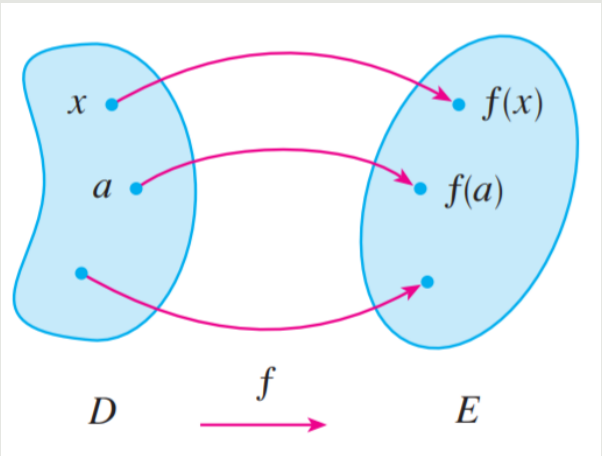
SUPERVISED LEARNING

Gradient Descent and Regression Problems in Machine Learning

MODELING

In general: it is the root of all sciences, a comprehensive framework of understanding reality. For many applications you can think of a model as a functional relationship between an input and an output.



Example of a Function by Arrow Diagram

The important elements of a function are

i. the *domain* of the function
ii. the *range* of the function
iii. the *definition* of the function.

We believe that when we have a lot of data collected (a lot of examples) we can train or update the model based on the minimization of the errors.

THE LEAST SQUARES METHOD

Carl Friedrich Gauss (1777-1855): was a German mathematician and physicist, widely considered one of the greatest mathematicians of all time. His contributions to a variety of fields have earned him the nickname "Prince of Mathematicians."

Had great contributions to mathematics and astronomy.

Proposed a rule to score the contributions of individual errors to overall error.

Reference

Portrait of Carl F. Gauss Source: gettyimages

LEAST SQUARES EXAMPLE

Assume we have five imprecise measurements, as shown below, what would be the correct answer?

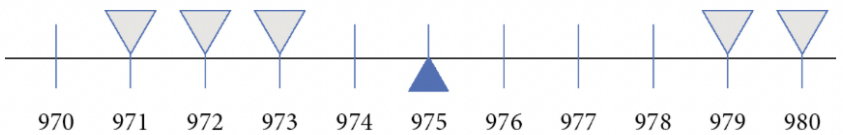


FIGURE 5: *Five measurements of the same length*

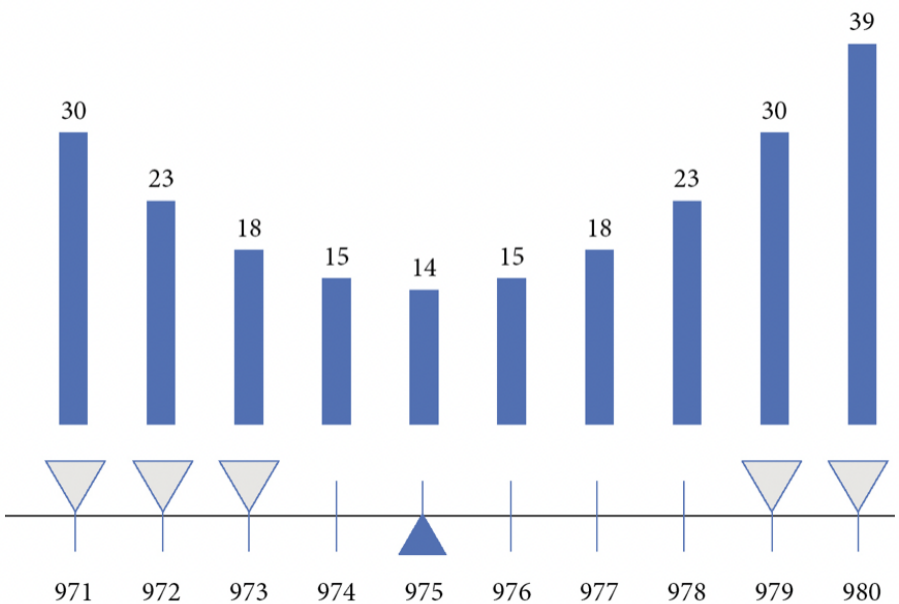


FIGURE 6: *Mean squared error (MSE) for ten possible values of the true length*

BIAS VS NOISE

Critical Thinking Prompts:

- What is the meaning of "noise" in this context?
- Can you describe a numerical situation where we may have bias?

