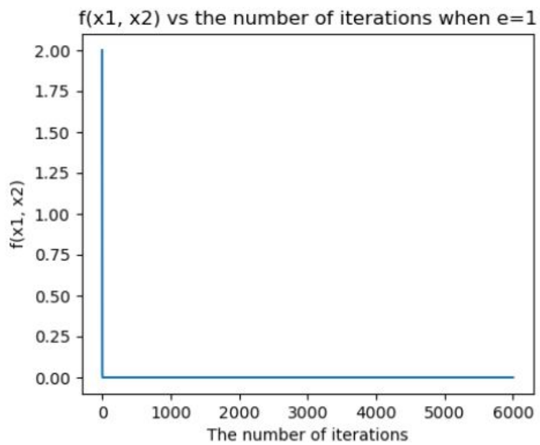
Since $0 < (1-\varepsilon)/(1+\varepsilon) < 1$ we have $\nabla^2 f \succ 0$, the unique minimizer is the solution of $\nabla f = 0$ which is $x_1 = x_2 = 0$.

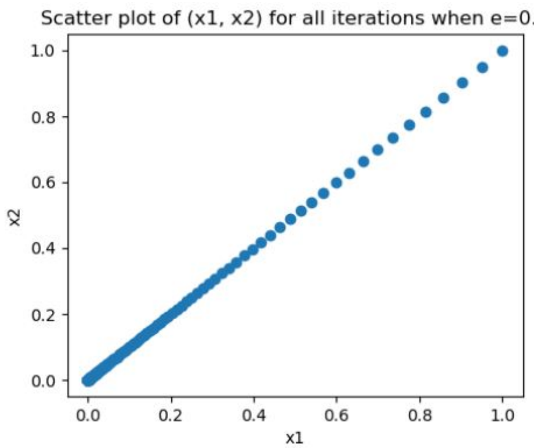
2. We must have
$$\begin{pmatrix} 2-m & 2\frac{1-\varepsilon}{1+\varepsilon} \\ 2\frac{1-\varepsilon}{1+\varepsilon} & 2-m \end{pmatrix} \succeq 0, \quad \begin{pmatrix} M-2 & -2\frac{1-\varepsilon}{1+\varepsilon} \\ -2\frac{1-\varepsilon}{1+\varepsilon} & M-2 \end{pmatrix} \succeq 0 \tag{16}$$
or equivalently
$$2-m \geq 0, \quad (2-m)^2 - \left(2\frac{1-\varepsilon}{1+\varepsilon}\right)^2 \geq 0 \quad \text{and} \quad M-2 \geq 0, \quad (M-2)^2 - \left(2\frac{1-\varepsilon}{1+\varepsilon}\right)^2 \geq 0 \tag{17}$$
The largest possible m is $2 - 2\frac{1-\varepsilon}{1+\varepsilon}$ and the smallest possible M is $2 + 2\frac{1-\varepsilon}{1+\varepsilon}$. Hence, $\kappa = M/m = 1/\varepsilon$

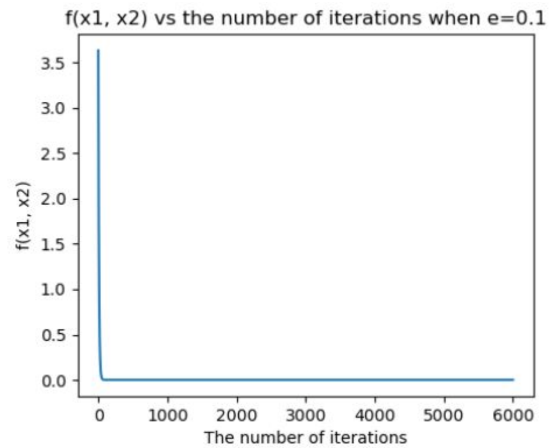
3. As $\varepsilon \to 0$, it holds $\kappa = 1/\varepsilon \to \infty$. Thus, we should expect gradient descent to converge slower.

