CS221 Problem Workout

Week 1

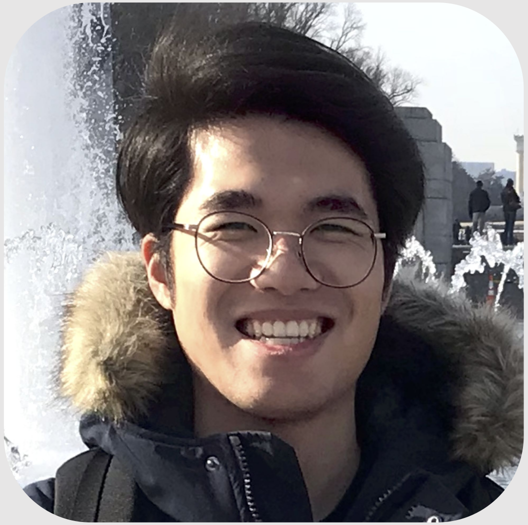
Stanford University

Introduction

Jenn Grannen

General OH: Mondays 3:00-4:30 Bytes
HW OH: Fridays 9:30-11:00 Huang Basement

Chen Wang



General OH: Fridays 3:00-4:30 Bytes
HW OH: Thursdays 10:30-12:00 Huang Basement

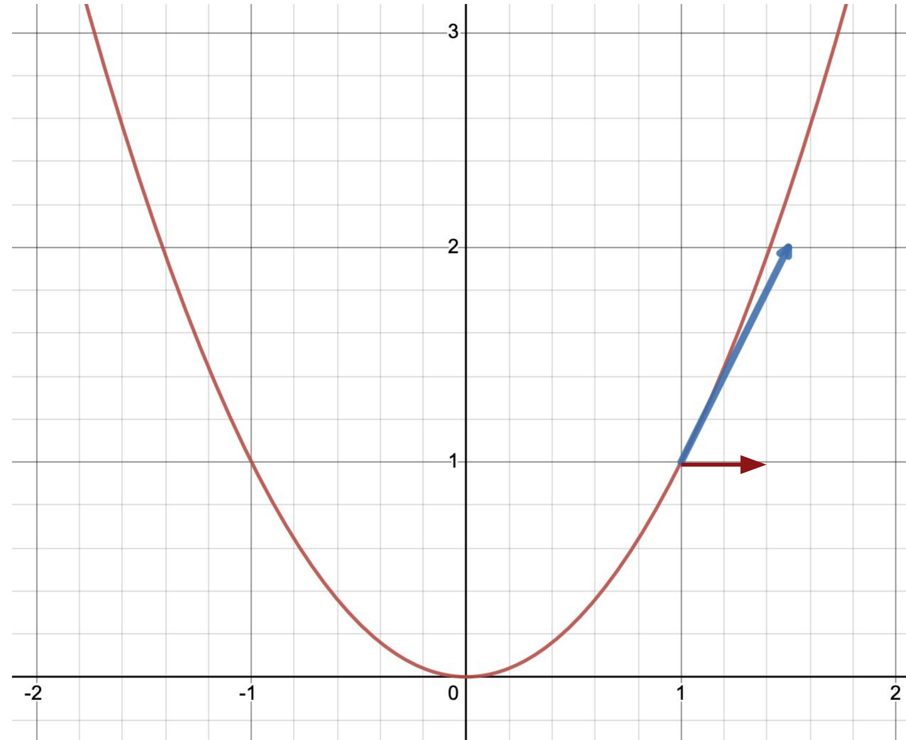
Computing the Gradient

- The gradient is the direction of greatest ascent
- With one variable it's the slope of the tangent line to the curve
- Example:

$$f(x) = x^2$$

$$\nabla f(x) = 2x$$

$$\nabla f(x)|_{x=1} = 2$$



Computing the Gradient

- How about for multiple dimensions?
- We have to take the derivative with respect to each variable.
- Example:

