

So far we reviewed:

- Vectors, matrices
- Operations on vectors and matrices: scalar multiplication, addition, dot product, matrix multiplication
- Matrices as operators (linear functions / transformations)
- Linearity and linear combinations
- Solving systems of linear equations, elimination
- Finding matrix inverse

## Recap: solutions to systems of linear equations

For a  $n \times m$  matrix  $A$

- Square,  $n = m$ 
  - Unique solution if  $A$  is full rank  $n = r$
  - Otherwise,
    - Infinite solutions if  $b$  is in the column space of  $A$
    - No solutions otherwise
- Rectangular,  $n < m$  (wide matrix)
  - Infinite solutions if  $b$  is in the column space of  $A$
  - No solutions otherwise
- Rectangular,  $n > m$  (tall/thin matrix)
  - Unique solution if  $b$  is in the column space of  $A$
  - No solutions otherwise

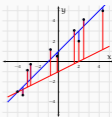
## Linear regression

Linear regression is about finding a linear model of the form,

$$y = w_1 x + w_0$$

where,

- $y$  is a numeric quantity we want to predict
- $x$  is a measurement/value helpful for predicting  $y$
- $w_0$  and  $w_1$  are the parameters that we want to learn from data
- both  $x$  and  $y$  can be vector valued



## Linear regression: and alternative view

this lecture

- Linear regression is also about finding the closest solution to a system of equations without a solution
- Given a dataset like

$x_1$	$x_2$	$y$
250.39	5.21	4913.19
332.18	3.77	59.67
312.47	1.26	154.42
272.01	7.01	166.27

- Find the closest solution to  $Xw = y$
- In other words, we solve  $Xw = p$ , where  $p$  is a vector that allows the system to be solved, and it the closest such vector to  $y$

## A simple example

- Let's take

$$x = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

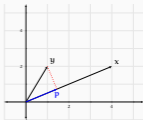
- We want to solve,

$$xw = y$$

- Instead we solve,

$$xw = p$$

where  $p$  is the orthogonal projection of  $y$  onto the line defined by  $x$



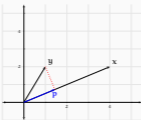
## Finding the projection

- $p$  is a scalar multiple (linear combination) of  $x$ :  $p = xw$
- We know that the length of  $p$  is the normalized dot product  $x^T y / \|x\|$
- We get the projection, if we multiply this with the unit vector in  $x$  direction

$$p = \frac{x \cdot x^T y}{\|x\| \|x\|} = \frac{xx^T y}{\|x\|^2} = \frac{xx^T y}{x^T x}$$

- $w$ , in this case is also easy:

$$w = \frac{x^T y}{x^T x}$$



## Finding the projection

a slightly different explanation

- Note that  $e = y - p$
- Since  $x$  and  $e$  are orthogonal

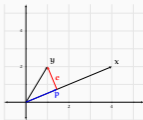
$$x^T (y - xw) = 0$$

$$x^T y - x^T x w = 0$$

$$w = \frac{x^T y}{x^T x}$$

- Since we defined  $p = xw$ ,

$$p = x \frac{x^T y}{x^T x} = \frac{xx^T y}{x^T x}$$



## Solution to the simple regression example

For our example,

$$x = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- Our 'training' gives us

$$w = \frac{x^T y}{x^T x}$$

- For future  $x$  values, the prediction of  $y$  is

$$y = wx$$

- $w = \frac{5}{2}$
- The model:

$$y = \frac{5}{2}x$$

Questions:

- what is the error  $e$  on the training instances?
- what is  $e^T x$ ?

## The other picture of the solution

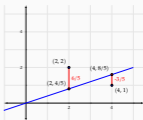
- The model:  $y = \frac{5}{2}x$
- Predictions:

$$p = \begin{bmatrix} 4 \times 2.5 \\ 2 \times 2.5 \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$$

- Error:

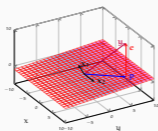
$$e = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 10 \\ 5 \end{bmatrix} = \begin{bmatrix} -9 \\ -3 \end{bmatrix}$$

- Is this a good model?



## Linear regression in higher dimensions

- In higher dimensional spaces we want the projection onto the column space of  $X$
- The error vector  $e$  is perpendicular to all column vectors of  $X$ ,  $x_i$
- Again, note that  $e = y - p$



## Deriving linear regression on higher dimensions

$$X^T (y - p) = 0 \quad \text{Error vector is orthogonal to columns}$$

$$X^T (y - Xw) = 0 \quad p \text{ is the weighted combination of columns}$$

$$X^T X w = X^T y \quad \text{Note: } X^T X \text{ is square}$$

$$w = (X^T X)^{-1} X^T y \quad \text{The final solution}$$

The projection of  $y$  onto columns space of  $X$  is

$$p = X(X^T X)^{-1} X^T y$$

## The intercept (bias) term

- The models we fit so far are 'linear',

$$\hat{y} = w_1x_1 + w_2x_2 + \dots + w_nx_n$$

they are forced to include  $y = 0$  for  $x = 0$

- In most (almost all) cases, this is too restrictive, we also want to learn an intercept term

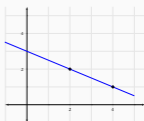
$$\hat{y} = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$

- A straightforward solution is to include an artificial column of 1s in the input matrix  $X$

$$X = \begin{bmatrix} 1 & 4 \\ 1 & 2 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

## Solution with the intercept term

- Solution:  $w_0 = 3, w_1 = -1/2$
- The model:  $y = 3 - 1/2x$
- Is this a better model?



## Regression in the real world

- In this lecture, we focused on finding the best fit to the data
- This may (very likely) result in *overfitting*
- To prevent overfitting, we
  - use regularization
  - never rely on performance on the training set**, success should only be measured on a *held-out* data set
- We will return to these concepts later

## Summary / next

- We reviewed regression as a way to find an approximate solution to a system of linear equations
  - We will come back to regression multiple times
- Next:
- Determinant, eigenvalues/eigenvectors, SVD

## Further reading

Any of the linear algebra references provided earlier.