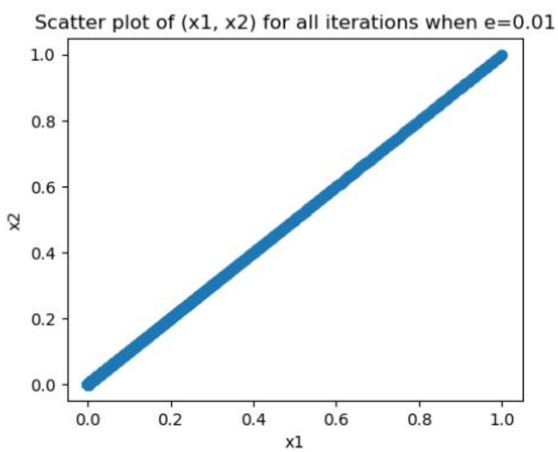
4. In the following figures we first verify that as $\varepsilon \to 0$ the gradient descent converges slower and then that for $\alpha = 1/M$ the algorithm converges.

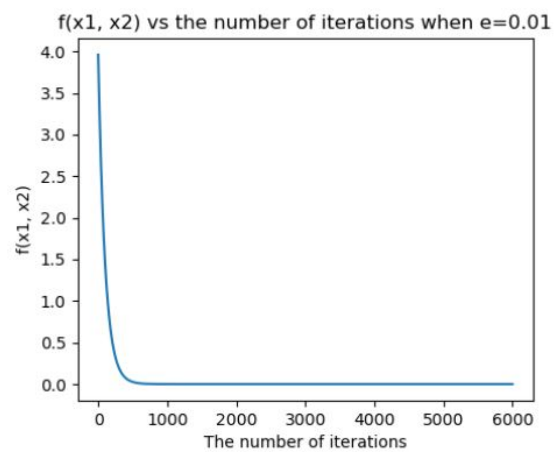


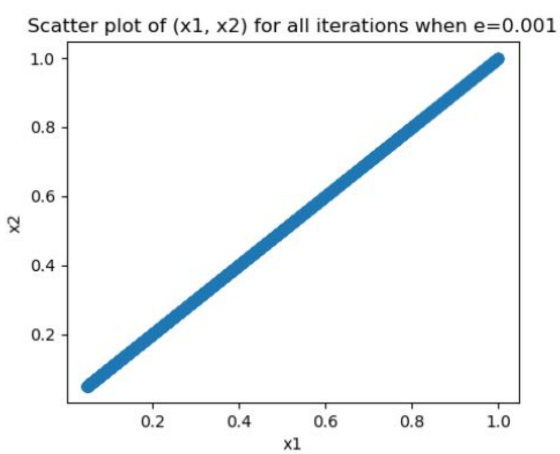


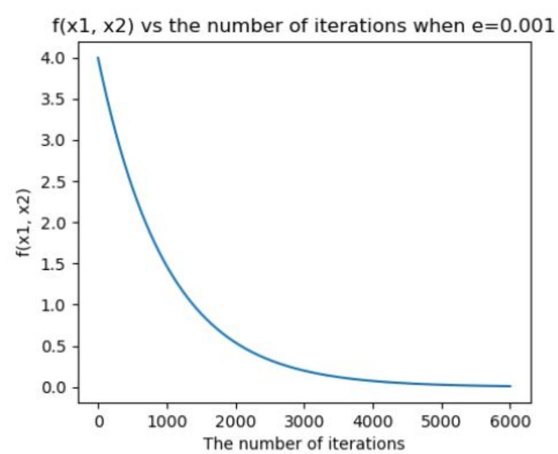

Scatter plot of (x1, x2) for all iterations when e=0.1
x2
x1

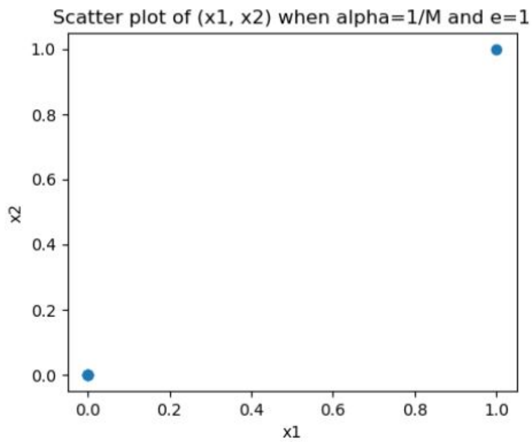

f(x1, x2) vs the number of iterations when e=0.1
f(x1, x2)
The number of iterations