$$f(x, y) = x^2 y$$

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

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

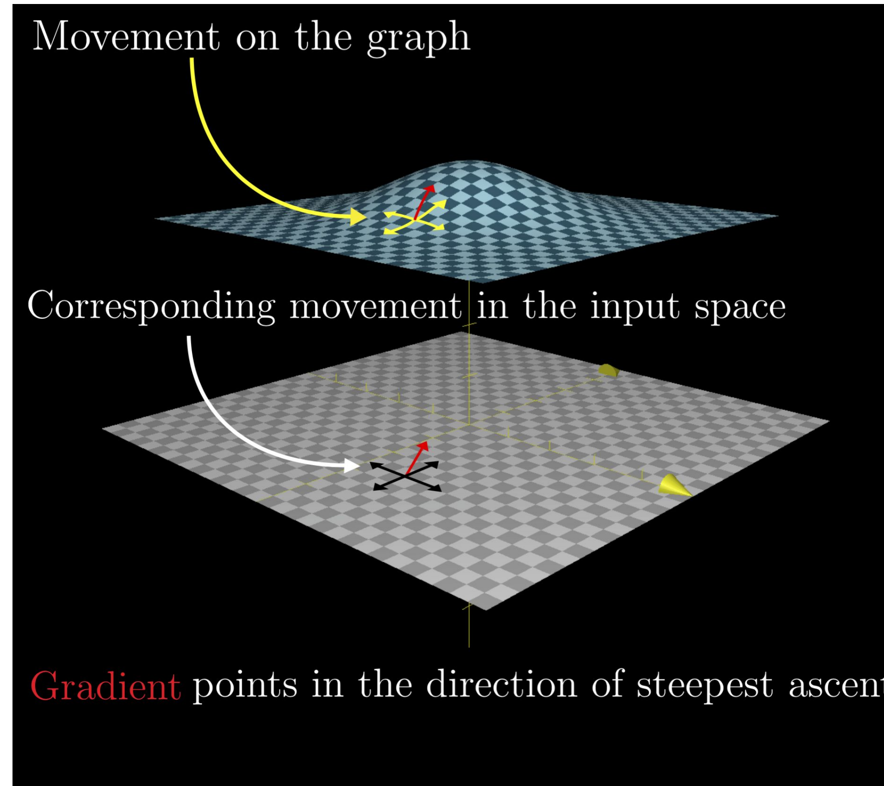


Image Credit: Khan Academy

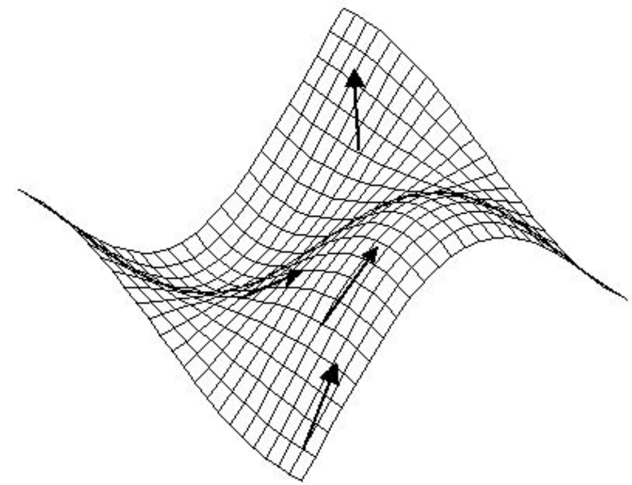
Computing the Gradient

- Sometimes we only care about the gradient with respect to a subset of variables.
- In this case we can treat the other variables as constants.
- Example:

  Compute the gradient with respect to w:

$$f(x, y, w) = \left(\frac{log(x)^{4y}}{x}\right) w$$

$$\nabla_w f(x, y, w) = \frac{log(x)^{4y}}{x}$$



Gradient Vectors Shown at Several Points on the Surface of cos(x) sin(y)

Image Credit: Saint John Fisher University

Preview: What is a Loss Function?

