Math 4650/MSSC 5650 - Homework 8

Instructor: Greg Ongie

Spring 2023

Problem 1 (5 pts)**.** Determine whether each function listed below is L-smooth for some $L > 0$ and/or μ-strongly convex for some $\mu > 0$. If it is L-smooth, specify the smallest possible value of L. If it is μ-strongly convex, specify the largest possible value of μ.

(a) $f(x) = x^2 + \sin(x)$

(b) $f(x) = x^3$

(c) $f(x) = \sqrt{1 + x^2}$

(d) $f(x, y) = x^2 + xy + y^2 + 5x + 3y - 1$

(e) $f(x, y) = \cos(x) - y^2$.

Solution 1. (a)

$$f'(x) = 2x + cos(x)$$
$$f''(x) = 2 - sin(x)$$

Since $|-sin(x)| \leq 1$ then $|f''(x)| = |2 - sin(x)| \leq 3 = L$. There fore $f(x)$ is 3-smooth.

Now let's check for μ-strongly convex.

Since $-1 \leq sin(x) \leq 1$ for all x, then $f''(x) = 2 - sin(x) \geq 1 = \mu$. So $f(x)$ is 1-strongly convex.

(b)

$$f'(x) = 3x^2$$
$$f''(x) = 6x \Rightarrow |f''(x)| < \infty$$

$f(x)$ is not L-smooth, since $f''(x)$ is not bounded above. $f(x)$ is not μ-strongly convex either, $f''(-1) = -6$, since it is not even convex.

(c)

$$f'(x) = \frac{x}{\sqrt{1+x^2}}$$

$$f''(x) = \frac{1}{(1+x^2)^{1/2}} - \frac{x^2}{(1+x^2)^{3/2}} = \frac{(1+x^2)^{3/2} - x^2(1+x^2)^{1/2}}{(1+x^2)^2} = \frac{1}{(1+x^2)^2}$$

Since $x^2 \geq 0$ and the x that gives the smallest denominator is 0, then $f''(x) \leq 1 = L$. Therefore, $f(x)$ is 1-smooth.

Since $\lim_{x\to\infty} \frac{1}{(1+x^2)^{3/2}} = 0 = \mu$, then $f(x)$ is not μ-strongly convex.

(d)

$$\nabla f(x,y) = \begin{bmatrix} 2x+y+5 \\ x+2y+3 \end{bmatrix}$$

$$\nabla^2 f(x,y) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \text{ with eigen values } \lambda_1 = 3, \lambda_2 = 1$$

$\|\nabla^2 f(x,y)\| = max|\lambda_1|, |\lambda_2| = 3$, so f is 3-smooth.

Let's check for μ-strongly convex.

Since $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \succ 0$, because all eigenvalues are positive, then f is 1-strongly convex, and $\mu = 1 = \lambda_{min}$

(e)

$$\nabla f(x,y) = \begin{bmatrix} -2sin(x) \\ -2y \end{bmatrix}$$

$$\nabla^2 f(x,y) = \begin{bmatrix} -cos(x) & 0 \\ 0 & -2 \end{bmatrix}$$

The eigenvalues of the Hessian are $\lambda_1 = -cos(x) \Rightarrow |\lambda_1 \leq 1$ and $\lambda_2 = -2 \Rightarrow |\lambda_2| = 2$

Then $\|\nabla^2 f(x,y)\| = max|\lambda_1|, |\lambda_2| = 2$, so f is 2-smooth.

Let's check for μ-strongly convex.

Since $\nabla^2 f(\pi) = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$ is indefinite, then this function is not convex, hence not μ-strongly convex.

Problem 2 (MATLAB, 10 pts)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the function

$$f(x_1, x_2) = \sqrt{1+x_1^2} + \sqrt{1+x_2^2}.$$

Note that $(x_1, x_2) = (0,0)$ is the unique global minimizer of f.

(a) Implement "Pure Newton's Method" in MATLAB to minimize this function (for guidance on how to code this, see pgs. 85-86 in Ch. 5 of Beck).

- Find an initial point $(x_0, y_0) \neq (0, 0)$ for which Newton's method converges to $(0, 0)$.
- Find an initial point (x_0, y_0) for which Newton's method *diverges*.

(b) Implement "Damped Newton's Method" in MATLAB (see pg. 89 in Ch. 5 of Beck). Show that the initial point (x_0, y_0) that led to divergent iterates in part (a) with "Pure Newton's Method" now converges to the minimizer $(0, 0)$ with "Damped Newton's Method".