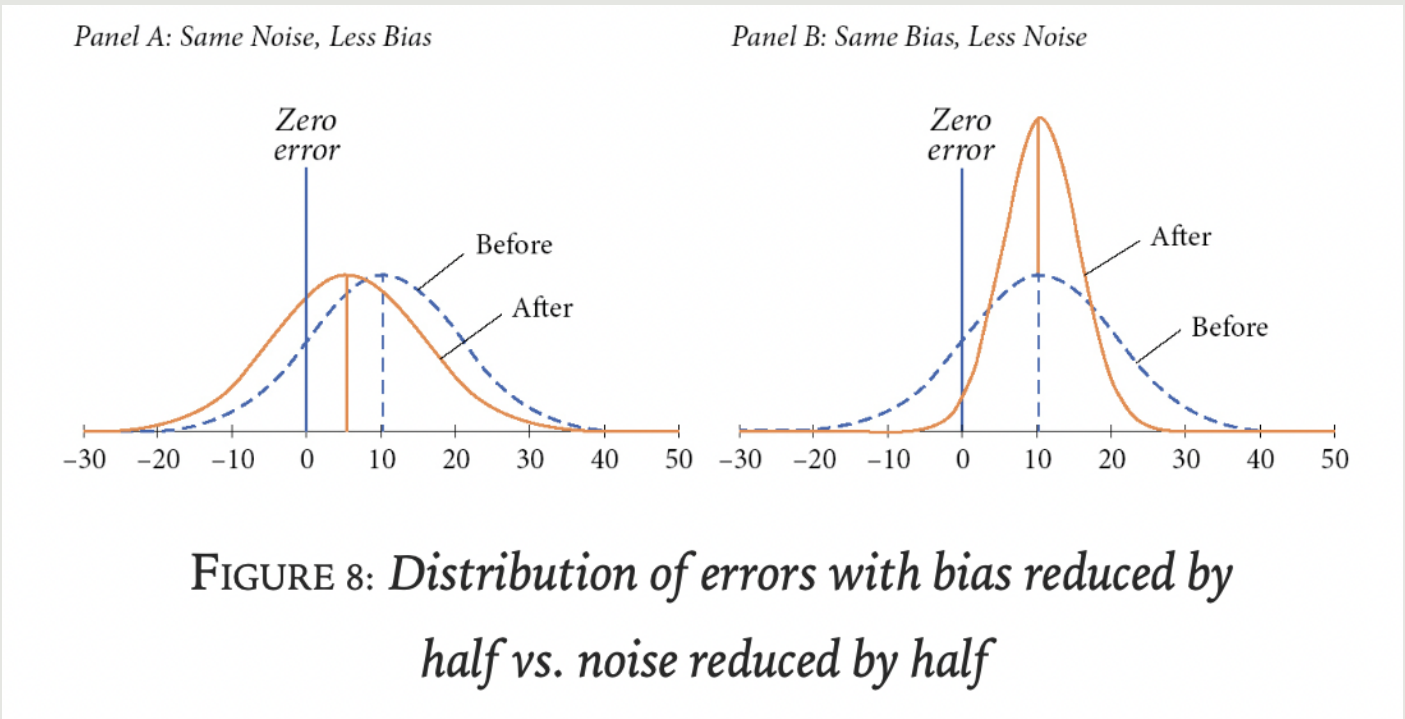


FIGURE 8: *Distribution of errors with bias reduced by half vs. noise reduced by half*

Source: "Noise, A Flaw in Human Judgement", D. Kahneman et al.

BIAS AND VARIANCE

The variance-bias tradeoff is a fundamental concept in machine learning and statistics that describes the balance between two sources of error that affect the performance of predictive models: bias and variance. Understanding this tradeoff is crucial for developing models that generalize well to new data.

BIAS

Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. High bias can cause an algorithm to miss relevant relations between features and target outputs (underfitting).

- **High Bias:** Assumptions in the model are too strong, making it overly simplistic.
- **Low Bias:** The model captures the true relationship more accurately.

