Math 4650/MSSC 5650 Final Exam, Spring 2023

Instructor: Greg Ongie

Due: Fri., May 12, 11:59 PM

Instructions:

1. You must explain all reasoning. It is not sufficient to just write the correct answer. Correct answers with no justification will receive no credit.

2. You must submit your completed exam to the D2L dropbox by 11:59pm on Friday, May 12. No late work will be accepted.

3. To speed up grading, **submit your exam as one ".pdf" file or one MS Word ".docx" file**. All code and plots need to be included in the single document.

4. **RULES FOR OUTSIDE SOURCES**
   It is permissible to use:

   (a) The textbook.

   (b) All resources on D2L, including posted lecture notes and supplemental notes, and HW solution guides.

   (c) *Your* course notes.

   (d) The internet for base-level questions: MATLAB syntax, basic definitions of terms, etc.

   (e) Me, via email, for clarification on the problem statement, but I will not give out hints for any of the problems.

   It is NOT permissible to consult:

   (a) Any classmates, other students, other professors, etc.

   (b) Anybody else's course notes (it is okay if you had previously borrowed someone else's notes because you missed a day, etc).

   (c) The internet for anything remotely approaching the actual question asked.

   - For example, it is okay to search the internet for "How to compute eigenvalues in MATLAB", but it is not okay to search the internet for the actual question.
   - If you are in doubt about what is okay, email me!

Any instances of plagiarism, working with classmates, or other cheating, will result in a score of 0 for the entire final exam.

This exam has five problems and is worth 100 points.

- **Math 4650 Students**: Complete the first four problems. Each problem is worth 25 points. Problem 5 may be attempted for up to 10 points extra credit.
- **MSSC 5650 Students**: Complete all five problems. Each problem is worth 20 points.

Problem 1. Find all critical points of the following functions and classify them according to whether they are local/global minimizers, local/global maximizers, or saddle points.

(a) $f(x, y) = 8x + 12y + x^2 - 2y^2$

(b) $f(x, y) = 100(y - x^2)^2 + (1 - x)^2$

(c) $f(x, y) = -\ln(1-x-y) - \ln(x) - \ln(y)$ over its domain $\{(x, y) : x > 0, y > 0, x + y < 1\}$.

Solution 1. (a)

$$\nabla f(x, y) = \begin{bmatrix} 8 + 2x \\ 12 - 4y \end{bmatrix}, \nabla^2 f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

To find critical points, let's set $\nabla f(x, y) = 0$

$$\begin{aligned} 8 + 2x = 0 &\Rightarrow x = -4 \\ 12 - 4y = 0 &\Rightarrow y = 3 \end{aligned}$$

So the point $(-4, 3)$ is the only critical point. Since $\nabla^2 f(x, y)$ is indefinite, then the point $(-4, 3)$ is a saddle point.

(b)

$$\nabla f(x, y) = \begin{bmatrix} 200(y - x^2)(-2x) - 2(1 - x) \\ 200(y - x^2) \end{bmatrix} = \begin{bmatrix} -400xy + 400x^3 - 2 + 2x \\ 200y - 200x^2 \end{bmatrix}$$

$$\nabla^2 f(x, y) = \begin{bmatrix} -400y + 1200x^2 + 2 & -400x \\ -400x & 200 \end{bmatrix}$$

To find critical points, set $\nabla f(x, y) = 0$

$$\begin{aligned} -400xy + 400x^3 - 2 + 2x &= 0 \\ 200y - 200x^2 &= 0 \end{aligned}$$

Let's look at the second equation first

$$200(y - x^2) = 0 \Rightarrow y - x^2 = 0 \Rightarrow y = x^2$$

Then plugging it to the first equation

$$-400x(x^2) + 400x^3 - 2 + 2x = 0 \Rightarrow 2x - 2 = 0 \Rightarrow x = 1$$

And $y = 1$. So the point $(1, 1)$ is the only critical point. Let's evaluate the Hessian at that point

$$\nabla^2 f(1,1) = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}$$

This Hessian has eigenvalues $\lambda_1 \approx 10001.6$ and $\lambda_2 \approx 0.3994$, which are both positive, therefore $\nabla^2 f(1,1) \succ 0$ and the point $(1,1)$ is at least a strict local minimizer. Let's check to see if it is a global minimizer.