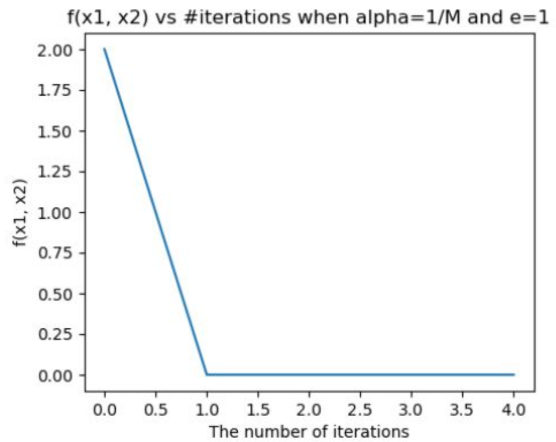

Scatter plot of (x1, x2) for all iterations when e=0.01
x2
x1

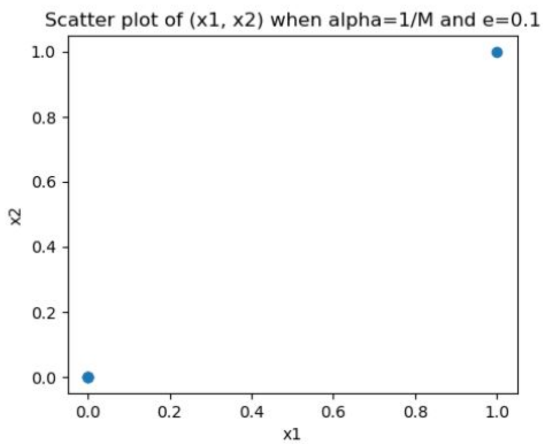

f(x1, x2) vs the number of iterations when e=0.01
f(x1, x2)
The number of iterations

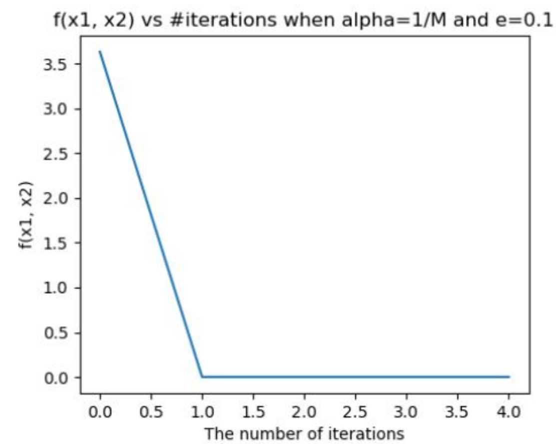

Scatter plot of (x1, x2) for all iterations when e=0.001
x2
x1

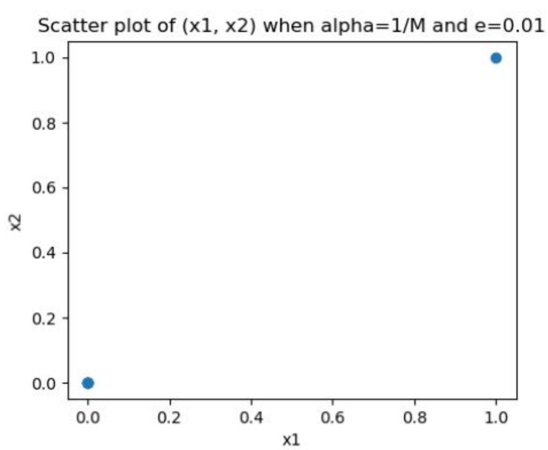

f(x1, x2) vs the number of iterations when e=0.001
f(x1, x2)
The number of iterations