VARIANCE

Variance refers to the error introduced by the model's sensitivity to the fluctuations in the training data. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

- **High Variance:** The model is too complex, capturing noise along with the underlying pattern.
- **Low Variance:** The model's predictions are stable across different training sets.

THE TRADEOFF

The tradeoff is the balance between these two sources of error. Increasing model complexity typically decreases bias but increases variance, and vice versa. The goal is to find the right balance to minimize the total error.

TOTAL ERROR

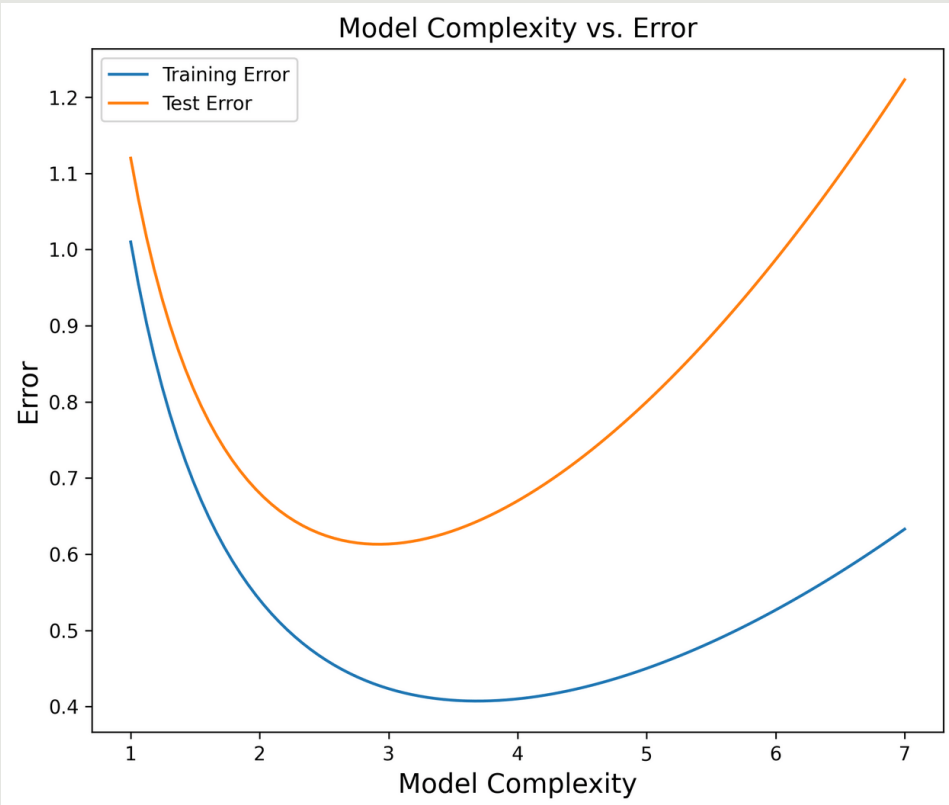
The total error in a model can be expressed as the sum of bias squared, variance, and irreducible error (noise):