Observe that $f(x,y) \geq 0$ for any (x,y) and $f(1,1) = 0$. Therefore, the point $(1,1)$ is a strict global minimizer.

(c)

$$\nabla f(x,y) = \begin{bmatrix} \frac{1}{1-x-y} - \frac{1}{x} \\ \frac{1}{1-x-y} - \frac{1}{y} \end{bmatrix}$$

$$\nabla^2 f(x,y) = \begin{bmatrix} \frac{1}{(1-x-y)^2} + \frac{1}{x^2} & \frac{1}{(1-x-y)^2} \\ \frac{1}{(1-x-y)^2} & \frac{1}{(1-x-y)^2} + \frac{1}{y^2} \end{bmatrix}$$

Then, let's find the critical points. By setting the gradient equal to 0.

$$\begin{bmatrix} \frac{1}{1-x-y} - \frac{1}{x} = 0 \\ \frac{1}{1-x-y} - \frac{1}{y} = 0 \end{bmatrix}$$

Start with solving the first equation for x

$$\frac{1}{(1-x-y)} - \frac{1}{x} = 0 \Rightarrow \frac{1}{(1-x-y)} = \frac{1}{x}$$
$$\Rightarrow 1 - x - y = x \Rightarrow x = \frac{1-y}{2}$$

Plug x in to the second equation

$$\frac{1}{(1 - \frac{1-y}{2} - y)} - \frac{1}{y} = 0 \Rightarrow \frac{1}{(1 - \frac{1-y}{2} - y)} = \frac{1}{y} \Rightarrow 1 - \frac{1-y}{2} - y = y$$
$$\Rightarrow -\frac{1-y}{2} = 2y - 1 \Rightarrow 3y = 1 \Rightarrow y = \frac{1}{3}$$

And $x = \frac{1-\frac{1}{3}}{2} = \frac{1}{3}$

So the point $(\frac{1}{3}, \frac{1}{3})$ is the only critical point. Now, let's ecaluate the Hessian at that point

$$\nabla^2 f(x,y) = \begin{bmatrix} \frac{1}{(1-\frac{1}{3}-\frac{1}{3})^2} + \frac{1}{\frac{1}{3}^2} & \frac{1}{(1-\frac{1}{3}-\frac{1}{3})^2} \\ \frac{1}{(1-\frac{1}{3}-\frac{1}{3})^2} & \frac{1}{(1-\frac{1}{3}-\frac{1}{3})^2} + \frac{1}{\frac{1}{3}^2} \end{bmatrix} = \begin{bmatrix} 9+\frac{1}{9} & 9 \\ 9 & 9+\frac{1}{9} \end{bmatrix}$$

Which has eigenvalues $\lambda_1 = \frac{37}{2}$ and $\frac{1}{2}$, and since both eigenvalues are positive, then $\nabla^2 f(\frac{1}{3}, \frac{1}{3}) \succ 0$. Which means that the point is at least a strict local minimimizer, let's check to see if it is a global minimizer.

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \frac{1}{(1-x-y)^2} + \frac{1}{x^2} & \frac{1}{(1-x-y)^2} \\ \frac{1}{(1-x-y)^2} & \frac{1}{(1-x-y)^2} + \frac{1}{y^2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} =$$

$$x^2(\frac{1}{(1-x-y)^2} + \frac{1}{x^2}) + 2y(\frac{1}{(1-x-y)^2})x + y^2(\frac{1}{(1-x-y)^2} + \frac{1}{y^2}) =$$

$$\frac{x^2}{(1-x-y)^2} + 1 + 2\frac{xy}{(1-x-y)^2} + \frac{y^2}{(1-x-y)^2} + 1 > 0$$

And this holds for all $\{(x,y) : x > 0, y > 0, x + y < 1\}$.

Since we are only looking inside its domain. We can see that this point is in fact a global minimizer. Since it is the only critical point and the function is strictly convex inside its domain.

Problem 2. In `problem2.m` you are given a "mystery function" $f : \mathbb{R}^{500} \to \mathbb{R}$ to minimize. This script loads MATLAB functions `f` and `grad` that return values and gradients of f, respectively. The mystery function f has the following properties:

- f is continuously differentiable.
- f is L-smooth, with $L = 4$.
- The minimum of f is 0, i.e., $0 = \min_{\mathbf{x} \in \mathbb{R}^{500}} f(\mathbf{x})$.