Solution 2. (a) .

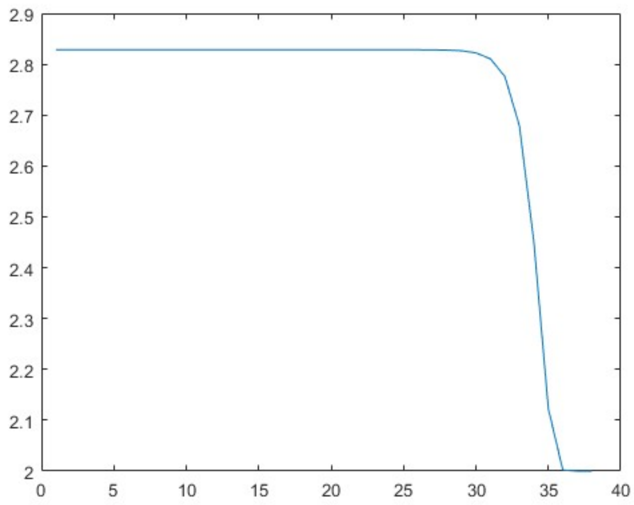
```
%% define variables
f=@(x)sqrt(1+x(1)^2)+sqrt(1+x(2)^2);
grad=@(x)[x(1)/sqrt(x(1)^2+1);x(2)/sqrt(x(2)^2+1)];
hessian = @(x)diag([1/(x(1)^2+1)^1.5,1/(x(2)^2+1)^1.5]);

%% Run main for-loop
x = [1;1];
maxiter = 1000;
tol = 1e-7;


cost = f(x); %initialize cost array
for k=1:maxiter
    x = x - hessian(x)\grad(x);
    cost = [cost,f(x)];

    if norm(grad(x)) < tol
        fprintf('algorithm converged at iteration k=%d,\nx1
           =%e\nx2=%e\n',k,x(1),x(2));
        break;
    end
end
%% plot cost
plot(cost);
```

- Choice of initial point that converges: $x_0 = 1, y_0 = 1$

  Output:

```
algorithm converged at iteration k=37,
x1=-7.257095e-13
x2=-7.257095e-13
```

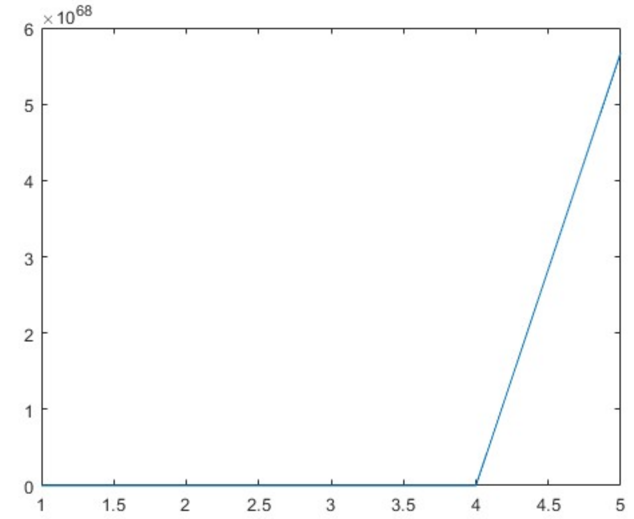
Plot of the cost:



- Choice of initial point that diverges: $x_0 = 7, y_0 = 7$
  Output:

```
algorithm converged at iteration k=5,
x1=-2.284671e+205
x2=-2.284671e+205
```

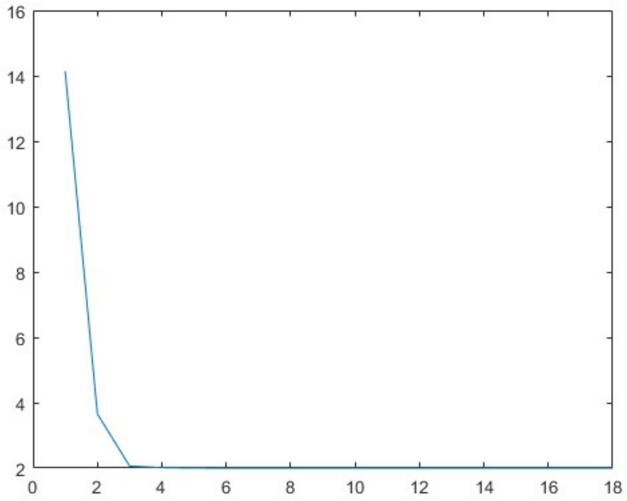
Plot of the cost:



(b) .

```
%% Run main for-loop
% DIVERGING X
x = [7;7];
maxiter = 1000;
tol = 1e-7;
alpha = 0.5;
beta = 0.5;

cost = f(x); %initialize cost array
for k=1:maxiter
    g = grad(x);
    h = hessian(x);
    t = 1;
    d=hessian(x)\grad(x);
    while(f(x-t*d)>f(x)-alpha*t*g'*d)
        t=beta*t;
    end
    x = x -t*d;
    cost = [cost,f(x)]; %store f(x_k) for plotting

    if norm(grad(x)) < tol
        fprintf('algorithm converged at iteration k=%d,\nx1=%e\nx2=%e\n',k,x(1),x(2));
        break;
    end
end
```

Plot of the cost:



Choice of initial point that diverges: $x_0 = 7, y_0 = 7$

Output:

```
algorithm converged at iteration k=17,
x1=-2.794702e-15
x2=-2.794702e-15
```

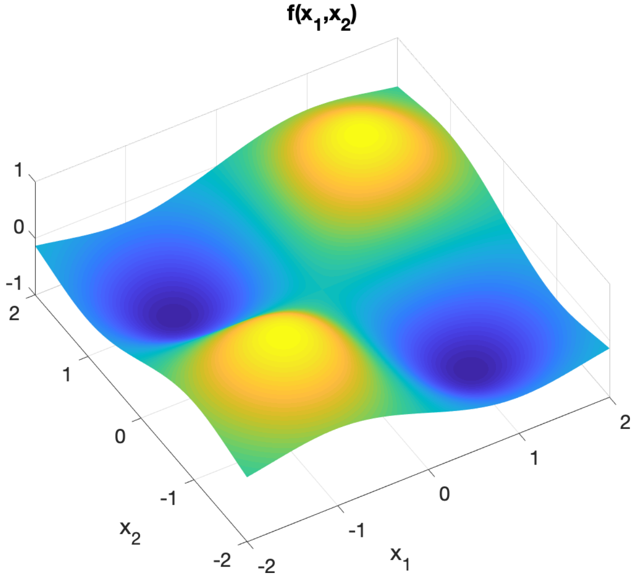
Problem 3 (MATLAB, 10 pts). Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \frac{1}{4}\Big(f_1(\mathbf{x}) + f_2(\mathbf{x}) + f_3(\mathbf{x}) + f_4(\mathbf{x})\Big)$$

where

$$f_i(\mathbf{x}) = b_i \exp(-\|\mathbf{x} - \mathbf{a}_i\|^2), \quad \text{for } i = 1, ..., 4,$$

with $\mathbf{a}_1 = [1, 1]^\top$, $\mathbf{a}_2 = [1, -1]^\top$, $\mathbf{a}_3 = [-1, 1]^\top$, $\mathbf{a}_4 = [-1, -1]^\top$ and $b_1 = 1$, $b_2 = -1$, $b_3 = -1$, $b_4 = 1$. This function is shown in the plot below. Note $f(\mathbf{x})$ has a saddle point at $\mathbf{x} = (0, 0)$.



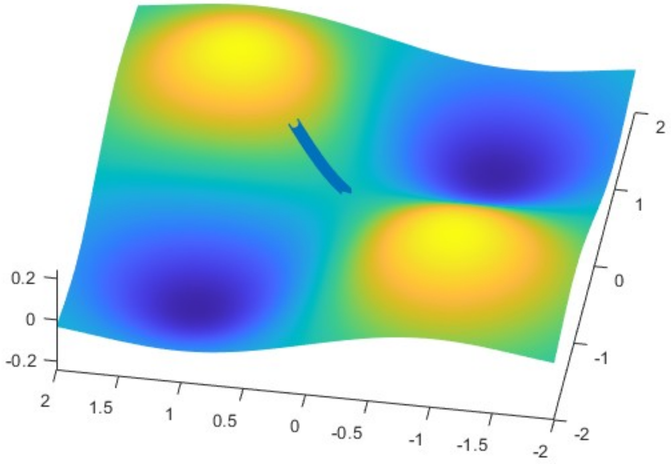
Your task is to compare the performance of gradient descent (GD) and stochastic gradient descent (SGD) in minimizing this function.