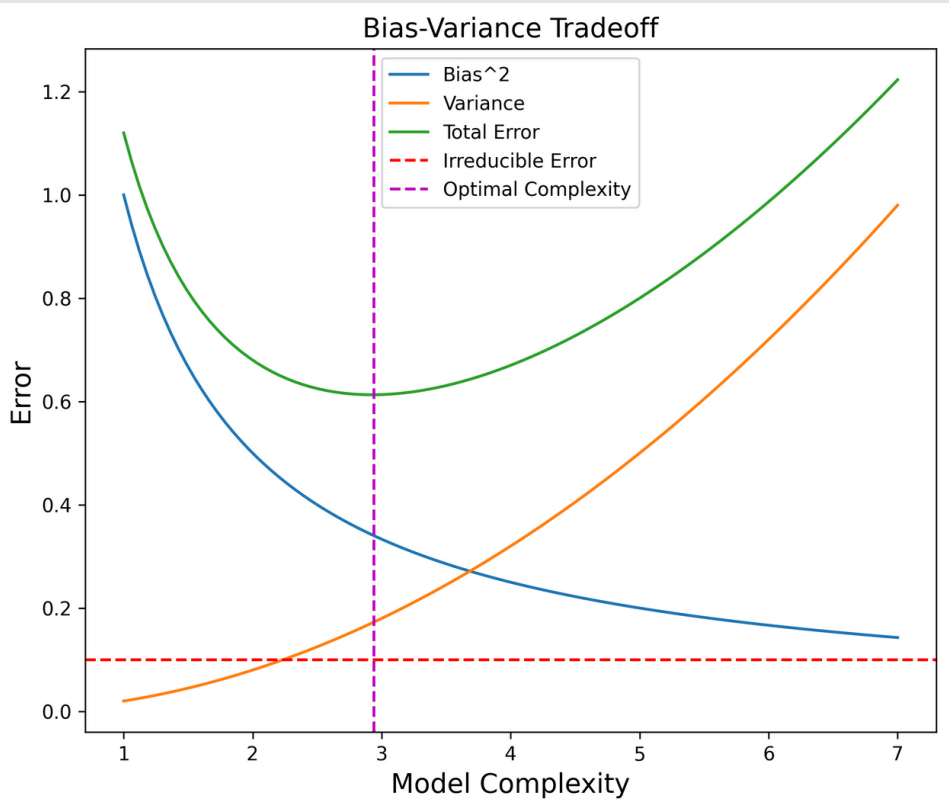
$$\text{Total Error} = (\text{Bias}^2) + \text{Variance} + \text{Irreducible Error}$$

The irreducible error is due to noise in the data itself and cannot be reduced by any model.

INTUITIVE VISUALIZATION

Here, we'll plot the training error and test error as functions of model complexity. Typically, as model complexity increases, the training error decreases, while the test error first decreases and then increases, forming a U-shaped curve.





INTERPRETATION

1. **Model Complexity vs. Error:**
   - The training error decreases as model complexity increases because the model can better fit the training data.
   - The test error decreases initially but starts to increase after a certain point, indicating overfitting. The test error curve shows a local minimum, reflecting the optimal model complexity where the tradeoff between bias and variance is balanced.
2. **Bias-Variance Decomposition:**
   - Bias squared decreases with increasing model complexity, as the model becomes more capable of capturing the underlying patterns.
   - Variance increases with model complexity, as the model starts to fit the noise in the training data.
   - The total error curve shows a local minimum at an optimal model complexity, balancing bias and variance. This minimum is marked with a vertical green dashed line.
   - The intersection point of the bias and variance curves is not necessarily the point of minimum total error.

MEAN SQUARED ERROR

From many *independent* imperfect measurements, we can easily derive a better one. Here, the accent goes on the concept of independence. We say that two events, A and B, are independent if

$$\mathrm{P}\left(A \cap B\right) = \mathrm{P}\left(A\right) \cdot \mathrm{P}\left(B\right)$$

Observation: While the above statement is correct, it is also deceivingly simple, and it's typically hard to use in practice.

Now, assume we have n independent imperfect, however unbiased, measurements of an object labeled as

$$x_1, x_2, \ldots x_n$$

and we assume that there is an ideal measurement x^* and we want to find it. Gauss' idea is to create an objective function that weighs more heavily larger errors. This can be formalized by assuming that we want to compute a value x that minimizes the sum of the squared errors/deviances:

$$\mathrm{SSE}\left(x\right) \triangleq \left(x - x_1\right)^2 + \left(x - x_2\right)^2 + \ldots \left(x - x_n\right)^2$$

$$\mathrm{MSE}\left(x\right) \triangleq \frac{\left(x - x_1\right)^2 + \left(x - x_2\right)^2 + \ldots \left(x - x_n\right)^2}{n}$$

THE BEST POINT ESTIMATE

Important: When the number of observations is fixed, n is just a constant.

So, in this case, we want to determine

$$\underset{x,\in,\mathbb{R}}{\mathbf{argmin}}\mathbf{SSE}(x)$$

that is to say, we want the value of x that yields the minimum of MSE.

To solve the problem, we take the derivative of the MSE with respect to x and set it equal to zero:

$$2(x - x_1) + 2(x - x_2) + \ldots 2(x - x_n) = 0.$$

When you solve this equation, you get that

$$x = \frac{x_1 + x_2 + \ldots x_n}{n}$$

This highlights the idea that the average is the best representative of a sample.