Your task is to find the best approximate minimizer `x` you can using *any algorithm of your choosing* given a fixed budget of 1000 gradient evaluations. Here "best" is measured by how close the final value of `f(x)` is to 0 (the minimum of f).

Ground rules:

- The initial iterate must be $\mathbf{x}_0 = \mathbf{0}$ (in MATLAB `x = zeros(500,1)`).
- You need to verify that `grad` was called at most 1000 times. This is checked at the end of the script.

- There is no limit on how many times `f` can be called.

You will be graded according to the following criteria:

$0.4990 >$ `f(x)` ≥ 0.0100	up to 60% points possible
$0.0100 >$ `f(x)` ≥ 0.0075	up to 80% points possible
$0.0075 >$ `f(x)` ≥ 0.0050	up to 100% points possible
$0.0050 >$ `f(x)`	up to 100% points possible, plus 5 points extra credit

To get full credit your write-up must include:

- A printout of your MATLAB code.
- Final values of `f(x)` and `grad('count')` (these are automatically output by the script)
- A log plot of the cost (this is automatically output by the script)
- A short discussion (2-4 sentences) on why you chose the algorithm you did and the parameter choices you settled on.

Solution 2. Print out of the code:

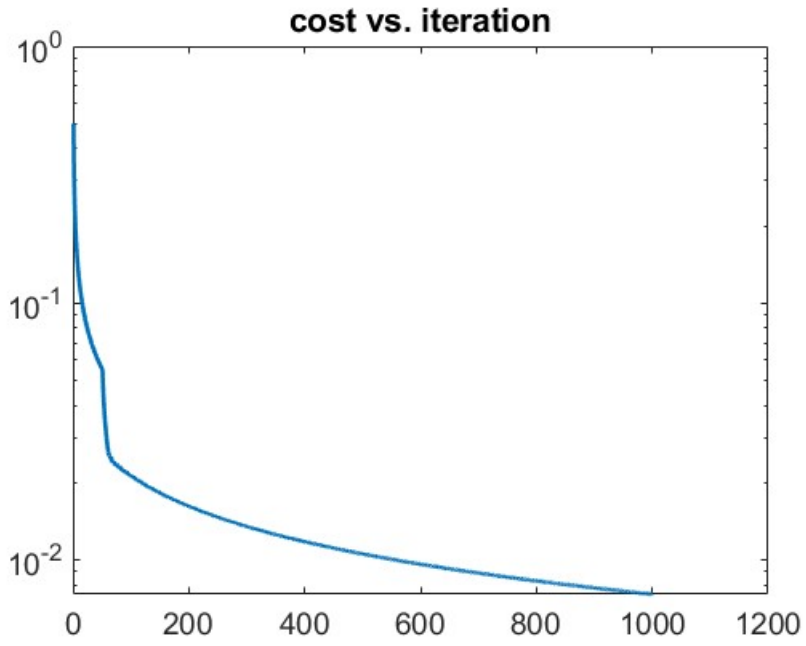
```
%note: the file mystery_function.p needs to be in the
%current folder for this script to work.
[f,grad] = mystery_function; %load mystery function and
   gradient
x = zeros(500,1); %fixed initialization

% define algorithm parameters here
s = 4;
alpha = 0.25;
beta = 0.552;
cost = f(x); %initialize cost array


for k=1:50
    g = grad(x);
    d = - g;
    t=1/4;
    x = x+t*d;
    cost = [cost,f(x)];
end
for k=51:1000
    g = grad(x);
    %your algorithm goes here
    if norm (g) < 1e-16
```

```
        break ;
    end
    d = - g; % steepest descent
    t = s;
    while (f(x) - f(x+t*d) < -alpha*t*g'*g)
        t = beta*t;
    end
    x = x+t*d;

    cost = [cost,f(x)];
end
fprintf('f(x) = %8.6f\n',f(x));
fprintf('number of gradient calls = %d\n',grad('count'));

figure(1); %plot the cost
semilogy(cost,'linewidth',2);
title('cost vs. iteration');
set(gca,'fontsize',14);
```

