Note: Some of the problems will have slightly different rubrics on Gradescope; refer to the Gradescope rubric as the official rubric.

1a (1 Point)

1. 1 Point for correct scatter plot
    a. 0.5 Points for call to plt.scatter
    b. 0.5 Points for having x and y as the first two arguments of plt.scatter (ok if points are different size, marker argument not set, or no axes labels)

1b (1 Point)

1. 0.5 Points for positive or linear association between x and y
2. 0.5 Points for sinusoidal noise or something similar (i.e. bumps/steps)

2a (2 Points)

1. 0.5 Points for correct average L2 loss function
2. 0.5 Points for correct (or clear indication of correct) partial derivative with respect to theta
3. 0.5 Points for setting the derivative equal to 0 and attempting to solve for the optimal theta
4. 0.5 Points for correct optimal theta value

2c (1 Point)

1. 1 Point for correct average loss curve, vertical line, and axes labels
    a. 0.5 Points for correct vertical line on the optimal theta value (plotting a vertical line at x = find_theta(x, y) is sufficient for this rubric item regardless of correctness)
    b. 0.5 Points for correct average loss curve AND axes labels (a call to visualize(x, y, thetas) is not sufficient for this rubric item if it is incorrect - i.e. no axes labels)

2d (1 Point)

1. 1 Point for correct scatter plot and linear model line
    a. 0.5 Points for correct linear model line (a call of plt.plot with first argument x and second argument linear_model(x, find_theta(x, y)) is sufficient for this rubric item regardless of correctness)
    b. 0.5 Points for correct background scatter plot (a call to scatter(x, y) is sufficient for this rubric item regardless of correctness)

2e (1 Point)

1. 1 Point for correct residual plot with horizontal line at y=0
    a. 0.5 Points for correct residual plot
    b. 0.5 Points for horizontal line at y=0

2f (1 Point)

1. 1 Point for a reasonable explanation of a relationship between residuals and x based on their plot in 2e (even if 2e plot is incorrect)

3b (3 Points)

1. 1 Point for correct partial derivative with respect to theta_1
    a. 0.5 Points for pulling out the 2
    b. 0.5 Points for correct chain rule and no extra terms
2. 2 Points for correct partial derivative with respect to theta_2
    a. 0.5 Points for pulling out the 2
    b. 0.5 Points for correct 1st chain rule
    c. 0.5 Points for correct 2nd chain rule
    d. 0.5 Points for no extra terms

4c (1 Point)

1. 1 Point for correct answer (Yes)

4d (1 Point)

1. 1 Point for correct static and decaying learning rate plots
    a. 0.5 Points for static learning rate plot
    b. 0.5 Points for decaying learning rate plot

4e (1 Point)

1. 1 Point for mentioning that decaying learning rate converges faster

5a (2 Points)

1. 2 Points for some reasonable explanation regarding the differences between a static and decaying learning rate using the loss history and the 3D visualizations as support

5b (4 Points)

1. 2 Points for some reasonable interpretation of the contour plots
2. 2 Points for a reasonable comparison of the contour and 3D plots and discussing some pros and cons of each

6a (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist)
   a. 1 Point for using the algebraic definition of convexity anywhere in the proof
   b. 0.5 Points for starting on one of the sides of the inequality and rewriting $h(x) = f(x) + g(x)$ on that side
   c. 0.5 Points for appropriate use of convexity in f
   d. 0.5 Points for appropriate use of convexity in g
   e. 0.5 Points for rewriting $f(x) + g(x) = h(x)$ to complete the proof

6b (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist)
   a. 0.5 Points for converting from $||Xw - y||^2$ to $(Xw - y)^T(Xw - y)$
   b. 1 Point for $(Xw - y)^T = w^TX^T - y^T$
   c. 0.5 Points for correct FOIL
   d. 1 Point for noticing that $(w^TX^Ty) = (w^TX^Ty)^T = y^TXw$ since this term is a scalar

6c (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist - i.e. using gradient chain rule)
   a. 2 Points for correct gradient ($2X^TXw - 2X^Ty$)

i. 1 Point for correct gradient of first term
    1. 0.5 Points partial credit for minor errors
ii. 1 Point for correct gradient of second and third terms
    1. 0.5 Points partial credit for minor errors

b. 1 Point for solving for w correctly (w_hat = (X^TX)^(-1)X^Ty)
    i. 0.5 Points partial credit for minor errors

6d (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist)
    a. 1 Point for correct expanded loss function including regularization (same as before + lambda*w^T*w)
        i. 0.5 Points partial credit for minor errors
    b. 1 Point for correct gradient (same as before + 2 * lambda * w)
        i. 0.5 Points partial credit for minor errors
    c. 1 Point for solving for the correct optimal w (w_hat = (X^TX + lambda * I)^(-1)X^Ty)
        i. 0.5 Points partial credit for minor errors

6e (3 Points)

1. 1 Point for mentioning that ridge regression guarantees invertibility (ok if no thorough explanation)
2. 1 Point for mentioning that ridge regression helps us reduce variance
3. 1 Point for mentioning that ridge regression will increase bias

6f (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist)
    a. 1 Point for swapping the order of the two summations
    b. 0.5 Points for applying definition of conditional probability
    c. 0.5 Points for pulling out b and yP(y) appropriately
    d. 0.5 Points for recognizing that the inner sum is 1 since it is the sum of all possible values of a probability distribution
    e. 0.5 Points for recognizing that the outer sum is E[Y]

6g (3 Points)

1. 3 Points for correct proof (Note: alternate solutions that do not apply to this rubric will probably exist)
    a. 1 Points for appropriate use of Var(X) = E[X^2]-E[X]^2 at the beginning and end of the proof
        i. 0.5 Points partial credit for minor errors
    b. 1 Point for appropriate use of linearity of expectation with the expansion of squares in the middle of the proof
        i. 0.5 Points partial credit for minor errors
    c. 1 Point for using E[XY] = E[X]E[Y] from independence
        i. 0.5 Points partial credit for minor errors