THE GRADIENT DESCENT METHOD

We want to minimize a cost function that depends on some coefficients. An example in 2-D is

$$L(w_1, w_2) := \frac{1}{n} \sum (y_i - w_1 \cdot x_{i1} - w_2 \cdot x_{i2})^2$$

Here we have a vector

$$\vec{w} := (w_1, w_2)$$

We can think of the vector $\vec{w}$ having (in this case) two components and a perturbation of $\vec{w}$ in some direction such as $\vec{v}$.

We consider the function $g(t) := L(\vec{w} + t \cdot \vec{v})$ we get some important ideas:

i. if $\vec{w}$ is ideal for the cost then $t = 0$ is min of the function g and so $g'(0) = 0$.

ii. if $\vec{w}$ is not minimizing the cost then we want to decrease the function g.

Here, we want that

$$g'(t) := \nabla L \cdot \vec{v}$$

to be negative (because we want to decrease the value of g).

LOOKING FOR SOLUTIONS

We want the gradient

$$g'(t) := \nabla L \cdot \vec{v}$$

to be negative (because we want to decrease the value of g).

Consider

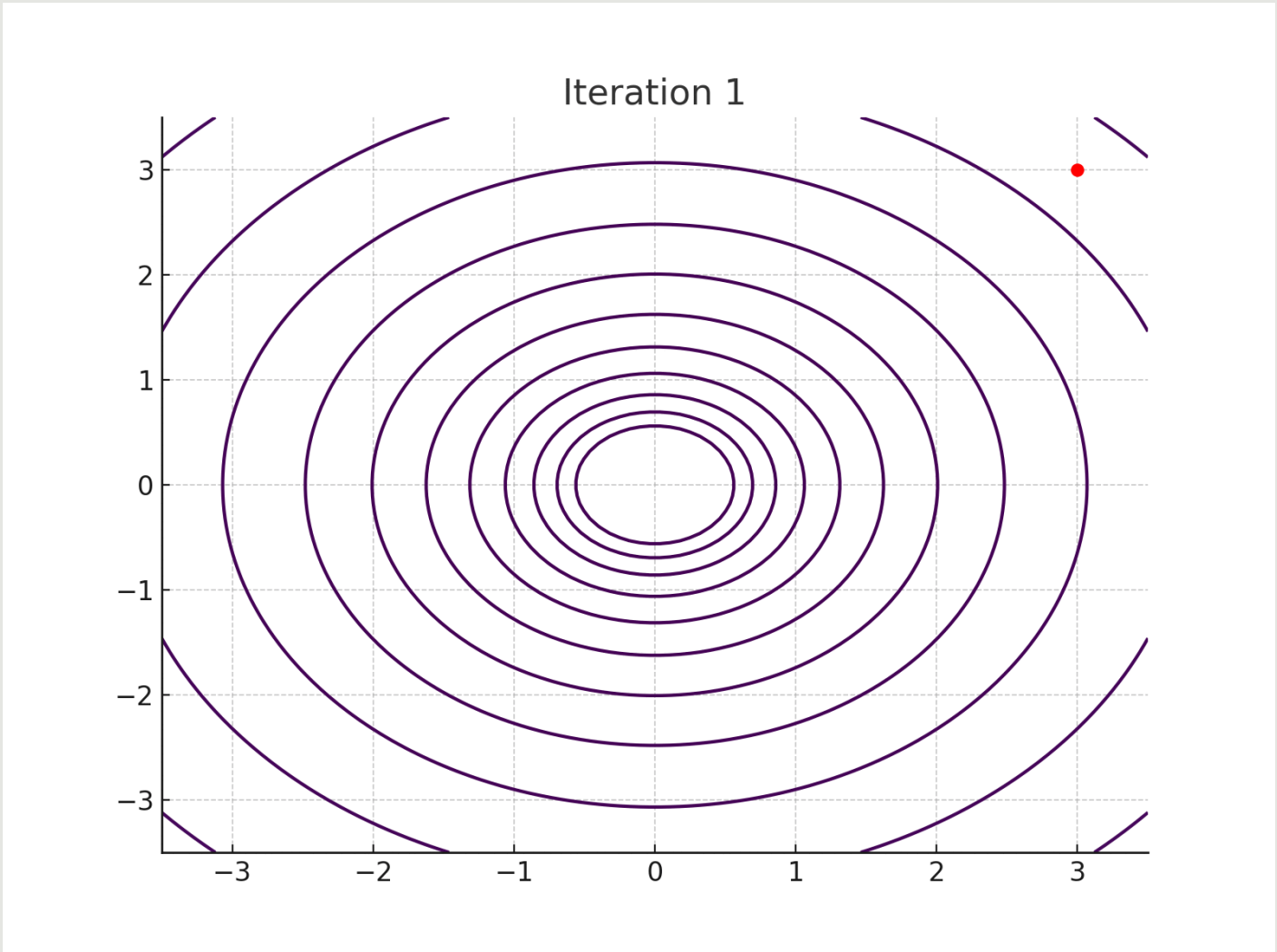
$$\vec{v} := -\nabla L$$

This means that the coefficients should be updated in the negative direction of the gradient, such as:

$$\vec{w}_{new} := \vec{w}_{old} - \mathrm{lr} \cdot \nabla L$$

This is called