Final values of `f(x)` and `grad('count')`:

```
f(x) = 0.007353
number of gradient calls = 1000
```

Log plot:



I chose to go with a combination of the gradient descent with constant step size and gradient descent with backtracking line search. The reason why I chose to combine these two methods is because when I started testing with only backtracking search, around the first 30 iterations it was not decreasing the cost function or decreasing too slowly, but after that it would go back converge to smaller values faster. Also, when trying only the constant step size it would not converge fast enough, but the first 50 iterations seemed pretty good. Then I combined them, by doing 50 iterations with a constant step size and the other 950 with backtracking line search. In order to choose my parameters I tested a bunch of times increasing and decreasing by small amounts too see which one would give the best result.

Problem 3. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^4 + y^4$. Note f has a unique global minimum at $(x, y) = (0, 0)$. This problem will have you examine the iterates of gradient descent with exact line search for minimizing this function.

(a) Derive a formula for the *exact line-search* step-size, i.e., given any point $(x, y) \in \mathbb{R}^2$ with $(x, y) \neq (0, 0)$ find a formula for the step-size t^* given by

$$t^* = \arg\min_{t \geq 0} f((x, y) - t\nabla f(x, y))$$

(b) Implement gradient descent with exact line search in MATLAB for this function by modifying the provided script `problem3.m`. Use the following settings: start from the initial iterate $(x_0, y_0) = (0.5, 1)$, and use the exit condition $\|\nabla f(x_k, y_k)\| \leq 10^{-16}$ (these are the default settings in the script).

In your write-up, include the following:

- A derivation of the exact line-search rule in part (a).
- A printout of your MATLAB code from part (b).
- A plot of the trajectory of the iterates output (code to make this plot is included in `problem3.m`).
- A few sentences (2-3) describing the trajectory of the iterates and a possible explanation of why they look the way they do.

Solution 3. (a)

$$\nabla f(x, y) = \begin{bmatrix} 4x^3 \\ 4y^3 \end{bmatrix}$$

$$\begin{aligned} t^* &= argmin_{t \geq 0} f((x, y) - t\nabla f(x, y)) \\ &= argmin_{t \geq 0} (x - 4tx^3)^4 + (y - 4ty^3)^4 \end{aligned}$$

Let's call $g(t) = (x - 4tx^3)^4 + (y - 4ty^3)^4$, then take its derivative

$$g'(t) = -16(x - 4tx^3)^3x^3 - 16(y - 4ty^3)^3y^3$$

and set it equal to 0

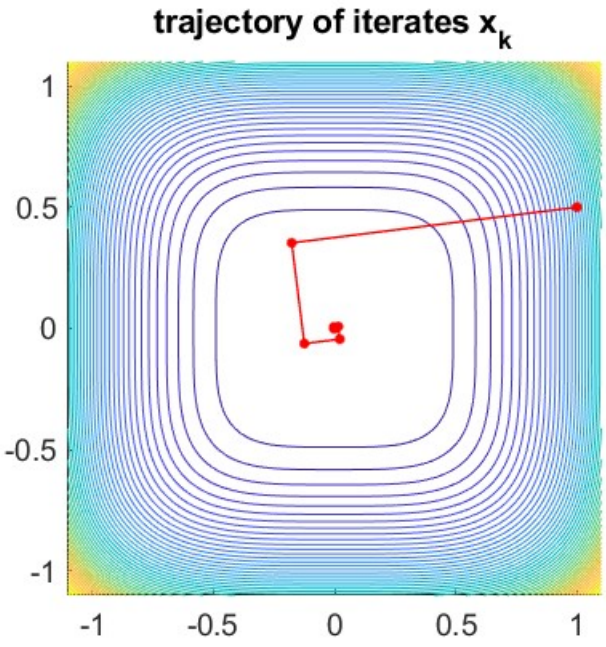
$$\begin{aligned}
-16(x - 4tx^3)^3x^3 - 16(y - 4ty^3)^3y^3 &= 0 \\
(x - 4tx^3)^3x^3 + (y - 4ty^3)^3y^3 &= 0 \\
(x - 4tx^3)^3x^3 &= -(y - 4ty^3)^3y^3 \\
((x - 4tx^3)^3x^3)^{\frac{1}{3}} &= (-(y - 4ty^3)^3y^3)^{\frac{1}{3}} \\
(x - 4tx^3)x &= -(y - 4ty^3)y \\
x^2 - 4yx^4 &= -y^2 + 4ty^4 \\
4tx^4 + 4ty^4 &= x^2 + y^2 \\
t(4x^4 + 4y^4) &= x^2 + y^2 \\
t &= \frac{x^2 + y^2}{(4x^4 + 4y^4)}
\end{aligned}$$

