James Zhao
HW5
A15939512

1 Q1

$$L(z) = \sum_{i=1}^{n} \|x^{(i)} - z\|^2$$

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{n} \frac{\partial}{\partial z_j} \|x^{(i)} - z\|^2 = \sum_{i=1}^{n} \frac{\partial}{\partial z_j} (x_1^{(i)} - z_1)^2 + ... + \frac{\partial}{\partial z_j} (x_j^{(i)} - z_j)^2 + ...$$

$$= \sum_{i=1}^{n} 0 + 2(x_j^{(i)} - z_j)(-1) + 0 = -2 \sum_{i=1}^{n} x_j^{(i)} - z_j$$

$$= \nabla = \begin{bmatrix} \frac{\partial L}{\partial z_1} \\ \vdots \\ \frac{\partial L}{\partial z_d} \end{bmatrix} = -2 \sum_{i=1}^{n} x^{(i)} - z$$

Solve for gradient = 0 (and cancel the -2):

$$\sum_{i=1}^{n} x^{(i)} - \sum_{i=1}^{n} z = 0$$

$$\sum_{i=1}^{n} x^{(i)} = \sum_{i=1}^{n} z = nz$$

$$z = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$$

2 2

2.1 a

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial}{\partial w_j}\left[(w \cdot x^{(i)})\right] + \frac{\partial}{\partial w_j}\frac{1}{2}c\|w\|^2$$

$$= \sum_{i=1}^{n} x_j^{(i)} + cw_j$$

$$\nabla L(w) = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix} = \sum_{i=1}^{n} x^{(i)} + cw$$

2.2 b

$$\nabla L(w) = \sum_{i=1}^{n} x^{(i)} + cw = 0$$

$$cw = -\sum_{i=1}^{n} x^{(i)}$$

$$w = -\frac{1}{c}\sum_{i=1}^{n} x^{(i)}$$

3 3

3.1 a

$$\nabla L(w) = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_4} \end{bmatrix} = \begin{bmatrix} 2w_1 + 2 \\ 4w_2 - 4 \\ 2w_3 - 2w_4 \\ 2w_4 - 2w_3 \end{bmatrix}$$

3.2 b

$$w_{t+1} = w_t - \eta \nabla L(w_t)$$

$$w_{t+1} = 0 - \eta \nabla L(0)$$

$$= \begin{bmatrix} 2 \\ -4 \\ 0 \\ 0 \end{bmatrix}$$

3.3 c

$$\nabla L(w) = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_4} \end{bmatrix} = \begin{bmatrix} 2w_1 + 2 \\ 4w_2 - 4 \\ 2w_3 - 2w_4 \\ 2w_4 - 2w_3 \end{bmatrix} = 0$$

$$w_1 = -1, w_2 = 1, w_3 = w_4$$

$$L((-1, 1, x, x)) = 1 + 2 + x^2 - 2x^2 + x^2 + 2(-1) - 4 + 4$$

$$L((-1, 1, x, x)) = 1$$

Minimum Value = **1**

3.4 d

There is not a unique solution. A solution exists for all $w = (-1, 1, x, x)$.

4 4

4.1 a

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial}{\partial w_j}\Big[y^{(i)} - w \cdot x^{(i)}\Big]^2 + \frac{\partial}{\partial w_j}\lambda\|w\|^2$$

$$= \sum_{i=1}^{n} -2\Big[y^{(i)} - w \cdot x^{(i)}\Big]x_j^{(i)} \quad + 2\lambda w_j$$

$$\nabla L(w) = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix} = -2\sum_{i=1}^{n}\Big[y^{(i)} - w \cdot x^{(i)}\Big]x^{(i)} + 2\lambda w$$

4.2 b

$$w_{t+1} = w_t - \eta_t \nabla L(w_t)$$

$$w_{t+1} = w_t - \eta_t\Big\{\; -2\sum_{i=1}^{n}\Big[y^{(i)} - w \cdot x^{(i)}\Big]x^{(i)} + 2\lambda w \;\Big\}$$

4.3 c

Update Equation:

$$w_{t+1} = w_t - \eta_t \nabla l(w_t; x^{(i)}, y^{(i)})$$

$$w_{t+1} = w_t - \eta_t\Big\{-2\Big[y^{(i)} - w \cdot x^{(i)}\Big]x^{(i)} + 2\lambda w\Big\}$$

Algorithm:
$w_0 = 0$
Cycle through points $(x^{(i)}, y^{(i)})$ until some stopping condition:

- $w_{t+1} = w_t - \eta_t\Big\{-2\Big[y^{(i)} - w_t \cdot x^{(i)}\Big]x^{(i)} + 2\lambda w_t\Big\}$

5 5

5.1 a

$\nabla^2 f(x) = 2 \geq 0$, convex.

5.2 b

$\nabla^2 f(x) = -2 \leq 0$, concave.

5.3 c

$\nabla^2 f(x) = 2 \geq 0$, convex.

5.4 d

$\nabla^2 f(x) = 0 = 0$, both.

5.5 e

$\nabla^2 f(x) = 6x$, hessian can change signs, neither.

5.6 f

$\nabla^2 f(x) = 12x^2 \geq 0$, convex.

5.7 g

$\nabla f(x) = \frac{1}{x} = x^{-1}, \nabla^2 f(x) = -x^{-2} \leq 0$, concave.

6 6

6.1 a

For my coordinate descent method, I will repeatedly call a routine until the gradient magnitude is less than a certain threshold. For this routine:

- (1) Initialize the weight vector to a d-dimensional vector sampled from a standard normal (seeded for reproducability)
- (2) while the gradient magnitude is greater than a certain value:
  - (3) Compute the gradient
  - (4) (i) Pick the coordinate with the largest absolute value. This, intuitively, is the component-direction of steepest ascent (and thus the negative is the direction of steepest descent)
  - (5) (ii) Update this coordinate by the magnitude of the gradient. Thus, when the model is close to optimized, the step size will naturally decrease in the corresponding direction.
    $w[\text{idx}] = w[\text{idx}] - \eta * sign(\nabla[\text{idx}]) * norm(\nabla)$

Because this method depends on the gradient, the function must be defined and differentiable at every point in R^d.

6.2 b