(a) Run the provided script `problem3.m` to plot the trajectory of gradient descent iterates starting from the initial point $\mathbf{x}_0 = [0.5, 0.5]^\top$ with a constant stepsize $t = 0.01$ for 1000 iterations (you don't need to modify anything).

(b) Now modify the script to implement stochastic gradient descent (SGD) to minimize this function. This can be done by replacing full gradient $\nabla f(\mathbf{x})$ in the main for loop with the partial gradient $\nabla f_i(\mathbf{x})$ where $i \in \{1, 2, 3, 4\}$ is a randomly selected index; in MATLAB this can be done with the command `i=randi(4)`. Plot the trajectory of iterates obtained using SGD starting from the initial point $\mathbf{x} = [0.5, 0.5]^\top$ with a constant stepsize $t = 0.01$ for 1000 iterations.

What algorithm (GD or SGD) found a better minimizer after 1000 iterations? What happens if you change the initial point to $\mathbf{x}_0 = [1, 0]$? Provide an explanation for what you observe.

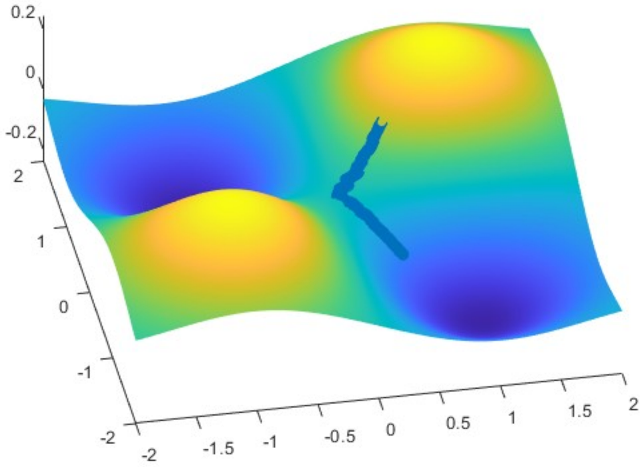
Solution 3. (a) .



(b) .

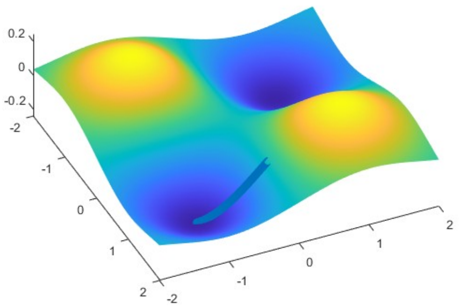
```
%define cost function and gradient
fa = @(x,a) exp(-norm(x-a)^2);
ga = @(x,a) -2*exp(-norm(x-a)^2)*(x-a);
A = [1 1 -1 -1;...
     1 -1 1 -1];
b = [1 -1 -1 1];

f = @(x) 0.25*(b(1)*fa(x,A(:,1))+b(2)*fa(x,A(:,2))+b(3)*fa(x
   ,A(:,3))+b(4)*fa(x,A(:,4)));
grad = @(x) 0.25*(b(1)*ga(x,A(:,1))+b(2)*ga(x,A(:,2))+b(3)*
   ga(x,A(:,3))+b(4)*ga(x,A(:,4)));


x = [0.5;0.5]; %initial point
t = 0.01; %GD stepsize

xar = x; %initialize x-array
cost = f(x); %initialize cost array
for k=1:1000
    i = randi(4);
    a = A(:,i);
    b_i = b(i);
    x = x - t*b_i*ga(x,a); %SGD update

    xar = [xar,x];
    cost = [cost,f(x)];
end

figure(1); %plot trajectories
hold on
[X,Y] = meshgrid(-2:0.01:2,-2:0.01:2);
Z = zeros(size(X));
for i=1:numel(X)
    x = [X(i);Y(i)];
    Z(i) = f(x);
end
figure(1);
surf(X,Y,Z);
shading interp
scatter3(xar(1,:),xar(2,:),cost); view(-15,50);
hold off

%%
figure(2)
plot(cost)
```

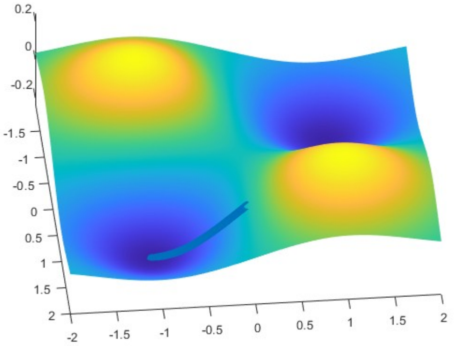
Plot of iteration on the function:



After 1000 iterations SGD did better than GD, since GD got stuck in the saddle point, while SDG kept going. Both did not reach the minimum of the function, but SGD with more iterations or better step size would more likely be able to reach the minimum.