(b) Print out of code:

```
f = @(x) x(1).^4+x(2).^4;
grad = @(x) [4*x(1).^3;4*x(2).^3]; %define grad f(x)

x = [1;0.5];        %inital point
cost = f(x);        %initialize cost array
xar = x;            %initialize array of iterates (for
   plotting)
tol = 1e-16;        %exit tolerance
for k=1:100
    t = (x(1).^2+x(2).^2)/(4*x(1).^4+4*x(2).^4);
    x = x - t*grad(x); %gradient descent

    cost = [cost,f(x)]; %store f(x_k) for plotting
    xar = [xar,x];

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

```
% plot trajectory of iterates
figure(1);cla;
[X,Y] = meshgrid(linspace(-1.1,1.1,100),linspace
   (-1.1,1.1,100));
Z = zeros(size(X));
for i=1:length(X(:))
    Z(i) = f([X(i),Y(i)]);
end
hold on
contour(X,Y,Z,50);
plot(xar(1,:),xar(2,:),'.-r','markersize',15,'linewidth',1);
axis square
hold off;
set(gca,'fontsize',14);
title('trajectory of iterates x_k');
```

Plot of trajectory:



As expected the trajectory of the iterates have a zig zag pattern. This happens because of the iterates being orthogonal to the previous one. Since we are going on the a descent direction up to the exact point where the function is minimized, then the next step has to be in a direction that is not a linear combination of the previous iterate, the orthogonal direction.

Problem 4. The *Gauss-Newton* method is an approximation to Newton's method for solving non-linear least squares problems of the form

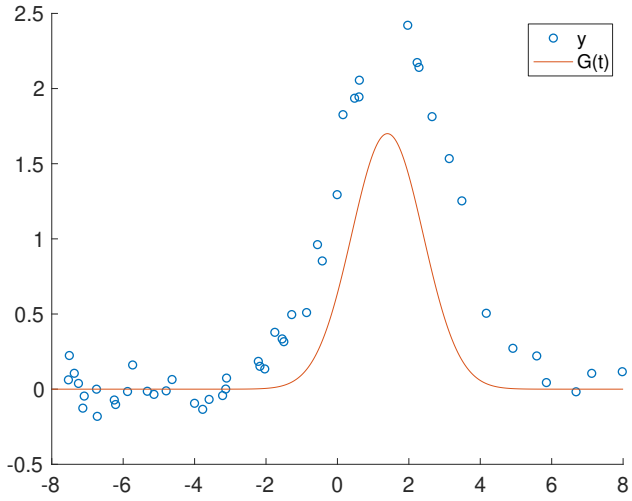
$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{g}(\mathbf{x}) - \mathbf{y}\|^2$$

where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector-valued function and $\mathbf{y} \in \mathbb{R}^m$ is a constant vector. The Gauss-Newton iterations are given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\mathbf{J}_\mathbf{g}(\mathbf{x}_k)^\top \mathbf{J}_\mathbf{g}(\mathbf{x}_k) + \varepsilon \mathbf{I})^{-1} \mathbf{J}_\mathbf{g}(\mathbf{x}_k)^\top (\mathbf{g}(\mathbf{x}_k) - \mathbf{y})$$

where $\mathbf{J}_\mathbf{g}(\mathbf{x}_k)$ is the Jacobian of $\mathbf{g}$ evaluated at the kth iterate $\mathbf{x}_k$, and $\varepsilon > 0$ is a parameter close to zero that ensures the matrix being inverted is nonsingular.

In this problem you will use the Gauss-Newton method to find the best fit of a Gaussian curve of the form $G(t) = a\exp(-b(t-c)^2)$ to scattered data points $(t_1, y_1), ..., (t_m, y_m)$. Note $G(t)$ depends on three parameters: a is the height of the Gaussian, b determines its width, and c determines its center. An illustration is shown below, where the data points (t_i, y_i) are shown in blue, and the curve $G(t)$ is shown in red.