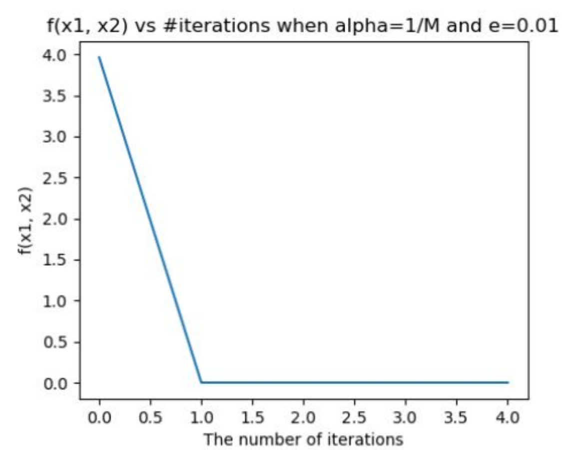

Scatter plot of (x1, x2) when alpha=1/M and e=1
x2
x1

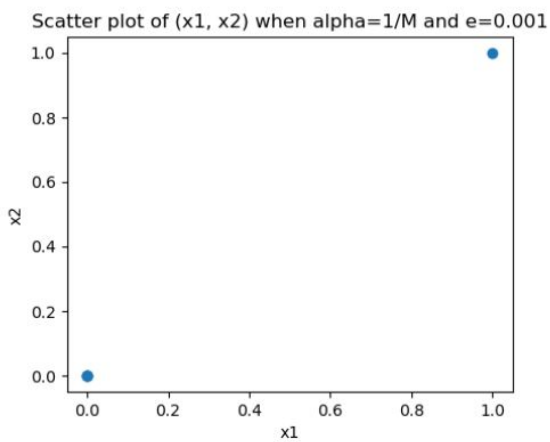

f(x1, x2) vs #iterations when alpha=1/M and e=1
f(x1, x2)
The number of iterations

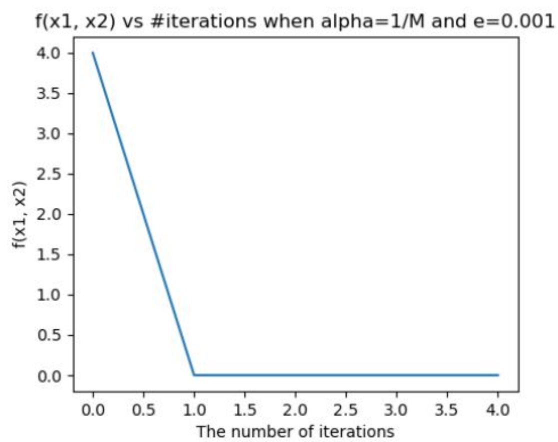

Scatter plot of (x1, x2) when alpha=1/M and e=0.1
x2
x1


f(x1, x2) vs #iterations when alpha=1/M and e=0.1
f(x1, x2)
The number of iterations


Scatter plot of (x1, x2) when alpha=1/M and e=0.01
x2
x1


f(x1, x2) vs #iterations when alpha=1/M and e=0.01
f(x1, x2)
The number of iterations


Scatter plot of (x1, x2) when alpha=1/M and e=0.001
x2
x1
0.0
0.2
0.4
0.6
0.8
1.0


f(x1, x2) vs #iterations when alpha=1/M and e=0.001
f(x1, x2)
The number of iterations
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