Now, if we change the initial point, we get the following:





The iterations are pretty much the same, and they both almost reached the minimum of the function. From this initial point, we avoid the saddle point, so GD performed better than it did before. Also, the SGD method did not have a problem with it before, and also did not have a problem with this initial point.

MSSC Students: Choose two of the following three problems to solve.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function that is bounded below, and let f^* denote its minimum value, i.e., $f^* = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. Then such a function f is said to satisfy the *Polyak-Lojasiewicz (P-L) inequality* if for some constant $\mu > 0$ it holds that

$$f(\mathbf{x}) - f^* \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2 \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

Problem 4 (MSSC)**.** Prove that if f is L-smooth and satisfies the P-L inequality for a constant $\mu > 0$, then gradient descent iterates $\mathbf{x}_k$ with stepsize $t = 1/L$ satisfy

$$f(\mathbf{x}_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(\mathbf{x}_0) - f^*).$$

(Hint: Start by plugging in the choices $\mathbf{y} = \mathbf{x}_k = \mathbf{x}_{k-1} - \frac{1}{L}\nabla f(\mathbf{x}_{k-1})$ and $\mathbf{x} = \mathbf{x}_{k-1}$ into the Decent Lemma. If you get stuck, there are many proofs of this on the web; search for "P-L inequality". It is OK to use one of these as guide, but please put the proof in your own words and cite any resources you use.)

Solution 4. The descent Lemma

$$f(\mathbf{x}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

Then let $\mathbf{y} = \mathbf{x}_k = \mathbf{x}_{k-1} - \frac{1}{L}\nabla f(\mathbf{x}_{k-1})$ and $\mathbf{x} = \mathbf{x}_{k-1}$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) + \nabla f(\mathbf{x}_{k-1})^\top(\mathbf{x}_k - \mathbf{x}_{k-1}) + \frac{L}{2}\|\mathbf{x}_{k-1} - \mathbf{x}_k\|^2$$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) + \nabla f(\mathbf{x}_{k-1})^\top(\mathbf{x}_{k-1} - \frac{1}{L}\nabla f(\mathbf{x}_{k-1}) - \mathbf{x}_{k-1}) + \frac{L}{2}\|\mathbf{x}_{k-1} - \frac{1}{L}\nabla f(\mathbf{x}_{k-1}) - \mathbf{x}_{k-1}\|^2$$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) + \nabla f(\mathbf{x}_{k-1})^\top(-\frac{1}{L}\nabla f(\mathbf{x}_{k-1}) + \frac{L}{2}\| - \frac{1}{L}\nabla f(\mathbf{x}_{k-1})\|^2$$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) - \frac{1}{L}(\nabla f(\mathbf{x}_{k-1}))^\top(\nabla f(\mathbf{x}_{k-1})) + \frac{L}{2}(\frac{-1}{L})^2\|\nabla f(\mathbf{x}_{k-1})\|^2$$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) - \frac{1}{L}\|\nabla f(\mathbf{x}_{k-1})\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_{k-1})\|^2$$

$$f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) - \frac{1}{2L}\|\nabla f(\mathbf{x}_{k-1})\|^2$$

From the P-L inequality we have

$$f(\mathbf{x}) - f^* \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2$$

$$\Rightarrow 2\mu(f(\mathbf{x}_{k-1}) - f^*) \leq \|\nabla f(\mathbf{x}_{k-1})\|^2$$