We can set this up as a non-linear least-squares problem where the optimization variable is $\mathbf{x} = (a, b, c) \in \mathbb{R}^3$ and $\mathbf{g}(a, b, c) \in \mathbb{R}^m$ is the vector of evaluations of the Gaussian curve with parameters a, b, c at the points $t_1, t_2, ..., t_m$, i.e.,

$$\mathbf{g}(a, b, c) = \begin{bmatrix} a\exp(-b(t_1 - c)^2) \\ a\exp(-b(t_2 - c)^2) \\ \vdots \\ a\exp(-b(t_m - c)^2) \end{bmatrix},$$

and $\mathbf{y} = (y_1, ..., y_m)$ is the vector of data values to be fit.

(a) Determine the Jacobian matrix $\mathbf{J}_\mathbf{g}(a, b, c)$ for the function $\mathbf{g}(a, b, c)$ above (Note: this matrix has dimensions $m \times 3$).

(b) Using the script `problem4.m` as a staring point, define the Jacobian of $\mathbf{g}$ in MATLAB and implement the Gauss-Newton algorithm. You may set $\varepsilon = 10^{-8}$.

(c) Starting from the initialization $(a, b, c) = (0, 0, 0)$, run 20 iterations of the Gauss-Newton method to find the best fit parameters a, b, c.

(d) Plot the resulting Gaussian curve $G(t)$ with the optimal values of a, b, c on top of the scattered data.

In your write-up, include the following:

- A derivation of the Jacobian found in part (a).
- A printout of your MATLAB code from part (b).
- The best-fit parameters a, b, c that you found.
- The plot requested in part (d).

Solution 4. (a)

$$\mathbf{J_g}(a,b,c) = \begin{bmatrix} \frac{\partial g_1(a,b,c)}{\partial a} & \frac{\partial g_1(a,b,c)}{\partial b} & \frac{\partial g_1(a,b,c)}{\partial c} \\ \frac{\partial g_2(a,b,c)}{\partial a} & \frac{\partial g_2(a,b,c)}{\partial b} & \frac{\partial g_2(a,b,c)}{\partial c} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_m(a,b,c)}{\partial a} & \frac{\partial g_m(a,b,c)}{\partial b} & \frac{\partial g_m(a,b,c)}{\partial c} \end{bmatrix} = \begin{bmatrix} (\nabla g_1(a,b,c))^\top \\ (\nabla g_2(a,b,c))^\top \\ \vdots \\ (\nabla g_m(a,b,c))^\top \end{bmatrix}$$

Let's look at the general term $g_i(a, b, c)$ and take its gradient $g_i(a, b, c) = ae^{-b(t_i-c)^2}$

$$\nabla g_i(a,b,c) = \begin{bmatrix} e^{-b(t_i-c)^2} \\ ae^{-b(t_i-c)^2}(-(t_i-c)^2) \\ ae^{-b(t_i-c)^2}(-2b(t_i-c))(-1) \end{bmatrix}$$

$$(\nabla g_i(a,b,c))^\top = \begin{bmatrix} e^{-b(t_i-c)^2} & -ae^{-b(t_i-c)^2}(t_i-c)^2 & 2bae^{-b(t_i-c)^2}(t_i-c) \end{bmatrix}$$

Then the Jacobian is

$$J_g(a,b,c) = \begin{bmatrix} e^{-b(t_1-c)^2} & -ae^{-b(t_1-c)^2}(t_1-c)^2 & 2bae^{-b(t_1-c)^2}(t_1-c) \\ e^{-b(t_2-c)^2} & -ae^{-b(t_2-c)^2}(t_2-c)^2 & 2bae^{-b(t_2-c)^2}(t_2-c) \\ \vdots & \vdots & \vdots \\ e^{-b(t_m-c)^2} & -ae^{-b(t_m-c)^2}(t_m-c)^2 & 2bae^{-b(t_m-c)^2}(t_m-c) \end{bmatrix}$$

(b)

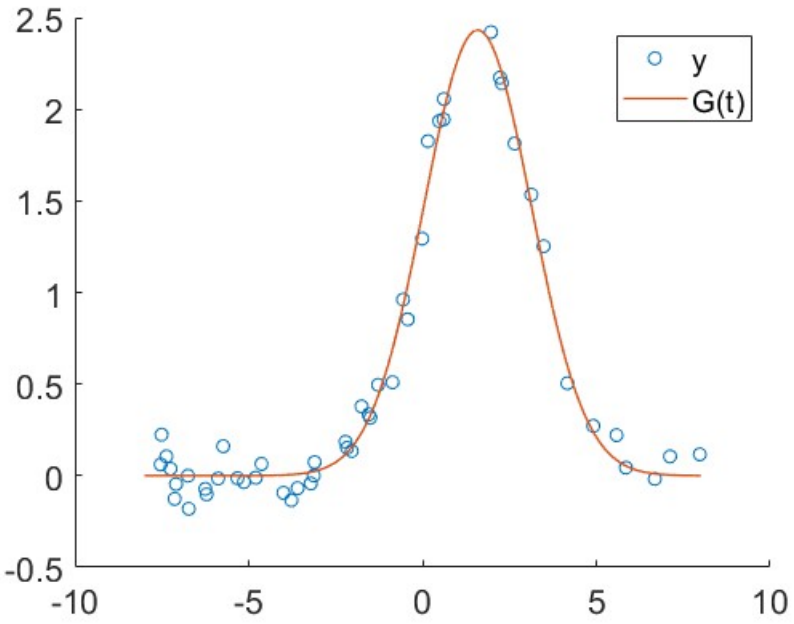
```
load problem4_data.mat
%loads variables t and y, both are m-by-1 vectors with m=50.
g = @(t,a,b,c) a*exp(-b*(t-c).^2); %define vector-valued
   function g(x)