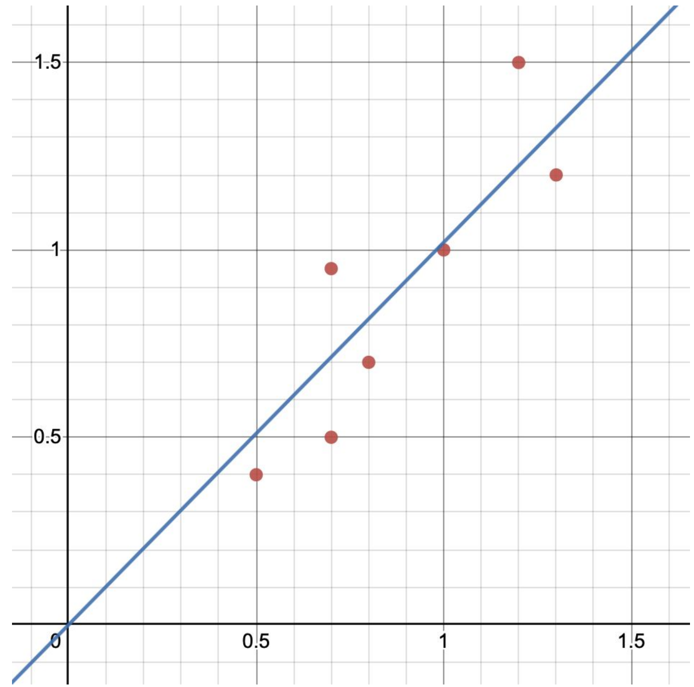
In Machine Learning we are finding functions that best approximate the mapping from inputs to outputs.

Example: Linear Regression

$$w = [w_0 \quad w_1]$$

$$f(x, w) = w_1 \cdot x + w_0$$

Want to find the best values of w_0 and w_1 such that f best fits the data points.

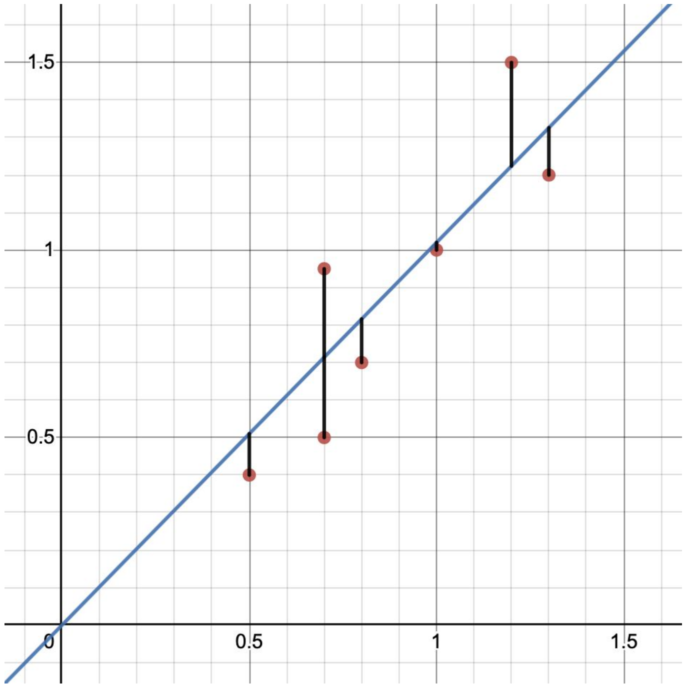


Preview: What is a Loss Function?

- How can we measure how good our current values of w are?
- Add up the (squared) distance between each data point and our current model prediction
- This is an example of a loss function

$$Loss(x, y, w) = (f_w(x) - y)^2$$

- Minimizing the Loss: demo



Problem 2

1) Problem 1: Gradient computation

(i) Let $\phi(x) : \mathbb{R} \mapsto \mathbb{R}^d$, $\mathbf{w} \in \mathbb{R}^d$, and $f(x, \mathbf{w}) = \mathbf{w} \cdot \phi(x)$. Consider the following loss function.