6 Problem 6

1. In order to prove this part, we use a property of smooth functions called Co-coercivity, which states that $\|\nabla g(x) - \nabla g(y)\|^2 \le L(\nabla g(x) - \nabla g(y))^\mathsf{T}(x-y)$ for any g being convex and L-smooth. First, we prove this property. Define $h(x) := g(x) - x^T \nabla g(y)$. By definition of convexity, $h(x)$ is convex when $g(x)$ is convex. In addition, we have $\nabla h(x) = \nabla g(x) - \nabla g(y)$. From this gradient formula, we can see h is L-smooth if g is L-smooth. In addition, h has the minimum at $x = y$ (since $\nabla h(y) = 0$). Therefore, we have $h(y) \le h(z)$ for any arbitrary z. We choose $z = x - \frac{1}{L}\nabla h(x)$. Then we can use the L-smoothness of h to show
$$\begin{aligned} h(y) \le h\left(x - \frac{1}{L}\nabla h(x)\right) &\le h(x) + \nabla h(x)^\mathsf{T}(x - (1/L)\nabla h(x) - x) + \frac{L}{2}\|x - (1/L)\nabla h(x) - x\|^2 \\ &= h(x) - \frac{1}{2L}\|\nabla h(x)\|^2 \end{aligned}$$
From the above property, we can directly show the following (the second inequality holds since the first inequality holds for arbitrary (x, y) such that we can exchange x with y)
$$\begin{aligned} g(y) - y^\mathsf{T}\nabla g(y) &\le g(x) - x^\mathsf{T}\nabla g(y) - \frac{1}{2L}\|\nabla g(x) - \nabla g(y)\|^2 \\ g(x) - x^\mathsf{T}\nabla g(x) &\le g(y) - y^\mathsf{T}\nabla g(x) - \frac{1}{2L}\|\nabla g(y) - \nabla g(x)\|^2 \\ &\to \frac{1}{L}\|\nabla g(x) - \nabla g(y)\|^2 + (\nabla g(y) - \nabla g(x))^\mathsf{T}(x-y) \le 0 \\ &\quad \|\nabla g(x) - \nabla g(y)\|^2 \le L(\nabla g(y) - \nabla g(x))^\mathsf{T}(y-x) \end{aligned}$$
The above inequality holds for any $L \ge 0$ (if $L = 0$, it is trivially true). Now, let $g(x) = f(x) - \frac{m}{2}\|x\|^2$. It is straightforward to verify the convexity of g as follows
$$\begin{aligned} g(y) = f(y) - \frac{m}{2}\|y\|^2 &\ge f(x) + \nabla f(x)^\mathsf{T}(y-x) + \frac{m}{2}\|y-x\|^2 - \frac{m}{2}\|y\|^2 \\ &= f(x) - \frac{m}{2}\|x\|^2 + (\nabla f(x) - mx)^\mathsf{T}(y-x) = g(x) + \nabla g(x)^\mathsf{T}(y-x) \end{aligned}$$
Similarly, we can use the L-smoothness property of f to show that g is $(L-m)$-smooth.
$$\begin{aligned} g(y) = f(y) - \frac{m}{2}\|y\|^2 &\le f(x) + \nabla f(x)^\mathsf{T}(y-x) + \frac{L}{2}\|y-x\|^2 - \frac{m}{2}\|y\|^2 \\ &= g(x) + \nabla g(x)^\mathsf{T}(y-x) + \frac{L-m}{2}\|y-x\|^2 \end{aligned}$$
Using the co-coercivity property of g, the following inequality holds
$$\|\nabla g(x) - \nabla g(y)\|^2 \le (L-m)(\nabla g(x) - \nabla g(y))^\mathsf{T}(x-y)$$
which is equivalent to
$$\|\nabla f(x) - \nabla f(y) - m(x-y)\|^2 \le (L-m)(\nabla f(x) - \nabla f(y) - mx + my)^\mathsf{T}(x-y).$$
One can verity that the above inequality directly leads to the desired conclusion.

2. For $\alpha = \frac{1}{L}$ and $\rho = 1 - \frac{m}{L}$, we only need to show that we can find some non-negative λ to make the matrix $\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix}$ negative semidefinite. Now we choose $\lambda = \frac{1}{L^2}$. Then we have
$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda \begin{bmatrix} -2mL & m+L \\ m+L & -2 \end{bmatrix} = \begin{bmatrix} -\frac{m^2}{L^2} & \frac{m}{L^2} \\ \frac{m}{L^2} & -\frac{1}{L^2} \end{bmatrix} = \frac{1}{L^2}\begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \tag{18}$$
The right side is clearly negative semidefinite due to the fact that $\begin{bmatrix} a \\ b \end{bmatrix}^\mathsf{T} \begin{bmatrix} -m^2 & m \\ m & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = -(ma-b)^2 \le 0$ for arbitrary (a, b). Therefore, the gradient method with $\alpha = \frac{1}{L}$ converges as $\|x_k - x^*\| \le \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|$.