a = 0;
b = 0;
c = 0;
x = [a;b;c]; %initialize optimization variable
epsilon = 1e-8;
for k=1:20
    a = x(1);
    b = x(2);
    c = x(3);

    r = g(t,a,b,c) - y; %residual vector

    % fill in rest of the algorithm
    % J = ... define the Jacobian
    J = [exp(-b*(t-c).^2);
        -a*(t-c).^2.*exp(-b*(t-c).^2);
        2*b*a*(t-c).*exp(-b*(t-c).^2)];
    % The line above was generating a 150x1 vector
    % But it should be 50x3, so reshape it
    J = reshape(J,[50,3]);
    n = size(J);
    n = n(2);
    % x = ... define the x update
    x = x - (J'*J+epsilon*eye(n))\(J'*(g(t,a,b,c)-y));
end

% plot for part (d)
a = x(1);
b = x(2);
c = x(3);
figure(1);
cla
hold all
scatter(t,y);
plot(-8:0.01:8,g(-8:0.01:8,a,b,c),'linewidth',1);
hold off
legend('y','G(t)');
```

```
set(gca,'fontsize',16);
```

(c) Best fit parameters:

```
a=2.43187186
b=0.20981714
c=1.58573637
```

(d) Plot of the Gaussian curve:



Problem 5 (MSSC)**.**

(a) A twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Prove that if f is convex then

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

(Hint: Start from Taylor's approximation theorem – see Theorem 1.24 in Beck)

(b) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, *convex*, and L-smooth. Assume that f has at least one global minimizer $\mathbf{x}^* \in \mathbb{R}^n$ and let $f^* = f(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$. Define $\{\mathbf{x}_k\}_{k \geq 0}$ to be the iterates obtained by running gradient descent with constant stepsize $t = 1/L$ starting from any initial point $\mathbf{x}_0 \in \mathbb{R}^n$. Show that if the iterates $\{\mathbf{x}_k\}_{k \geq 0}$ are bounded[1] then $f(\mathbf{x}_k) \to f^*$ as $k \to \infty$.

(Hint: You may use the fact that under these assumptions we have $\|\nabla f(\mathbf{x}_k)\| \to 0$ as $k \to 0$; see Theorem 4.25 in Beck.)

Solution 5. (a) From Taylor's approximation theorem, we have

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x})$$

So plugging this back to the original equation, we have

$$f(\mathbf{x}) + \nabla(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

$$f(\mathbf{x}) + \nabla(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geq 0$$

And this equation only holds if

$$\frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \geq 0$$

Let the vector $\mathbf{y} - \mathbf{x} = \mathbf{c}$, so

$$\frac{1}{2}\mathbf{c}^\top \nabla^2 f(\mathbf{z})\mathbf{c} \geq 0 \Rightarrow \mathbf{c}^\top \nabla^2 f(\mathbf{z})\mathbf{c} \geq 0$$

And since f is convex, this holds for all $\mathbf{c} \in \mathbb{R}^n, \mathbf{z} \in \mathbb{R}^n$, by definition of a positive semi definite matrix. Therefore $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ if f is convex.

[1] i.e., there exists a $B > 0$ such that $\|\mathbf{x}_k\| \leq B$ for all $k \geq 0$

(b) From the descent lemma for L-smooth functions

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