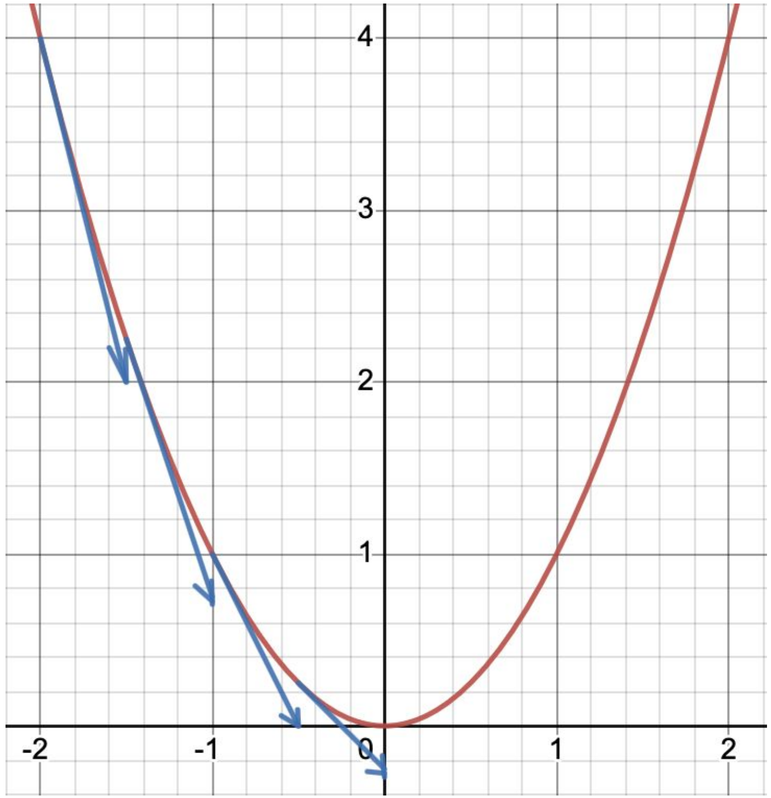
$$\text{Loss}(x, y, \mathbf{w}) = \frac{1}{2} \max\{2 - (\mathbf{w} \cdot \phi(x))y, 0\}^2. \tag{1}$$

Compute its gradient $\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w})$.

Gradient Descent

- How can we minimize the loss?
- The gradient points in the direction of steepest ***ascent***
- If we move in the opposite direction we go in the direction of steepest descent
- Gradient Descent Weight Updates:

$$w := w - \eta \nabla_w Loss(w)$$



Stochastic Gradient Descent

- Pick out random data points to use for our loss computation at each step instead of all data points