Learning Rate

INTUITIVE VISUALIZATION



EXAMPLES OF LOSS FUNCTIONS AND GADIENTS

THE LOSS FUNCTION (SSE) AND ITS GRADIENT FOR N OBSERVATIONS

$$L(w_1, w_2) := \sum_{i=1}^{n} (y_i - w_1 \cdot x_{i1} - w_2 \cdot x_{i2})^2$$

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{n} 2 \cdot (y_i - w_1 \cdot x_{i1} - w_2 \cdot x_{i2}) \cdot (-x_{i1})$$

A 2D EXAMPLE WITH PARTIAL DERIVATIVES AND ONE OBSERVATION

Suppose $L(w_1, w_2) := (y - w_1 \cdot x_1 - w_2 \cdot x_2)^2$

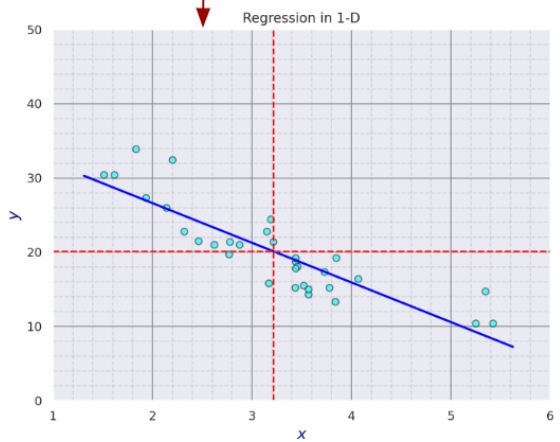
What are the partial derivatives of L with respect to w_1 and w_2?

$$\frac{\partial L}{\partial w_1} = 2 \cdot (y - w_1 \cdot x_1 - w_2 \cdot x_2) \cdot (-x_1)$$

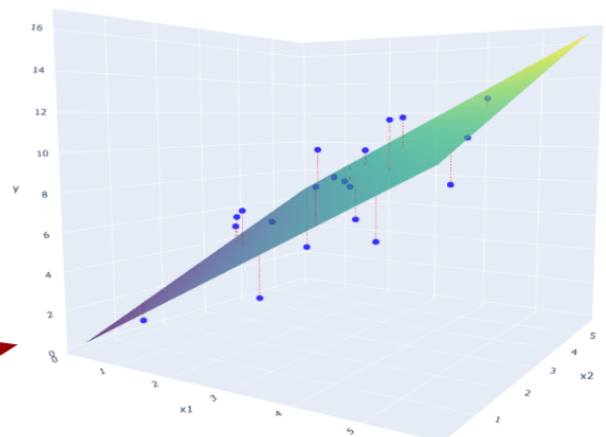
$$\frac{\partial L}{\partial w_2} = 2 \cdot (y - w_1 \cdot x_1 - w_2 \cdot x_2) \cdot (-x_2)$$

LINEAR REGRESSION

In 2 dimensions: $y = \beta_0 + \beta_1 x$



In 3 dimensions: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



In 2 dimensions: $y = \beta_0 + \beta_1 x$

In 3 dimensions: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

In p dimensions:
(data has p variables)

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Bias Term

Linear Combination of Variables