Therefore $f(\mathbf{x}_k) \leq f(\mathbf{x}_{k-1}) - \frac{\mu}{L}(f(\mathbf{x}_{k-1}) - f^*) \leq f(\mathbf{x}_{k-1}) - \frac{1}{2L}\|\nabla f(\mathbf{x}_{k-1})\|^2$

$$\begin{aligned}
\Rightarrow f(\mathbf{x}_k) - f^* &\leq f(\mathbf{x}_{k-1}) - \frac{\mu}{L}(f(\mathbf{x}_{k-1}) - f^*) - f^* \\
f(\mathbf{x}_k) - f^* &\leq f(\mathbf{x}_{k-1}) - \frac{\mu}{L}f(\mathbf{x}_{k-1}) + \frac{\mu}{L}f^* - f^* \\
f(\mathbf{x}_k) - f^* &\leq (1 - \frac{\mu}{L})(f(\mathbf{x}_{k-1}) - f^*)
\end{aligned}$$

Then if we apply this recursively as $k \to k-1 \to k-2 \to \cdots \to 0$ we have

$$f(\mathbf{x}_k) - f^* \leq (1 - \frac{\mu}{L})^k(f(\mathbf{x}_0) - f^*)$$

I used the following proof as a guide: https://angms.science/doc/NCVX/PolyakLojasiewiczIQ.pdf

Problem 5 (MSSC)**.** Prove that if f is μ-strongly convex then f satisfies the P-L inequality with constant μ. (Hint: Start from the definition of μ-strongly convex as

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

and minimize both sides over $\mathbf{y} \in \mathbb{R}^n$. If you get stuck, there are many proofs of this on the web. It is OK to use one of these as guide, but please put the proof in your own words and cite any resources you use.)

Solution 5. Since f is strongly convex, then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Let's minimize both sides with respect to y, first, the left hand side

$$min_y\{f(\mathbf{y})\} = f(\mathbf{x}^*)$$

Now, in order to minimize the right hand side, let's take the gradient, set it equal to 0 and solve for y. Let's start expanding to make taking the gradient easier:

$$\begin{aligned}
&f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top\mathbf{y} - \nabla f(\mathbf{x})^\top\mathbf{x} + &\frac{\mu}{2}\mathbf{y}^\top\mathbf{y} - \mu\mathbf{y}^\top\mathbf{x} + \frac{\mu}{2}\mathbf{x}^\top\mathbf{x}
\end{aligned}$$

Now taking the gradient with respect to y

$$\Rightarrow \nabla f(\mathbf{x})^\top + \mu\mathbf{y} - \mu\mathbf{x}$$

Set it equal to 0 and solve for y

$$\begin{aligned}\nabla f(\mathbf{x})^\top + \mu\mathbf{y} - \mu\mathbf{x} = 0 \\ \mu\mathbf{y} = \mu\mathbf{x} - \nabla f(\mathbf{x})^\top \\ \mathbf{y} = \mathbf{x} - \frac{\nabla f(\mathbf{x})^\top}{\mu}\end{aligned}$$

Then, this is our minimizer, so let's plug it to the original equation.

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{x} - \frac{\nabla f(\mathbf{x})^\top}{\mu} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \frac{\nabla f(\mathbf{x})^\top}{\mu} - \mathbf{x}\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(-\frac{\nabla f(\mathbf{x})^\top}{\mu}) + \frac{\mu}{2}\| - \frac{\nabla f(\mathbf{x})^\top}{\mu}\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}) - \frac{1}{\mu}\|\nabla f(\mathbf{x})^\top\|^2 + \frac{\mu}{2}(\frac{-1}{\mu})^2\|\nabla f(\mathbf{x})^\top\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}) - \frac{1}{\mu}\|\nabla f(\mathbf{x})^\top\|^2 + \frac{1}{2\mu}\|\nabla f(\mathbf{x})^\top\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})^\top\|^2 \\
f(\mathbf{x}^*) - f(\mathbf{x}) &\geq -\frac{1}{2\mu}\|\nabla f(\mathbf{x})^\top\|^2 \qquad \text{multiply by } -1 \\
f(\mathbf{x}) - f(\mathbf{x}^*) &\geq \frac{1}{2\mu}\|\nabla f(\mathbf{x})^\top\|^2
\end{aligned}$$

Which is exactly the same as the P-L inequality. Therefore, if f is μ-strongly convex then f satisfies the P-L inequality with constant μ.

Problem 6 (MSSC)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^2 + 3\sin^2(x)$. Note f the global minimum of f is $f^* = 0$ which is attained at $x = 0$. Show that f is *not* strongly convex, but that it does satisfy the P-L inequality for some $\mu > 0$. (Hint: To find a good choice of $\mu > 0$, it might help to first plot left and right sides of the P-L inequality on the same graph and plug in guesses of μ; however, the plot is not a proof.).