- Why?
  - More efficient
  - Can help escape shallow local minima

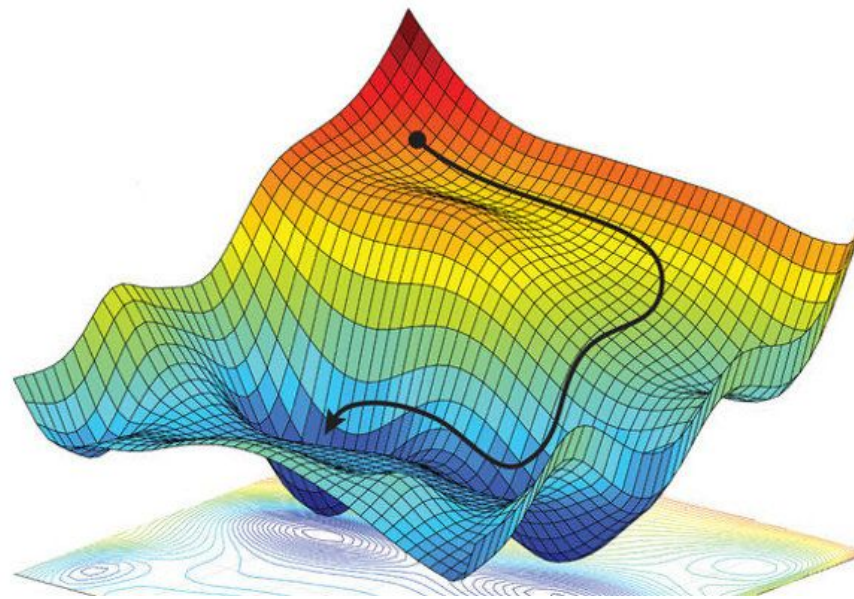
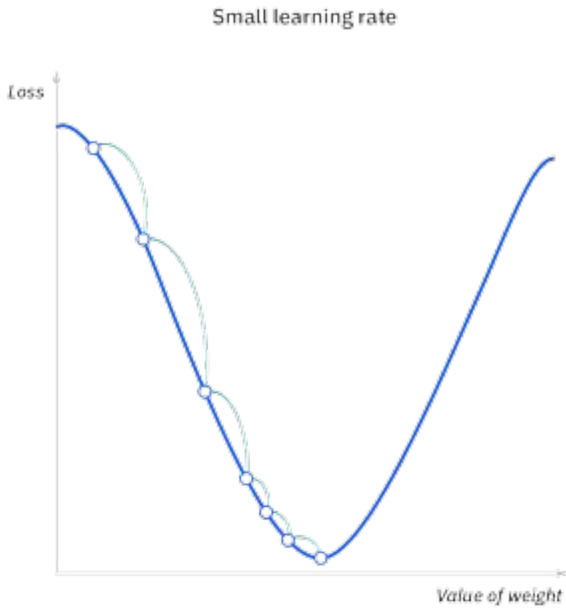


Image Credit: [Er Raqabi El Mehdi]

Step Size

$$w := w - \eta \nabla_w Loss(w)$$



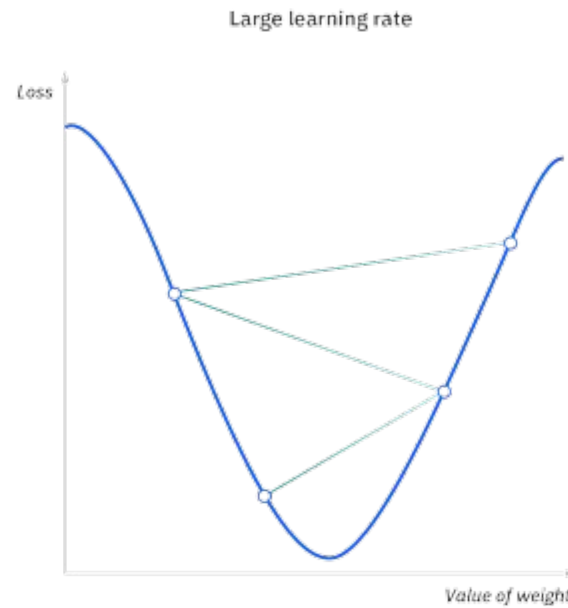


Image Credit: IBM

Problem 1 (i)

3) Problem 3: Gradient and Gradient Descent

(i) Let $\phi(x) : \mathbb{R} \mapsto \mathbb{R}^d$, $\mathbf{w} \in \mathbb{R}^d$. Consider the following loss function.

$$\text{Loss}(x, y, \mathbf{w}) = \begin{cases} 1 - 2(\mathbf{w} \cdot \phi(x))y & \text{if } (\mathbf{w} \cdot \phi(x))y \leq 0 \\ (1 - (\mathbf{w} \cdot \phi(x))y)^2 & \text{if } 0 < (\mathbf{w} \cdot \phi(x))y \leq 1 \\ 0 & \text{if } (\mathbf{w} \cdot \phi(x))y > 1, \end{cases}$$

where $y \in \mathbb{R}$. Compute the gradient $\nabla_{\mathbf{w}} \text{Loss}(x, y, \mathbf{w})$.

