Gradients and Hessians of Linear/Quadratic Functions

- $f(w) = a^\top w$
  - $\nabla_w f(w) = a$
  - $\nabla_w^2 f(w) = 0$
- $f(w) = w^\top A w$
  - $\nabla_w f(w) = (A^\top + A)w$ ($= 2Aw$ if A symmetric)
  - $\nabla_w^2 f(w) = (A^\top + A)$ ($= 2A$ if A symmetric)
- See "Linear/Quadratic Gradients" on course page

Practice Problems

Consider the dataset below, which has 5 training examples and 2 features:

$$X = \begin{bmatrix} 2 & 1 \\ 10 & 4 \\ 10 & 9 \\ 5 & 8 \\ 8 & 10 \end{bmatrix}, \quad y = \begin{bmatrix} 8.0 \\ 6.5 \\ 4.0 \\ 7.5 \\ 2.0 \end{bmatrix}.$$

Suppose that you have the following test example:

$$\hat{x} = \begin{bmatrix} 9 & 4 \end{bmatrix}.$$

(a) Suppose we use a linear regression model with coefficients given by

$$w = \begin{bmatrix} 2/3 \\ -1/4 \end{bmatrix}.$$

What prediction $\hat{y}$ would we make for the test example?

Practice Problems

Consider a linear regression problem with 5 features, with learned weights w and two test cases, $\tilde{x}_1$ and $\tilde{x}_2$, shown below:

$$w = \begin{bmatrix} 2 \\ 1300 \\ 3 \\ 3 \\ -5 \end{bmatrix}, \quad \tilde{x}_1 = \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix}, \quad \tilde{x}_2 = \begin{bmatrix} a+2 \\ b \\ c \\ d \\ e+1 \end{bmatrix}.$$

Using the above weights w, the prediction for the test vector $\tilde{x}_1$ is $\hat{y}_1 = 10$. Find the prediction for the test vector $\tilde{x}_2$.

Practice Problems

(a) Consider the following objective, which considers a weighted worst-case error with a penalty on the absolute value of the weights,

$$f(w) = \max_{i \in \{1,2,\ldots,n\}} \{v_i |w^T x_i - y_i|\} + \lambda \sum_{j=1}^{d} |w_j|,$$

where λ is a non-negative scalar. Re-write this objective function in matrix and norm notation. You can use V as a diagonal matrix with the elements v_i along the diagonal.

Practice Problems

(b) Consider the L2-regularized *tilted* least squares objective,

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2 + \sum_{j=1}^{d} v_j w_j,$$

where the λ is a non-negative scalar and the v_j are real-valued "tilting" variables. Write down a linear system whose solution minimizes this (convex and quadratic) objective function. You can use v as a vector containing the v_j values.