Now let's run gradient descent with stepsize $t = \frac{1}{L}$, and $\mathbf{x}_{k+1} = \mathbf{x}_k - t\nabla f(\mathbf{x}) = y, \mathbf{x} = \mathbf{x}_k$

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top)(-t\nabla f(\mathbf{x}_k)) + \frac{L}{2}\| - t\nabla f(\mathbf{x}_k)\|^2 \\
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -t\|f(\mathbf{x}_k)\|^2 + \frac{Lt^2}{2}\|\nabla f(\mathbf{x}_k)\|^2 \\
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -\frac{1}{L}\|f(\mathbf{x}_k)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \\
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2
\end{aligned}$$

We have $lim_{k\to\infty}(\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2) = 0$, then $lim_{k\to\infty} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) = 0$.
Since f is convex, then from part a we have:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)$$

From above $f(\mathbf{x}_{k+1}) + \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k)$, combining both equations gives:

$$\begin{aligned}
f(\mathbf{x}^*) &\geq f(\mathbf{x}_{k+1}) + \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 + \nabla f(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k) \\
f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) &\geq \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2 + \nabla f(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k) \\
f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) &\geq \frac{L}{2}\left(\|\frac{1}{L}\nabla f(\mathbf{x}_k)\|^2 + \frac{2}{L}f(\mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)\right) \\
f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) &\geq
\end{aligned}$$

$$\frac{L}{2}\left((\frac{1}{L}\nabla f(\mathbf{x}_k))^\top((\frac{1}{L}\nabla f(\mathbf{x}_k)) + \frac{2}{L}\nabla f(\mathbf{x}_k))^\top(\mathbf{x}^* - \mathbf{x}_k)\right) + (\mathbf{x}^* - \mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k) - (\mathbf{x}^* - \mathbf{x}_k)^\top(\mathbf{x}^* - \mathbf{x}_k)$$

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

Now using the gradient step again $\mathbf{x}_k = \mathbf{x}_{k+1} + \frac{1}{L}\nabla f(\mathbf{x}_k)$, then

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

Since $lim_{k\to\infty}\mathbf{x}_{k+1} = lim_{k\to\infty}\mathbf{x}_k$ then

$$lim_{k\to\infty} f(\mathbf{x}^*) - f(\mathbf{x}_{k+1}) \geq lim_{k\to\infty}\frac{L}{2}\left(-2(\mathbf{x}_{k+1})^\top(x^*) + \|\mathbf{x}_{k+1}\|^2 - \|\mathbf{x}_k\|^2 + 2(\mathbf{x}_k)^\top(\mathbf{x}^*)\right)$$
$$lim_{k\to\infty} f(\mathbf{x}^*) - f(\mathbf{x}_k) \geq lim_{k\to\infty}\frac{L}{2}\left(-2(\mathbf{x}_k)^\top(x^*) + \|\mathbf{x}_k\|^2 - \|\mathbf{x}_k\|^2 + 2(\mathbf{x}_k)^\top(\mathbf{x}^*)\right)$$
$$lim_{k\to\infty} f(\mathbf{x}^*) - f(\mathbf{x}_k) \geq 0$$

We showed that $lim_{k\to\infty} f(\mathbf{x}^*) - f(\mathbf{x}_k) \geq 0$ and $lim_{k\to\infty} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq 0$. Therefore $f(\mathbf{x}_k) \to f(\mathbf{x}^*)$ as $k \to \infty$