Problem 1 (ii)

(ii) Write out the Gradient Descent update rule for some function $\text{TrainLoss}(\mathbf{w}) : \mathbb{R}^d \mapsto \mathbb{R}$.

Problem 1 (iii)

(ii) Let $d = 2$ and $\phi(x) = [1, x]$. Consider the following training loss function.

$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{2}\Big(\text{Loss}(x_1, y_1, \mathbf{w}) + \text{Loss}(x_2, y_2, \mathbf{w})\Big). \tag{13}$$

Compute $\nabla_w \text{TrainLoss}(\mathbf{w})$ for the following values of $x_1, y_1, x_2, y_2, \mathbf{w}$.

$$\mathbf{w} = \left[0, \frac{1}{2}\right],$$
$$x_1 = -2,\ \ y_1 = 1,$$
$$x_2 = -1,\ \ y_2 = -1.$$

Problem 1 (iv)

(iii) Now, let's define the Gradient Descent update rule for some function $\text{TrainLoss}(\mathbf{w}) : \mathbb{R}^d \mapsto \mathbb{R}$. The rule helps us update the weights $\mathbf{w}$.

$$\mathbf{w} := \mathbf{w} - \eta \nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w}), \text{ where } \eta \text{ is the step size.} \tag{17}$$

Perform two iterations of Gradient Descent to minimize the objective function $\text{TrainLoss}(\mathbf{w}) = \frac{1}{2}\Big(\text{Loss}(x_1, y_1, w) + \text{Loss}(x_2, y_2, w)\Big)$ with values for x_1, y_1, x_2, y_2 from part (iii), using the weights update equation above. Use initialization $\mathbf{w}^0 = \left[0, \frac{1}{2}\right]$ and step size $\eta = \frac{1}{2}$.

Problem 4 (i)

Problem 2: More gradient computations

(i) Compute the gradient of the loss function below.

$$\text{Loss}(x, y, \mathbf{w}) = \sigma(-(\mathbf{w} \cdot \phi(x))y), \tag{4}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the logistic function.

Problem 4 (ii)

(ii) Suppose we have the following loss function.

$$\text{Loss}(x, y, \mathbf{w}) = \max\{1 - \lfloor(\mathbf{w} \cdot \phi(x))y\rfloor, 0\}, \tag{10}$$

where $\lfloor a \rfloor$ returns a rounded down to the nearest integer. Determine what the gradient of this function looks like, and whether gradient descent is suitable to optimize this loss function.

Looking Ahead: Linear Classification

- Now our weight vector defines a decision plane

- If the dot product with our weight vector is positive we assign a positive label to our data point, otherwise negative.

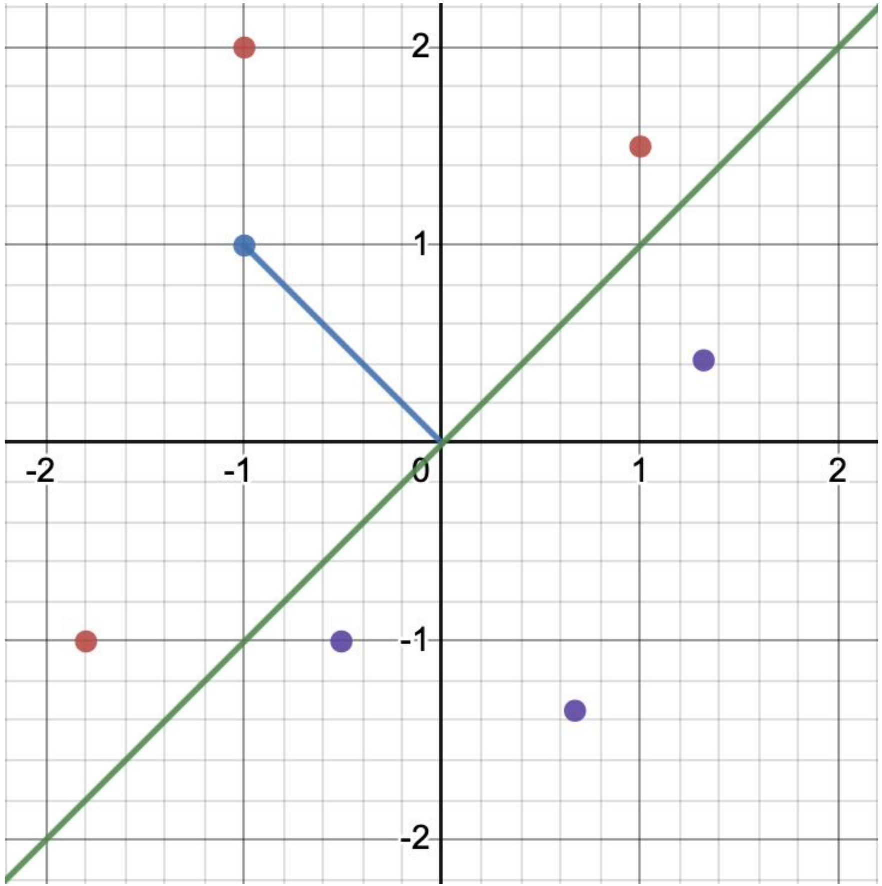
- Perpendicular to the weight vector is the decision plane.

$$w = [w_0 \quad w_1]$$

$$f_w(x) = sign(w \cdot x)$$

- Example:

$$w = [-1 \quad 1]$$
$$x_1 = [1 \quad 1.5]$$
$$f_w(x_1) = sign(-1 + 1.5) = sign(0.5) = +$$



Problem 3 (i)

4) **Problem 4 (Extra): Vector visualization**

Recall that we can visualize a vector $\mathbf{w} \in \mathbb{R}^d$ as a point in d-dimensional space. Let us now visualize some vectors in 2 dimensions on pen and paper.

(i) Consider $\mathbf{x} \in \mathbb{R}^2$. Draw the line (i.e. the "decision boundary") that separates between vectors having a positive dot product with weights $\mathbf{w} = [3, -2]$ and those having a negative dot product. Shade the part of the 2D plane that contains vectors satisfying $\mathbf{w} \cdot \mathbf{x} > 0$.

Hint: It might help to write out the expression for the dot product and seeing the relation between x_1 and x_2 that leads to a positive dot product. You could also use the geometric interpretation of the dot product.

Problem 3 (ii)

(ii) Repeat the above for $\mathbf{w} = [2, 0]$ and $\mathbf{w} = [0, 2]$.

Problem 3 (iii)

(iii) A small twist: visualize the set of vectors where $\mathbf{w} \cdot \mathbf{x} \geq 1$ for $\mathbf{w} = [3, -2]$.

Problem 3 (iv)

(iv) Consider the following element-wise inequality notation. For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$,

$$\mathbf{a} \leq \mathbf{b} \iff a_i \leq b_i \ \forall i = 1, 2, \ldots d. \tag{18}$$

Suppose we have a matrix $A \in \mathbb{R}^{2\times 2}$ and a vector $\mathbf{b} \in \mathbb{R}^2$ as follows.

$$A = \begin{bmatrix} 3 & -2 \\ 2 & 0 \end{bmatrix}, \mathbf{b} = [1, 0]. \tag{19}$$

Visualize the set of vectors where $A\mathbf{x} \geq \mathbf{b}$. Hint: A matrix vector product is a collection of dot products, and the above set can be obtained by the intersection of two of the sets constructed in the previous questions.

Any final questions?

Thank You